ORGANIZAÇÃO

TIAGO CRUZ DE FRANÇA JOSÉ LUIZ T. NOGUEIRA

SBC
Sociedade Brasileira de Computação

MINICURSOS DA

ERSI 2019

VI ESCOLA REGIONAL DE SISTEMAS DE INFORMAÇÃO

DUQUE DE CAXIAS - RJ

UFRRJ
UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

UNIVERSIDADE UNIGRANRIO

EDITORA: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO

Coordenadores
Tiago Cruz de França
José Luiz Thomaselli Nogueira
João Francisco Antunes

MINICURSOS DA

VI ESCOLA REGIONAL DE SISTEMAS DE INFORMAÇÃO

DUQUE DE CAXIAS - RJ

Realização
Sociedade Brasileira de Computação - SBC
Universidade do Grande Rio - UNIGRANRIO
Universidade Federal Rural do Rio de Janeiro - UFRRJ

Dados Internacionais de Catalogação na Publicação (CIP)

E612

Escola Regional de Sistemas de Informação (6. : 2019: Duque de Caxias, RJ).

Minicursos da VI Escola Regional de Sistemas de Informação [recurso eletrônico] : 06 a 09 de novembro de 2019, UNIGRANRIO – Duque de Caxias / coordenadores Tiago Cruz de França, José Luiz Thomaselli Nogueira, João Francisco Antunes. – Rio de Janeiro : SBC, 2019.

1 recurso eletrônico. (157 p.)

Inclui bibliografia.

ISBN 978-85-7669-488-5 (e-book)

1. Tecnologia da informação. 2. Informática. 3. Escola Regional de Sistemas de Informação (ERSI). I. França, Tiago Cruz de. II. Nogueira, José Luiz Thomaselli. III. Antunes, João Francisco. IV. Sociedade Brasileira de Computação. V. Universidade do Grande Rio (Unigranrio). VI. Título.

CDD 004

Ficha catalográfica elaborada pela bibliotecária Tatiana Ribeiro – CRB-7/6199

Prefácio

A Escola Regional de Sistemas de Informação do estado do Rio de Janeiro (ERSI-RJ) é um evento que reúne profissionais, professores e estudantes interessados em aprender e discutir problemas, soluções e conceitos relacionados a Sistemas de Informação. Os minicursos são atividades de curta duração (quatro horas) que fazem parte da programação da ERSI-RJ. Eles abordam temas relacionados a Sistemas de Informação com objetivo de proporcionar ao público da escola um ambiente de aprendizagem e de discussão de tendências e desafios na área de Sistemas de Informação.

Este ano foram selecionados cinco minicursos entre quatorze propostas submetidas à ERSI-RJ. A seleção foi realizada por um comitê formado por vinte e seis avaliadores. Todos professores e profissionais de Sistemas de Informação. Os critérios de seleção das propostas foram: relevância para o evento, expectativa de público, atualidade e conteúdo.

Os temas dos minicursos selecionados abordam algoritmos de recomendação, análise de sentimentos, LGPD (Lei Geral de Proteção de Dados) e Bancos de Dados, ensino de matemática a alunos com deficiência visual e fusão de dados em ambientes inteligentes. O material de referência dos minicursos compõem a segunda edição de livro de minicurso da ERSI-RJ.

No primeiro capítulo, "Conceitos, Implementação e Dados Privados de Algoritmos de Recomendação"), os autores adotaram uma abordagem prática e simples para apresentar diferentes algoritmos de recomendação de informação. Em seguida, eles discutiram a importância desses algoritmos e os desafios de privacidade relacionados à recomendação de informação. No segundo capítulo, "Introdução à Análise de Sentimentos com *Word Clouds*", os autores introduziram conceitos de análise de sentimentos em textos, apresentando exemplos e estratégias de visualização dos resultados das análises. No terceiro capítulo, "LGPD em Ambientes de Bancos de Dados nas Organizações", as autoras apresentaram os principais pontos da lei, sua influência em ambientes de bancos de dados nas organizações e apresentaram exemplos práticos de suporte operacional para atender princípios da LGPD. No quarto capítulo, "Técnica de Ensino de Matemática para Alunos com Deficiência Visual com suporte Informatizado", os autores apresentaram uma nova metodologia para ensino de matemática para alunos deficientes visuais usando o computador. O quinto capítulo, "Fusão de dados para Ambientes Inteligentes", apresentaram os principais conceitos de fusão de dados para serviços de ambientes inteligentes e apresentaram exemplos práticos de implementação.

Acreditamos que este material será útil em aulas de Sistemas de Informação; em discussões sobre novas abordagens de pesquisa suportando trabalhos atuais e futuros; e para apoiar a prática profissional. Parabenizamos e agradecemos aos autores dos minicursos. Agradecemos também ao Comitê de Seleção de Propostas pela dedicação e eficiência; ao Comitê Editorial pelo empenho; e a CESI (Comissão Especial de Sistemas de Informação) da SBC pelo apoio para publicação deste livro.

Tiago Cruz de França (UFRRJ)
Coordenador de Minicurso da ERSI-RJ 2019

VI Escola Regional de Sistemas de Informação do Estado do Rio de Janeiro

06 a 09 de Novembro de 2019

Duque de Caxias – RJ – Brasil

MINICURSOS

Promoção

Sociedade Brasileira de Computação – SBC

Coordenação da Escola Regional de Sistemas de Informação do Estado do Rio de Janeiro 2019

Thiago Silva de Souza – UNIGRANRIO
Daniel de Oliveira – UNIGRANRIO

Coordenação de Painéis e Palestras

Raphael Carlos Santos Machado – INMETRO
Rafael Elias de Lima Escalfoni – CEFET/RJ

Coordenação de Sessões Técnicas e Pôsteres

Claudio Miceli de Farias – NCE/UFRJ
Kele Teixeira Belloze – CEFET/RJ

Coordenação de Minicursos

Tiago Cruz de França – UFRRJ
José Luiz Thomaselli Nogueira – UNIGRANRIO
João Francisco Antunes – UNIGRANRIO

Coordenação da Exposição de Robótica e IoT

Marco Antônio de Melo Britto – UNIGRANRIO
Alayne Duarte Amorim – Colégio Pedro II

Coordenação da Hackathon

Miguel Gabriel P. de Carvalho – UNIGRANRIO
Daniel de Oliveira – UNIGRANRIO

Comitê Institucional

Anderson Silva do Nascimento – UNIGRANRIO
João Francisco Antunes – UNIGRANRIO
Miguel Gabriel P. de Carvalho – UNIGRANRIO
Natália Joana Silva de Oliveira – UNIGRANRIO

Comitê de Programa

Alana Morais - IESP
Alessandro Copetti - UFF
André Luiz Leal - UFRRJ
Angelica Dias - UFRJ
Bernardo Peralva - UERJ
Breno de França - UNICAMP
Bruna Diirr - UNIRIO
Carlos Eduardo Pantoja – CEFET/RJ
Daniel Cruz de França – UFPB
Daniel Paiva - UFF
Daniel Schneider - UFRJ
Davi Viana - UFMA
Diego Brandão - CEFET/RJ
Eliezer Dutra - CEFET/RJ e UNIRIO
Emanuele Jorge - IFRJ
Fabiana Mendes - UnB
Fabio Gomes Rocha - UNIT
Fábio Silveira Vidal - IFTO
Flavio Horita - UFABC
Geiza Hamazaki - PUC-Rio
Gizelle Vianna - UFRRJ
Henrique Sousa – UFRRJ
Isabel Cafezeiro – UFF
José Ricardo Cereja - UNIRIO
Juliana França - UFRRJ
Laci Mary Manhães - UFF
Leonardo Azevedo - IBM Research, Brazil
Luis Orleans - UFRRJ
Luiz Felipe Mendes - UFJF
Marcelo Fornazin - UFF
Marco Antonio Araujo - UFJF e IF Sudeste-MG
Mario Dantas - UFJF
Nilton Rizzo - UFRRJ
Paulo Sérgio Santos - UNIRIO

Priscila Goliatt - UFJF
Rafael Escalfoni - CEFET/RJ
Rafael Lima de Carvalho - UFTO
Raimundo José Macário Costa - UFRRJ
Rita Suzana Pitangueira Maciel - UFBA
Rodrigo Monteiro - UFF
Rodrigo Santos - UNIRIO
Scheila de Avila e Silva - UCS
Tiago Cruz de França - UFRRJ

Comitê de Seleção de Propostas de Minicursos

Alana Morais - IESP
Alessandro Cerqueira -Univeritas-RJ
André Luiz de Castro Leal - UFRRJ
Bruna Diirr - UNIRIO
Bruno Nascimento - UFRJ
Daniel Oliveira - UNIGRANRIO
Danilo S. Carvalho - UFRJ
Diego Pessoa - IFPB
Edgar Sarmiento - UNSA (Peru)
Eduardo Goncalves - ENCE/IBGE
Eduardo Hargreaves - Petrobras
Emanuele Jorge - IFRJ
Gizelle Vianna - UFRRJ
Helvio Junior - UFRJ
Jesus Talavera Portocarrero - Nuance Communications
Lívia Ruback - UFRRJ
Marcelo Cruz - UFRRJ
Marcos Arrais - PUC-MG
Rafael Bernardo Teixeira - UFRRJ
Rafael Escalfoni - CEFET-RJ
Raimundo Costa - UFRRJ
Renata Araujo - Mackenzie
Robson Silva - UFRRJ
Tadeu Classe - UNIRIO
Talita Ribeiro - UFRJ
Tiago Cruz de França - UFRRJ

Comitê Editorial

Tiago Cruz de França – UFRRJ
André Luiz Leal – UFRRJ
João Francisco Antunes – UNIGRANRIO

Arte da capa de Matheus Nunes Ritton – UFRRJ

Sumário

Capítulo 1

Conceitos, Implementação e Dados Privados de Algoritmos de Recomendação

Leonardo Herdy Marinho, Rodrigo Campos, Rodrigo Pereira dos Santos, Mônica Ferreira da Silva e Jonice Oliveira

Abstract

Through the recommendation algorithms, it is possible to suggest items relevant to users, increasing the proximity to their interest. These facilities also aim to reduce the time that would be spent searching for desired items. These algorithms can be applied in many scenarios, presenting relevant results in solving various real-word problems. In this context, the purpose of this chapter is to simplify and present the concepts of recommendation algorithms, demonstrating how these techniques work. Concepts and challenges involving data privacy in these algorithms are also presented. Finally, this chapter introduces Python programming language operations and applies the recommendation techniques of the collaborative filtering approach, using the cosine similarity.

Resumo

Por meio dos algoritmos de recomendação, é possível sugerir itens relevantes para usuários, aumentando a proximidade com o interesse dos mesmos. Essas facilidades visam também reduzir o tempo que seria dispensado na busca de itens desejados. Esses algoritmos podem ser aplicados em muitos cenários, apresentando resultados relevantes na solução de diversos problemas. Nesse contexto, o objetivo deste capítulo é simplificar e apresentar os conceitos sobre algoritmos de recomendação, demonstrando como essas técnicas funcionam. São apresentados ainda, conceitos e desafios envolvendo a privacidade de dados nesses algoritmos. Por fim, esse capítulo apresenta operações com linguagem de programação Python e aplica as técnicas de recomendação da abordagem filtragem colaborativa, utilizando a similaridade cosine.

1.1. Introdução

Os algoritmos de recomendação implementam filtros de informação visando apresentar itens ou objetos como: páginas web, filmes, músicas, livros, medicamentos, lojas e artigos que provavelmente são do interesse do usuário. Algoritmos de recomendação são amplamente utilizados por grande parte das gigantes redes lojistas, sites focados em entretenimento, redes sociais, players de música e mais uma gama de prestadores de serviços ou vendedores de produtos. É um tema em alta em toda a comunidade científica e mercadológica. O conjunto desses algoritmos e técnicas é chamado de sistema de recomendação.

Os algoritmos de recomendação podem agir sem o acionamento do usuário e por essa razão, alguns usuários podem não notar que determinado *website* ou sistema tenha um algoritmo de recomendação. Nos últimos anos, muitas aplicações que envolvem comércio eletrônico e buscas utilizam algum tipo de mecanismo de recomendação.

Os algoritmos de recomendação têm desempenhado um importante papel na solução do problema de sobrecarga de informações [Ricci et al. 2015]. Entretanto, é preciso considerar também que esses algoritmos apresentam riscos de privacidade e algumas preocupações devem ser observadas na implementação e utilização de dados pelos recomendadores [Feng et al. 2018].

Nesse contexto, esse capítulo explora conceitos dos algoritmos de recomendação, buscando introduzir o assunto e suas principais técnicas. São tratados exemplos do cotidiano que utilizam esses algoritmos, bem como quais são os principais tipos de algoritmos, organizados após uma revisão da literatura. Considerando a utilização da linguagem Python em diversas soluções de recomendação, bem como a variedade de bibliotecas para ciência de dados, esse capítulo introduz as principais operações da linguagem Python. Dessa forma, implementamos um algoritmo de recomendação utilizando métodos de similaridade “cosine”.

Esse capítulo está organizado da seguinte forma: a Seção 1.2 aborda os conceitos dos algoritmos de recomendação, destacando o surgimento e histórico, bem como os principais conceitos do processo de busca e recuperação da informação; a Seção 1.3 apresenta exemplos da aplicação dos algoritmos de recomendação na indústria, bem como os diferenciais a nível de usuário de cada uma das soluções; a Seção 1.4 apresenta os principais tipos de abordagens de recomendação e as diferenças investigadas na literatura; a Seção 1.5 aborda alguns problemas e soluções para atender a privacidade de dados nos recomendadores; a Seção 1.6 tem um enfoque em conhecimentos básicos sobre Python; na Seção 1.7 é apresentado como implementar um recomendador de filmes em Python utilizando um *dataset* pré-selecionado; as principais oportunidades de pesquisa e desafios em sistemas de recomendação são abordados na Seção 1.8; e por fim, a Seção 1.9 conclui o capítulo com considerações finais.

1.2. Algoritmos de Recomendação

Os usuários da internet buscam constantemente por informações que possam auxiliá-los em seu cotidiano (Figura 1.2). Com a grande massa de dados, produtos, serviços e opções que estão disponíveis, encontrar o que realmente é procurado se tornou uma tarefa árdua [Cazella et al. 2008].

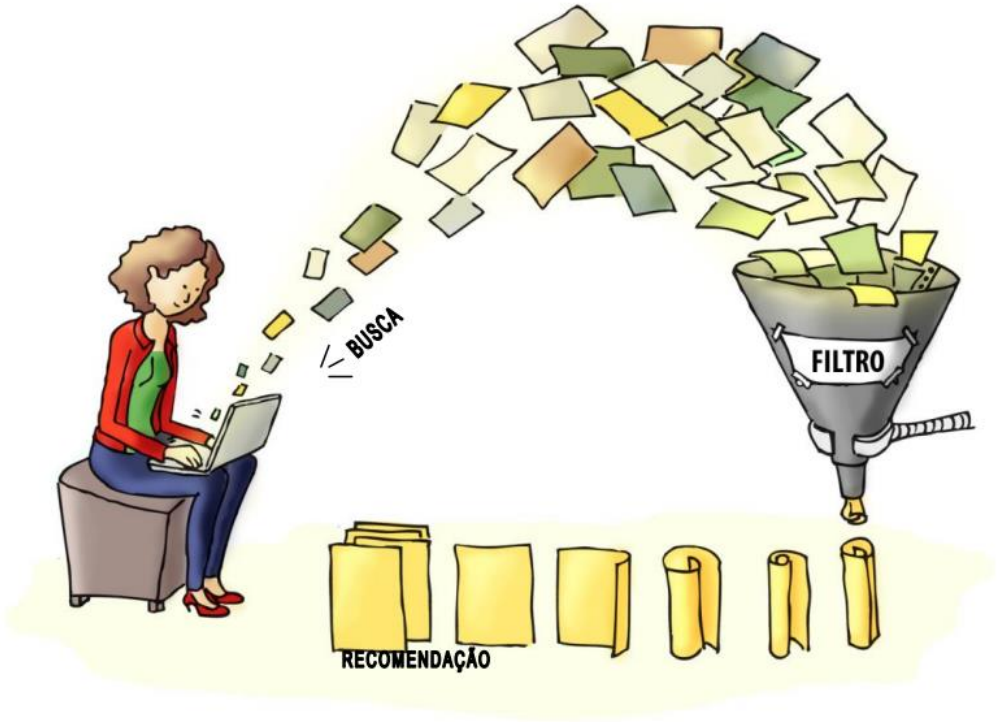


Figura 1.2. Funcionamento de um algoritmo de recomendação

Fonte: [Motta et al. 2011]

Os recomendadores surgem com o objetivo de propor soluções para essa tarefa de recuperar uma informação de acordo com a necessidade do usuário. Essa seção apresenta como esses sistemas evoluíram desde o surgimento, bem como os conceitos do processo de busca de uma determinada informação. Por fim, é apresentado como funciona a lógica de recomendação.

1.2.1. Dados Históricos

A necessidade de construir algoritmos que recomendem algo é bem recente se compararmos com o tempo histórico que a computação existe. O nascimento de tal necessidade teve seu início com o aumento do poder de armazenamento de dados ocorrido principalmente no início da década de 90, que dificultou aos usuários encontrarem de forma rápida e fácil algo que buscam em uma pesquisa realizada na *web*. Anteriormente a análise realizada pelo usuário era bastante simples, afinal, as buscas retornavam um valor substancial relativamente pequeno de itens que eram facilmente triados pelo usuário em poucos minutos. Tal realidade mudou no início dos anos 90 com o advento da Internet e a interconexão entre milhares senão milhões de pessoas, serviços, sensores e mais uma gama de geradores de dados. A cada dia mais dados eram gerados e armazenados, tornando assim, complexa a tarefa de encontrar o que era relevante ao usuário para ser exibido logo nas primeiras páginas de resultados das buscas.

Com o surgimento de serviços de música online, vídeo, venda de produtos e afins, aumentou o desafio de não tomar o tempo dos usuários com conteúdo irrelevante ao gosto deles. As grandes companhias que investiam no “setor online” entenderam que o desperdício do tempo dos usuários ocasionava a redução do consumo online, ou seja,

quanto mais tempo um usuário desperdiçasse para encontrar algo que estava buscando maiores eram as chances de desistência. Isso desencadeou a corrida para o "algoritmo perfeito", mais conhecido como o algoritmo de recomendação, que trouxesse o mais rápido possível exatamente o que o usuário esperava encontrar.

Um dos primeiros sistemas com algoritmos de recomendação conhecido foi o RINGO [Shardanand e Maes 1995]. Segundo Motta et al. (2011), o RINGO trabalhava com a recomendação de músicas a partir do perfil do usuário. As recomendações eram criadas partindo das informações explicitamente fornecidas pelos próprios usuários, ou seja, os usuários precisavam dizer que não gostavam de um ritmo *r* para que o algoritmo não exibisse músicas desse ritmo, por exemplo. O algoritmo ajustava as recomendações de músicas continuamente a partir do uso prolongado, além de realizar sucessivas recomendações analisando o feedback dos usuários. O Grouplens criado aproximadamente na mesma época que o RINGO, na década de 90, implementou uma arquitetura genérica para as recomendações de notícias. Mais tarde, o sistema do Grouplens foi evoluído para o MovieLens [Konstan et al. 1997] que consistia em trabalhar com a sugestão de filmes geradas a partir da correlação entre a avaliação dos usuários sem utilizar características ou parâmetros definidos previamente pelo usuário. Esses trabalhos foram de suma importância para consolidar o marco inicial da implementação e evolução dos atuais algoritmos de recomendação. Ao final desse capítulo, utilizamos um conjunto de dados do MovieLens para explorar recomendações de filmes.

A tecnologia construída para o funcionamento do RINGO posteriormente foi transformada em um produto chamado Firefly, um site de recomendação de músicas, artistas e livros. Após inúmeras parcerias, o Firefly difundiu o sistema de recomendação e a filtragem colaborativa. Posteriormente grandes empresas adotaram a ideia da tecnologia de recomendação do Firefly. como o Yahoo, Amazon e eBay.

1.2.1. Busca e Recuperação da Informação

Recuperação de Informação (RI) é geralmente considerada um sinônimo de recuperação de documentos, mas na verdade, os processos dos sistemas de RI produzem a recuperação do que seria o objetivo principal do usuário. As abordagens desenvolvidas para esse fim são aplicáveis a uma série de tarefas relacionadas ao processamento de informações existentes, tanto na recuperação de dados quanto na recuperação de conhecimento [Jones e Willett 1997]. Dentre essas possibilidades de aplicação, estão os algoritmos de recuperação

A recuperação de dados, no contexto de um sistema de RI, consiste em identificar quais documentos em uma determinada coleção contêm as palavras-chave da consulta do usuário. No entanto, o usuário está interessado em recuperar determinadas informações sobre um assunto, ou seja, dados em um contexto desejado, e não apenas em recuperar dados que satisfaçam uma consulta. Devido a isso, a estratégia de recuperação surge para resolver essa tarefa subjetiva. Essa estratégia consiste no uso de um algoritmo, no processo de recuperação e classificação de sistemas de RI, que identifica o coeficiente de similaridade entre uma consulta de usuário em um determinado contexto e um conjunto de documentos coletados [Santos e Bräscher 2017].

O termo recuperação de informação refere-se a uma pesquisa que pode abranger qualquer forma de informação: dados estruturados, texto, vídeo, imagem, som, notas musicais ou sequências de DNA [Grossman e Frieder 2004]. Em [Baeza-Yates e Ribeiro-Neto 2013], é destacado que a RI fornece aos usuários acesso fácil a informações de interesse, lidando com representação, armazenamento, organização e acesso a itens de informação que devem fornecer aos usuários facilidade de acesso a informações. Portanto, o objetivo principal da RI é recuperar todos os documentos relevantes às necessidades de informações do usuário e, ao mesmo tempo, recuperar o mínimo possível de documentos irrelevantes. Portanto, o principal desafio não é apenas extrair informações dos documentos, mas ter uma noção de sua relevância.

As estratégias da RI para atribuir uma medida de similaridade entre uma consulta e um documento são baseadas no senso comum de quão mais frequente um termo é encontrado em um documento e em uma consulta, mais relevante é o documento para consulta. Uma estratégia de recuperação consiste em um algoritmo que processa uma consulta Q e um conjunto de documentos D1, D2, ..., Dn e identifica o coeficiente de similaridade para cada um dos documentos 1 <= i <= n. As estratégias de recuperação de informações são: Modelo de Vetor Espacial, Recuperação Probabilística, Modelos de Linguagem, Redes de Inferência, Indexação Booleana, Indexação Semântica Latente, Redes Neurais e Algoritmos Genéticos e Recuperação de Conjuntos Fuzzy [Grossman e Frieder 2004].

O sistema de RI consiste em obter a coleção de documentos - prontos ou coletados na web. Depois disso, o sistema de RI rapidamente armazena e indexa a coleção de documentos, para recuperar e classificar. A estrutura de índice mais conhecida é a lista invertida que consiste em todas as palavras da coleção. Cada palavra tem uma lista dos documentos em que está presente. Após o processo de indexação, o processo de recuperação é iniciado, procurando documentos que atendam à consulta do usuário.

Posteriormente, a consulta é analisada sintaticamente e expandida com formas variantes de palavras de consulta. A consulta expandida usa um índice para recuperar um subconjunto dos documentos. Esses documentos recuperados são classificados e os que estão no topo da classificação são retornados ao usuário. O processo de classificação consiste em identificar os documentos com maior probabilidade de serem relevantes para o usuário. E considerando a subjetividade inerente à escolha do critério de classificação, torna-se necessário um processo de avaliação para analisar a qualidade dos resultados produzidos pelo sistema de RI.

1.2.2. Lógica de Recomendação

Algoritmos de recomendação analisam em um espaço muito vasto e denso de dados o que os usuários desejam, para isso os algoritmos levam em consideração a análise crítica com base no perfil dos usuários, padrões comportamentais, itens visualizados e mais uma série de outros fatores que fazem o algoritmo entender quais resultados da busca são em teoria mais relevantes para quem está realizando a mesma. Os resultados muitas vezes não são completamente precisos, mas trazem dentre os primeiros itens da lista o que provavelmente o usuário deseja. A cada ano tais algoritmos são aperfeiçoados de modo que a acurácia se mantém cada vez mais elevada. Na Figura 1.1 temos um exemplo simples de como um sistema de recomendação funciona na prática.

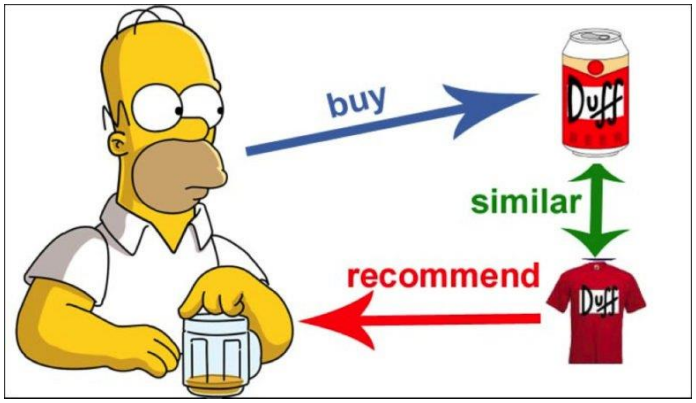


Figura 1.1. Exemplo do funcionamento da recomendação

Após o usuário, que na Figura 1.1 é representado pelo personagem de desenho animado Homer Simpson, comprar uma cerveja da marca Duff o algoritmo de recomendação compara com outros produtos que dispõe em sua base de dados e percebe que existe um outro produto que pode ser interessante ao usuário, logo, este define uma similaridade entre a bebida e a camisa Duff. Ao ter um alto valor de similaridade entre dois itens de acordo com o que o algoritmo estima que o usuário terá interesse, ele recomenda a camisa com a marca da cerveja. Nesse contexto, as recomendações buscam sempre o resultado mais assertivo, mas um algoritmo não necessariamente conseguirá prever todos os casos possíveis. Se Homer tivesse comprado uma camisa Duff naquele dia, o algoritmo teria mostrado uma recomendação muito boa, porém o personagem não compraria a mesma tendo em vista que já tem uma daquele modelo. O algoritmo acertou no gosto do usuário, mas errou na hora da exibição já que não tinha em sua base de dados como prever que o usuário já havia obtido aquele produto anteriormente. Assim funcionam diversos tipos de algoritmos de recomendação. Há sempre a possibilidade de encontrarem dificuldades como essa no processo. No geral, tais algoritmos são bastante satisfatórios já que poupam horas de pesquisas e navegação em páginas, mas ainda existem desafios em abertos para contemplarem todos os cenários possíveis.

Esses desafios envolvem uma dificuldade de se ter acesso às informações que apoiam decisões corretas em um grande conjunto de dados, como os de empresas de vendas. As técnicas de RI auxiliam nesta árdua tarefa. O primeiro aspecto a ser observado é identificar o que um usuário está buscando. Basicamente um usuário menciona um grupo de palavras e/ou símbolos. Em seguida, os documentos/páginas/registros identificados como os mais relevantes são recomendados por algoritmos de busca e recuperação. O princípio dos sistemas de recomendação é baseado em relevância, logo, "o que é relevante para mim também pode ser relevante para alguém com interesses e gostos similares". A grande maioria das técnicas de recomendação trabalham com as avaliações realizadas pelos indivíduos, assim, é possível identificar usuários que avaliaram itens (com notas, por exemplo) de forma similar, passando a entender, portanto, os gostos coletivos. Nesse contexto, o algoritmo consegue criar médias para as pessoas que tem gosto similares (Figura 1.3). Quando um novo indivíduo que se assemelha a um determinado grupo entra em cena, o algoritmo o sugere algo que os outros pertencentes do grupo demonstram interesse. Assim, as chances de acerto são potencializadas partindo da premissa que os gostos do grupo são próximos.

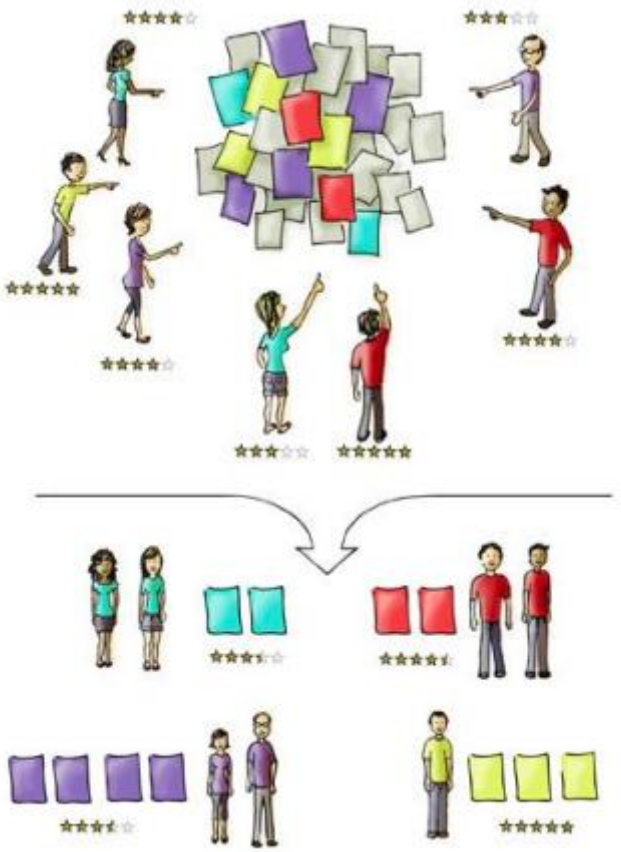


Figura 1.3. Média por grupos similares

Fonte: [Motta et al. 2011]

1.3. Exemplos de Uso

Sistemas de recomendação são amplamente utilizados por grande parte dos sites que nos rodeiam. Quando lemos notícias online, de alguma forma os sistemas de recomendação estão presentes. Quando acessamos as redes sociais esses sistemas podem existir sugerindo publicações que julgam mais relevantes para os usuários. O mesmo pode ocorrer em lojas virtuais, com a presença de sistemas de recomendação exibindo produtos que estão no raio de interesse maior dos clientes. Nesta seção mostraremos dois exemplos de algoritmos de recomendação aplicados.

1.3.1. Portal de Notícias G1

Sites de notícias são excelentes exemplos da aplicação de recomendação. Desde o início da democratização da internet as notícias são disseminadas em sites ou veículos similares. Exibir aos leitores conteúdos que são relevantes, tornam maior a permanência dos mesmos na página. Nesse sentido, a fidelização e as receitas aumentam, tendo em vista que quanto mais tempo um usuário navegar pelo site mais anúncios intercalados com as matérias normalmente serão exibidos. Nada impede de que junto às notícias sejam exibidas recomendações de outras notícias, eventos ou qualquer conteúdo que o

algoritmo da página julgar interessante para o leitor. Um grande veículo de comunicação jornalística no Brasil é o portal de notícias G1[1].

No caso do G1, é possível encontrar recomendações ao navegar por uma notícia, por exemplo. Ao final de cada página da notícia, existe uma recomendação para que o leitor siga lendo uma outra notícia que o algoritmo de recomendação presente no site entende como relevante para aquele leitor naquele momento. Temos tal exemplo demonstrado na Figura 1.4.

Figura 1.4. Exemplo de recomendação de notícias

Fonte: Captura de tela realizada pelas pessoas autoras no dia 31 de outubro de 2019.

1.3.2. Netflix

Seguindo com os exemplos, também é de suma importância citar a Netflix como uma plataforma que faz uso de sistemas de recomendação.A Figura 1.5 mostra que na tela principal de usuários da Netflix há uma seção explícita de recomendação chamada "Sugestões para você", com diversas séries exibidas para um usuário x.

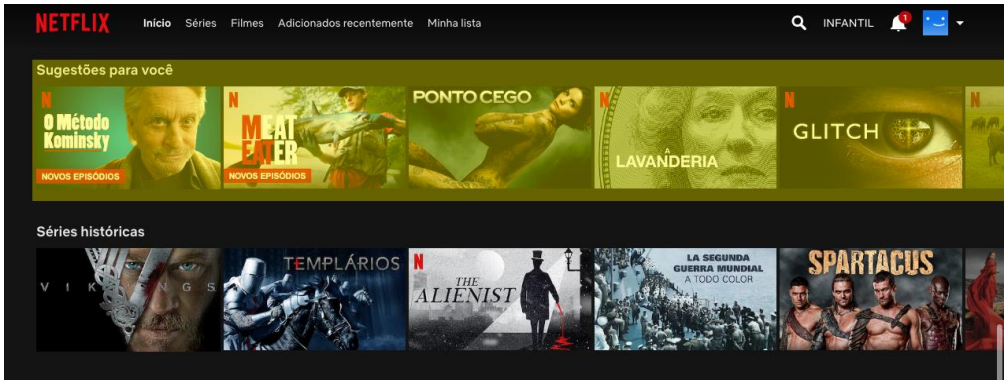


Figura 1.5. Sugestões personalizadas para usuário x da Netflix

Fonte: Captura de tela realizada pelas pessoas autoras no dia 31 de outubro de 2019.

[1] https://g1.globo.com

Atualmente, a Netflix é uma das maiores plataformas para quem deseja assistir filmes, séries, documentários e similares. Seu enorme sucesso se deu pela habilidade de manter seus usuários engajados com o conteúdo que oferecem. Isso se deu também com a alta precisão do algoritmo de recomendação que construíram, que permite sugerir o que um usuário talvez goste para assistir posteriormente. É possível notar na Figura 1.6 que quando mudamos para um usuário y na plataforma, automaticamente outras séries completamente diferentes são inseridas na lista de sugestões visando oferecer ao usuário um conteúdo mais próximo de seu perfil. Portanto, essas séries são completamente distintas das séries mostradas na Figura 1.5 para o usuário x.

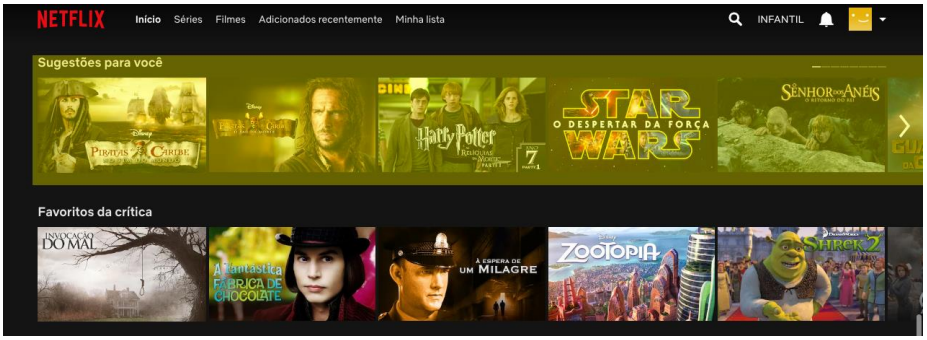


Figura 1.6. Sugestões personalizadas para usuário y da Netflix

Fonte: Captura de tela realizada pelas pessoas autoras no dia 31 de outubro de 2019.

É interessante destacar que tais algoritmos utilizam de várias estratégias para definir qual filme ou série devem recomendar. No caso da Netflix, utilizam-se dos dados de navegação do usuário como: palavras inseridas na busca, filmes clicados, filmes assistidos por completo anteriormente, filmes "abandonados" pela metade, conteúdos curtidos pelo usuário, conteúdos adicionados na lista de preferências e mais uma série de dados que coletam ao longo da permanência do usuário na plataforma.

A Netflix é uma das plataformas que exibe ao usuário o coeficiente de relevância empregado pelo seu algoritmo. Sempre que abrimos uma série ou filme, por exemplo, a plataforma exibe um índice de relevância deste item para o usuário. Como exemplo, a Figura 1.7 exibe o resumo do filme "Piratas do Caribe no Fim do Mundo". Além da descrição, a plataforma mostra para o usuário x que esse filme é 82% relevante de acordo com os dados por eles analisados.

Figura 1.7. Relevância de 82% para o usuário x da Netflix

Fonte: Captura de tela realizada pelas pessoas autoras no dia 31 de outubro de 2019.

Já para o usuário y (Figura 1.8), a plataforma não mensura a relevância tendo em vista que aquele filme, em teoria, não seria do gosto do usuário de acordo com seu perfil.

Figura 1.8. Relevância não mensurada para o usuário y da Netflix

Fonte: Captura de tela realizada pelas pessoas autoras no dia 31 de outubro de 2019.

Com isso, tem-se que a maior parte do catálogo sugerido e resultados de busca utilizam técnicas de recomendação por similaridade de conteúdo. Ao pesquisar por "Piratas do Caribe", são retornados abaixo dos resultados esperados uma série de outros filmes que não são necessariamente ligados ao tema do filme, mas que trazem diversas similaridades como: atores, conteúdo da sinopse, filmes que pessoas que pesquisaram por "Piratas do Caribe" assistiram após recomendação do algoritmo, e mais uma série de fatores. A Figura 1.9 mostra as recomendações que vieram após a busca do usuário x por "Piratas do Caribe" no fim do mundo.

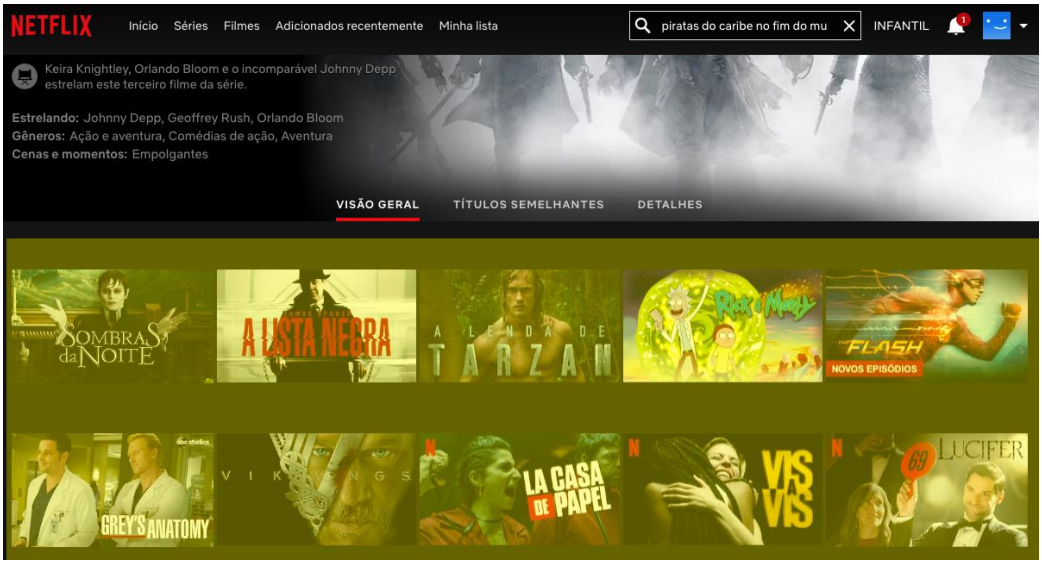


Figura 1.9. Sugestões após a busca por "Piratas do Caribe no Fim do Mundo" na Netflix

Fonte: Captura de tela realizada pelas pessoas autoras no dia 31 de outubro de 2019.

1.4. Tipos de Algoritmos de Recomendação

Segundo Herlocker et al. (2000), por muitos anos os cientistas têm direcionado esforços para buscar soluções e meios de amenizar o problema ocasionado com a sobrecarga de informações geradas através por soluções que reconhecem e categorizam informações automaticamente. Recomendações não são triviais para serem realizadas. Tanto no mundo físico quando no mundo digital é preciso observar uma série de fatores que influenciam um tipo de recomendação. Oferecer comida de cachorro para as pessoas degustarem em um supermercado não é algo interessante, na verdade, seria um grande fracasso. Mas oferecer café por exemplo pode ser uma ótima opção. É possível perceber o quão falha ou bem-sucedida é essa recomendação de "O que oferecer para degustação humana", mas para as máquinas não é. Café e ração são produtos como quaisquer outros. É preciso ensinar para a máquina o sentido das coisas para que ela não sugira um pneu para degustação, por exemplo. Ou, que ela não sugira um desenho animado infantil para uma pessoa que prefira filmes de terror.

Ao longo dos anos foram desenvolvidas técnicas por pesquisadores que refinam os resultados das buscas utilizando diversos tipos de filtros e parâmetros. Tais filtragens são realizadas de acordo com as características de um item, assim como o comportamento dos usuários/consumidores perante aquele item. Logo começaram a entender padrões que poderiam ser utilizados para melhorar as recomendações e torná-las realmente úteis e de alta relevância para quem a recebesse. Para isso, criaram-se técnicas para buscar e recuperar informações, filtrar itens com base nas descrições dos conteúdos e também analisar o comportamento dos usuários perante o item. Assim, foi possível construir um modelo colaborativo para entender melhor como é a interação dos usuários com os itens e gerar modelos estatísticos baseados em grupos similares.

Nos sistemas de recomendação são utilizadas em geral uma das três técnicas de filtragem de informação citadas a seguir: filtragem baseada em conteúdo, filtragem colaborativa, também conhecida como filtragem social e filtragem híbrida [Cazella, Sílvio César et al. 2010].

1.4.1. Filtragem Baseada em Conteúdo

Alguns softwares têm como objetivo gerar de forma automática descrições dos conteúdos dos itens e comparar estas descrições com os interesses dos usuários, verificando assim se o item é ou não relevante [Balabanović e Shoham 1997].

A filtragem baseada em conteúdo (Figura 1.10) consiste em realizar análises comparativas envolvendo o conteúdo de itens distintos, como por exemplo, palavras similares, temas similares, gêneros similares e afins. A análise de diversos componentes do conteúdo de um item torna possível identificar a similaridade entre dois ou mais itens através de um coeficiente numérico. A recomendação baseada na filtragem por conteúdo utiliza informações anteriores do usuário em relação a um item para recomendar itens similares. São recomendados, por exemplo, itens que mais se aproximam aos outros itens avaliados de forma positiva por determinado usuário.

A indexação da frequência de termos, ou seja, o quanto determinados termos se repetem no conteúdo de um item é bastante utilizada nessa abordagem, sendo as informações dos documentos e necessidades dos usuários descritas por vetores que armazenam a frequência com que as palavras ocorrem em um dado documento (como

em [Campos et al. 2019]) ou em uma consulta realizada pelo usuário [Cazella, Sílvio César et al. 2010].

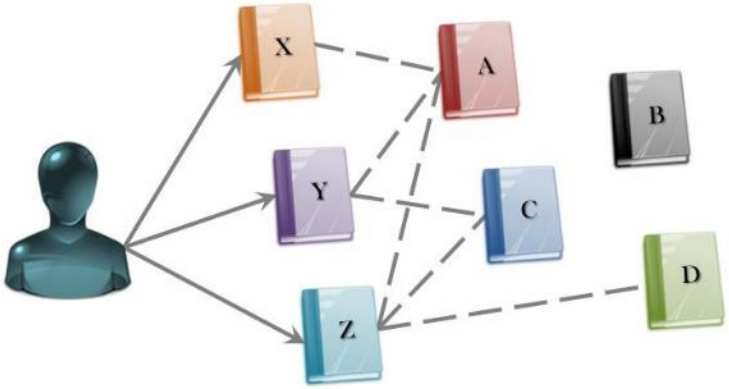


Figura 1.10. Representação da abordagem de filtragem por conteúdo

Fonte: [Costa et al. 2013]

1.4.2. Filtragem Colaborativa

A diferença da abordagem entre a filtragem colaborativa para a baseada em conteúdo está no fato de que, enquanto a filtragem colaborativa considera as preferências de vários usuários analisadas em seu perfil, a baseada em conteúdo considera padrões semelhantes relacionados às experiências apenas do usuário interessado [Aggarwal 2016]. Dado o item i e usuário u, a abordagem de filtragem colaborativa tem como princípio o fato de que, se um usuário u avalia itens semelhantes a outro usuário u', há uma grande chance de a próxima avaliação de um usuário u para um novo item seja semelhante ao do usuário u'. Portanto, se outros usuários avaliaram dois itens de maneira semelhante, existe uma tendência do usuário u avaliá-los da mesma maneira. A abordagem de filtragem colaborativa pode ser dividida em duas etapas; a abordagem em vizinhança e a baseada em modelos [Desrosiers e Karypis 2011]. O primeiro pode ser realizado de duas maneiras:

- Baseado em conteúdo: considere o exemplo com os usuários u e u'. A semelhança entre esses usuários os torna vizinhos. Sempre que o sistema avalia o interesse de u pelo item i, ele verifica a classificação dada pelos vizinhos [Desrosiers e Karypis 2011];
- Baseado em itens: nessa abordagem, o sistema consulta outros itens semelhantes ao item i. Dados esses itens semelhantes, toda vez que o sistema avalia o interesse de u pelo item i, ele considera as avaliações feitas pelo usuário u para esses itens semelhantes. Para definir se dois itens são semelhantes, o sistema verifica se muitos usuários classificaram esses itens da mesma maneira [Desrosiers e Karypis 2011]. É também chamado de abordagem item a item [Koren 2008].

Embora as possibilidades da abordagem de vizinhança sejam baseadas em item ou usuário, o modelo propõe mesclar itens e usuários no mesmo espaço de fatores latentes, permitindo inferir automaticamente com base no feedback do usuário [Koren 2008]. Apesar de ter várias técnicas nessa abordagem [Desrosiers e Karypis 2011], existem algumas muito comuns, como *clustering*, classificação, modelo latente, Markov Decision Process (MDP) e Matrix Factorization (MF) [Do et al. 2010].

A técnica para descobrir de forma automática as relações entre um dado usuário e seus "vizinhos mais próximos" consiste em (1) calcular a similaridade do usuário alvo em relação aos outros usuários; (2) selecionar um grupo de usuários com maiores similaridades para considerar na predição; e (3) normalizar as avaliações computando as predições, ponderando as avaliações dos usuários mais similares [Cazella, Silvio César et al. 2010]. Seguindo a lógica do exemplo de recomendação representado pela Figura 1.10, a Figura 1.11 mostra de forma representativa como é uma recomendação no modelo de filtragem colaborativa.

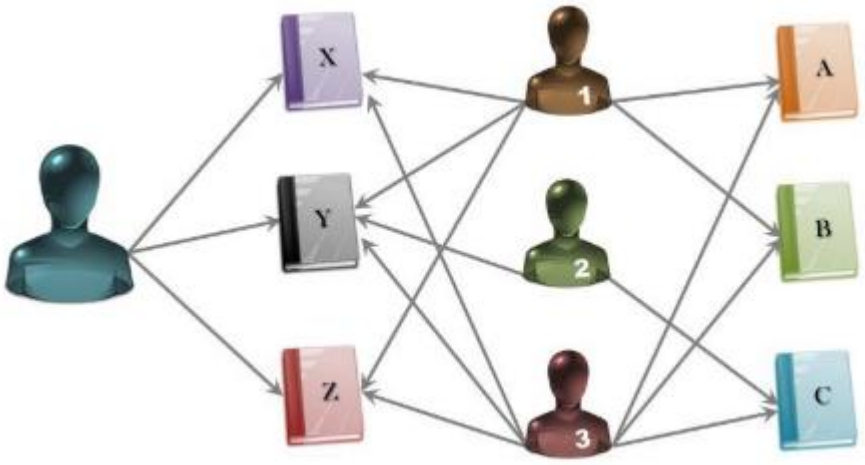


Figura 1.11. Representação da abordagem de filtragem colaborativa

Fonte: [COSTA et al, 2013]

No exemplo da Figura 1.11, um dado usuário alvo gostou dos livros X, Y e Z. É possível notar que seus vizinhos mais próximos são os usuários 1, 2 e 3, tendo em vista que também leram os mesmos livros X, Y, e Z, logo, provavelmente esses usuários têm gostos similares. Portanto, o sistema pode recomendar os livros A, B e C para o usuário alvo, tendo em vista que estes livros foram consumidos pelos vizinhos do usuário alvo.

1.5. Dados Privados em Sistemas de Recomendação

Com o gigante volume de dados armazenados sobre as preferências, itens visualizados, itens comprados, itens ignorados e mais uma gama de dados sensíveis que os sistemas armazenam para refinar as recomendações dos usuários, existe uma preocupação vinculada à segurança. Isso inclui questionamentos sobre como esses dados são tratados e utilizados por estes sistemas. Dados dos quais se armazenados ou utilizados de forma não correta podem expor centenas de milhares de usuários.

A privacidade de dados em algoritmos de recomendação pode incluir o princípio de Privacidade Diferencial. A partir desse recurso matemático, é possível atribuir aos algoritmos de recomendação um valor do quanto esse algoritmo preserva a privacidade dos usuários. Uma dificuldade existente para a obtenção de privacidade é que a privacidade diferencial garante a privacidade de usuários, mas pode diminuir a utilidade do algoritmo de recomendação [Jiang et al. 2019]. Essa teoria de privacidade diferencial começou a ser implementada em sistemas de recomendação de filtragem colaborativa muito recentemente com a proposta de Mcsherry e Mironov (2009), fazendo parte da categoria de "*data transformation technology*", segundo Feng et al. (2018).

Outra categoria descrita em [Feng et al. 2018] para privacidade de dados em sistemas de recomendação é a de recomendação distribuída [Qi et al. 2017]. As soluções dessa categoria podem envolver processos para ocultar informações reais de usuários, ou aplicar dimensões de qualidade através da "*List Scheduling Heuristic*" (LSH). Existe também a possibilidade do uso de encriptação homomórfica. Armknecht e Strufe (2011) sugerem utilizar essa técnica para atender as necessidades de construir um recomendador que não necessite de informações de preferências de usuários, cujo provedor de recomendação não precise revelar informações internas, mas que seja possível retornar uma recomendação correta para o usuário.

Os aspectos de privacidade existem até mesmo em cenários que envolvam dados abertos, exigindo uma atenção por parte dos sistemas de recomendação. Um exemplo disso pode ser observado nos Ecossistemas MOOCs [Campos et al. 2018] (entende-se que MOOC, por definição traduzida, significa "Cursos Online Abertos e Massivo"). Nesse cenário, apesar dos dados de cursos dos provedores e dados dos alunos matriculados serem abertos, há restrições que exigem uma camada de autorização por parte dos recomendadores interessados em extrair esses dados. Essa camada deve permitir que a recomendação envolvendo dados pessoais de um determinado aluno nos MOOCs apenas seja processada caso esse aluno se autentique e autorize a extração dos seus dados.

Apesar das soluções existentes, há uma necessidade eminente de novas abordagens que considerem os aspectos de dados privados, principalmente em sistemas de recomendação baseados em conteúdo. Essa abordagem utiliza apenas dados do usuário interessado em receber a recomendação. Na existência de restrições de acesso a esses dados, a recomendação se tornaria ainda mais limitada, podendo comprometer inclusive o resultado final dos itens retornados. A solução para esse problema pode partir, por exemplo, da integração dos dados do usuário interessado com outras bases de dados (como com dados das redes sociais) e, em seguida, inferência de um perfil do usuário. Essas estratégias permitem, inclusive, a assertividade de um algoritmo de recomendação no caso do *cold-start*, ou seja, usuários ou itens que são novos nas plataformas, não possuindo dados suficientes para uma primeira recomendação.

Um enfoque que os sistemas de recomendação baseados em conteúdo podem ter é na recomendação de publicidade direcionada ou na entrega de cupom direcionada. Para esses casos, Wang et al. (2018) analisam trabalhos que buscam criar mecanismos de proteção e privacidade para usuários. Através desse levantamento é possível identificar, inclusive, a importância da proteção de dados para quem fornece a recomendação. Wang et al. (2018) mencionam que no caso de entrega de cupom direcionada, os cupons não podem ser utilizados por outros usuários além do usuário alvo da recomendação. Acrescentam ainda que, além das técnicas de criptografia, mencionada anteriormente, outras soluções possíveis para uma camada de privacidade nessa abordagem de recomendação seria anonimização, teoria de games, ofuscação ou segmentação local.

1.6. Operações com Python

Essa seção inclui elementos técnicos para compreender algumas operações básicas com a linguagem Python e como elas são utilizadas para lidar com dados, cálculos e exibição de resultados. Esses conhecimentos são fundamentais para a compreensão do algoritmo

de recomendação sugerido na Seção 1.7. Portanto, são introduzidos conceitos como: comentários, exibição de valores, variáveis e tipos de dados, operações aritméticas, operações lógicas, controle de fluxo e funções. Vale ressaltar que não são abordadas informações sobre a instalação local do Python ou configuração de ambiente de desenvolvimento tendo em vista que todo o desenvolvimento e exemplos aplicados neste capítulo são realizados na plataforma do Google Colab[2]. Essa plataforma dispõe de um ambiente previamente configurado e pronto para qualquer usuário. Todos os exemplos empregados nessa seção foram testados e desenvolvidos com o Python na versão 3.6, por ser uma das versões mais recentes e estáveis disponíveis no momento em que este capítulo foi desenvolvido.

1.6.1. Comentários

A linguagem Python assim como a grande maioria das linguagens de programação dispõe de comentários para auxiliar na organização do código. Ao comentar uma linha ela é ignorada pelo interpretador da linguagem quando o código for executado, assim, servindo apenas para programadores lerem e entenderem o que foi comunicado naquele código. Neste exemplo são abordadas as maneiras de se utilizar o comentário em linha, por ser o mais utilizado e o que é empregado em nossos exemplos. Vale ressaltar que também existem comentários de blocos, ou seja, é possível comentar várias linhas de uma única vez.

O comentário de linha serve para criar apenas uma linha comentada. No Python os comentários de linha são criados utilizando o caractere “#”. A Figura 1.12 exemplifica a criação de um comentário na primeira linha de um arquivo no Google Colab.

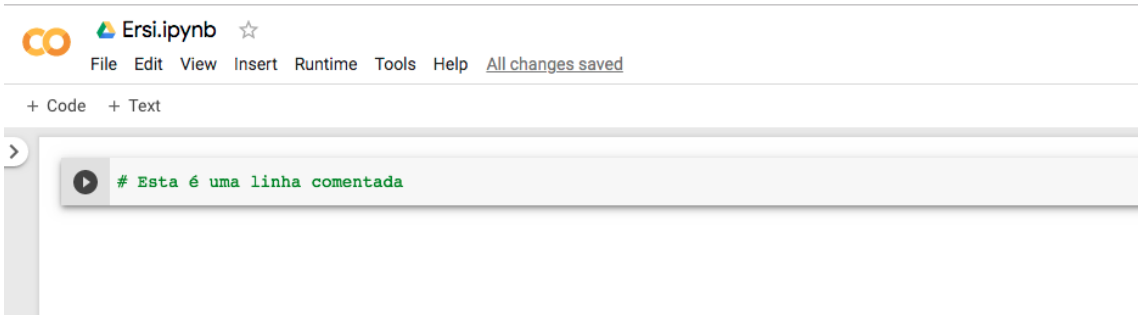


Figura 1.12. Exemplo de comentário de linha no Google Colab

Fonte: Captura de tela realizada pelas pessoas autoras no dia 03 de novembro de 2019.

Normalmente os comentários são utilizados para descrever blocos ou linhas de códigos visando tornar simples a compreensão e a manutenção daquele bloco ou linha por outros programadores. Comentários são amplamente utilizados para documentar funções e cabeçalhos de arquivos. Nos blocos de código das próximas seções, os comentários são utilizados para detalhar passo a passo o que faz cada linha de código escrita, visando auxiliar no entendimento do conteúdo aplicado.

[2] https://colab.research.google.com/

1.6.2. Exibir Textos e Valores

O Python, assim como outras linguagens, dispõe de um comando em especial que serve para exibir valores na tela para que um usuário do sistema ou um programador consiga imprimir valores para o entendimento do que está ocorrendo em determinado momento da execução do algoritmo. Esse recurso é possível através do comando “print”. A Figura 1.13 demostra como utilizar o comando “print” para exibir o texto “Hello World” (que no Python é tratado como um tipo *String*).

```
print("Hello World")
```

Figura 1.13. Código Python para exibição de texto

O comando “print” não se limita à exibição de textos. Também é possível exibir outros tipos de valores, como é abordado nas próximas sessões.

1.6.3. Variáveis e Tipos de Dados

Variáveis são amplamente utilizadas na programação para armazenar valores. Pode-se entendê-las como “caixinhas” que guardam valores na memória do computador. As variáveis no Python têm tipagem dinâmica, ou seja, não é preciso especificar o tipo do valor que estamos atribuindo para ela no ato de declarar a variável. Basicamente, criamos uma variável de tipo genérico e é possível atribuir a ela um texto ou valor numérico, por exemplo. Em linguagem com a tipagem não dinâmica é preciso declarar que determinada variável recebe apenas texto, por exemplo, não sendo possível inserir valores diferentes dos textuais. A Figura 1.14 demonstra como é a atribuição de valores de diferentes tipos a uma variável chamada “a”.

```
a = 1 # Número inteiro
a = 2.65 # Número decimal
a = 1E3 # Notação científica, equivale a 10^3 = 1000.0
a = "Escola regional de sistemas 2019" # Texto (string atribuída com aspas duplas)
a = 'ERSI2019' # Texto (string atribuída com aspas simples)
a = True # Tipo lógico [booleano] (verdadeiro -> True, falso -> False)
a = [1, 2, 3] # Lista de valores numéricos
a = {"a": 1, "b": 2, "c": {"x": 10}} # Dicionário de dados (pares de chave e valor)
```

Figura 1.14. Código Python para exemplificar declaração e exibição de variáveis

No Python o próprio interpretador testa e verifica o tipo da variável para tratá-la da melhor e mais performática forma possível. Na criação de uma variável, é possível declará-la com valores inteiros, decimais, textuais, listas, dicionários e afins e isso torna bem mais simples o uso de variáveis se comparado com as linguagens que exigem tipos fixos. Utilizando as variáveis, é possível realizar as mais diversas operações matemáticas, assim como receber objetos maiores como listas e valores provenientes de arquivos ou banco de dados.

1.6.4. Operações Aritméticas

As operações aritméticas são importantes para diversas linguagens de programação. Com tais operações, é possível calcular praticamente todo tipo de conta matemática básica. Pode-se entender como operadores básicos aritméticos os operadores de soma, subtração, multiplicação e divisão. Pode-se também estender os operadores aritméticos para os operadores um pouco mais avançados como os de exponenciação e cálculo de parte inteira. A Figura 1.15 exemplifica a realização das quatro operações básicas (soma, subtração, multiplicação e divisão) em Python e também como atribuir valores das operações em variáveis para posteriormente exibi-los através do comando "print". No exemplo da Figura 1.15 também é demonstrado como o comando "format" pode auxiliar na apresentação de resultados de uma forma mais agradável e simples de entender visualmente, mesclando texto e valores numéricos.

```
a = 5
b = 2
soma = a + b
# O método format serve para criar uma string que contem campos entre chaves que são substituídos pelos valores que estão dentro do método. No exemplo abaixo, o ".format()" recebe a variável soma
print("Soma: {}".format(soma))
subtracao = a - b
print("Subtração: {}".format(subtracao))
multiplicacao = a * b
print("Multiplicação: {}".format(multiplicacao))
divisao = a / b
print("Divisão: {}".format(divisao))
```

Figura 1.15. Código Python para exemplificar operações aritméticas básicas

Como mencionado anteriormente, os operadores aritméticos não se resumem aos básicos das quatro operações. Também existem mais alguns, dentre os quais ressaltamos os de exponenciação e de parte inteira. A Figura 1.16 ilustra como utilizar tais operadores.

```
a = 5
b = 2
exponenciacao = a**b
print("Valor exponencial: {}".format(exponenciacao))
parte_inteira = a // b
print("Parte inteira: {}".format(parte_inteira))
```

Figura 1.16. Código Python para exemplificar operações aritméticas avançadas

1.6.5. Operações de Comparação

As operações de comparação retornam resultados booleanos, ou seja, *True* (para verdadeiro) ou *False* (para falso). Os comparadores permitem testar se um valor x é igual a um valor y, por exemplo. A Figura 1.17 ilustra como utilizar os operadores de comparação do Python.

```
a = 10
b = 5

# Operador de maior que >
maior_que = a > b
print("{} é maior que {}? {}".format(a, b, maior_que))

# Operador de menor que <
menor_que = a < b
print("{} é menor que {}? {}".format(a, b, menor_que))

# Operador de maior ou igual >=
maior_igual = a >= b
print("{} é maior ou igual a {}? {}".format(a, b, maior_igual))

# Operador de menor ou igual >=
menor_igual = a <= b
print("{} é menor ou igual a {}? {}".format(a, b, menor_igual))

# Operador de igualdade ==
igual_a = a == b
print("{} é igual a {}? {}".format(a,b, igual_a))

# Operador de diferença !=
diferente_de = a != b
print("{} é diferente de {}? {}".format(a,b, diferente_de))
```

Figura 1.17. Código Python para exemplificar operações de comparação

É possível observar na Figura 1.17 que além de verificar igualdade entre dois valores, os comparadores permitem ainda testar se são “diferentes”, “maior que”, “menor que”, “maior ou igual a” e “menor ou igual a”. Com os operadores de

comparação é possível saber se a nota dada por um determinado usuário a um item é maior que a nota que esse mesmo usuário deu para um outro item, ou ainda se a nota de ambos é igual. Diversas combinações podem ser realizadas por sistemas de recomendação para analisar se dado item é mais ou menos similar a outros para que seja criada uma recomendação. Assim, os operadores de comparação cooperam ativamente para os sistemas de recomendação.

1.6.6. Operações Lógicas

Os operadores lógicos unem expressões lógicas formando assim, uma nova expressão que é composta por duas ou mais subexpressões. O resultado lógico (verdadeiro ou falso) de expressões compostas é a relação entre as subexpressões resultando em uma única resposta. A Figura 1.18 demonstra as diferenças no uso dos operadores “and” e “or”.

```
# Operadores and e or


print("Operador and")
print(True and False)
print(True and True)
print(False and True)
print(False and False)


print("\n\nOperador or")
print(True or False)
print(True or True)
print(False or True)
print(False or False)
```

Figura 1.18. Código Python para exemplificar operações lógicas

O Python implementa os valores lógicos utilizando a lógica *booleana*, implementando os valores *True* ou *False*. Ambos precisam necessariamente ser informados com a primeira letra maiúscula. Ao analisar a Figura 1.18 é possível perceber que sempre que existie pelo menos um valor *False*, o operador “and” logo retorna que a expressão é falsa. Já o operador “or” verifica se pelo menos um dos dois valores é positivo. Caso seja, o Python interpreta que a expressão pode ser considerada verdadeira.

Os valores lógicos são de fato a base de toda a computação, até porque, temos que, o bit está ligado ou o bit está desligado. O valor lógico em sua forma mais primitiva assume o número 1 quando for um valor verdadeiro e o valor 0 quando for um valor falso. Toda expressão avaliada na computação de maneira geral, resulta em um valor lógico, isto é, ou a expressão é verdadeira ou falsa.

Além disso, é possível combinar expressões comparativas utilizando operadores booleanos, assim, buscando resultados mais complexos que posteriormente podem ser utilizados em condicionais, por exemplo. A Figura 1.19 demonstra como combinar comparadores de igualdade com os operadores “and” e “or” buscando um resultado único booleano por expressão.

```
# Operações avançadas com operadores lógicos
a = 1
b = 2
c = 3
d = 3
ab = a == b
cd = c == d
print(ab and cd)
print(ab or cd)
```

Figura 1.19. Código Python para exemplificar Exemplo de operações relacionais e lógica combinada

1.6.7. Controle de Fluxo

O controle de fluxo é uma parte importantíssima das linguagens de programação, uma vez que dão total autonomia ao desenvolvedor para definir se algo deve ou não ser realizado. O controle de fluxo só é possível devido às operações lógicas que tornam possível realizar testes de verdadeiro ou falso para dadas expressões. Os conceitos de controle de fluxo podem ser divididos em duas partes: o controle de fluxo através de condicionais e através de laços de repetição. Sobre o controle de fluxo através de condicionais, a Figura 1.20 demonstra como desenvolver um código para tomar certas decisões dependendo de condições prévias.

```
# Utilização de condicionais (if, else, elif)
a = 2
b = False


if b == True:
  print("Eba, o B é positivo!")
elif a > 0:
  print("Eba, o A é maior do que 0")
else:
  print("Poxa... Nenhuma das condições foi verdadeira")
```

Figura 1.20. Código Python para exemplificar o uso de condicionais

Uma das capacidades de uma linguagem de programação é a de decidir o fluxo por onde os comandos são executados. O controle condicional é necessário em praticamente todo programa de computador, sendo basicamente a capacidade de tomar decisões com base em certas condições.

A linguagem Python tem algumas palavras reservadas que são dedicadas especialmente para criar fluxos de execuções condicionais, como: “if”, “else” e “elif”. Basicamente, os condicionais nos permitem selecionar certos blocos de código que são ou não executados dependendo da condição inserida para ser analisada.

Na Figura 1.20, o comando “if” verifica se a primeira condição é válida, ou seja, se a variável “b” possui o valor *True*. Como “b” tem o valor False, a condição foi considerada falsa. O interpretador então pula para o próximo comando, sendo esse o “elif”. A condição do “elif” é verdadeira, logo, o interpretador exibe a mensagem "Eba, o A é maior do que 0". Caso a variável “a” não fosse maior do que 0, o interpretador Python pularia para a próxima instrução que seria o “else”, e consequentemente exibiria o texto "Poxa... Nenhuma das condições foi verdadeira". O comando “else” sempre é executado quando todas as condições acima não forem verdadeiras. O comando “elif” do Python tem a mesma função do comando “elseif” em outras linguagens como o PHP e JavaScript, por exemplo.

Já o controle de fluxo através dos laços de repetição, é a capacidade de repetição de comandos até que uma dada condição seja satisfeita. Existem vários tipos de laços (*loops*) em Python. O mais comum e amplamente utilizado é o “for”. Normalmente o “for” é utilizado com objetos iteráveis, tais como listas e intervalos. Para simplificar, são utilizados dois exemplos: um com uma lista e outro com um intervalo. A Figura 1.21 demonstra um *loop* “for” percorrendo cada um dos elementos de uma lista. A condição de parada é o final da lista.

```
# Exemplo de utilização do laço de repetição for com lista
for i in [0, 1, 'oi', 3, 4.5, 'batata']:
  print(i)
```

Figura 1.21. Código Python para exemplificar um *loop* percorrendo uma lista

É interessante notar que não há diferença se a lista é de um determinado tipo único (como uma lista exclusivamente de número inteiros), ou se a lista tem valores mistos (como inteiros, *strings* e decimais). No exemplo da Figura 1.21, o *loop* percorre item por item passando o valor do item atual para a variável “i”. Já na Figura 1.22, pode-se notar um exemplo de utilização de laço de repetição “for” com um intervalo.

```
# Exemplo de utilização do laço de repetição for com intervalo
for i in range(5):
  print(i)
```

Figura 1.22. Código Python para exemplificar o uso de laços de repetição com intervalo

Os *loops* podem possibilitar, por exemplo, que os sistemas de recomendação percorram os valores advindos da base de dados para obter insumos, podendo sugerir

itens aos usuários. Sem a navegação dos itens realizada pelos laços seria pouco provável a viabilidade de implementar um bom algoritmo de recomendação.

1.6.8. Funções

Funções são blocos de código que realizam determinadas tarefas que normalmente precisam ser executadas várias vezes dentro de uma aplicação. Quando surge a necessidade de executar diversas vezes o mesmo código, criamos funções para que estas instruções não precisem ser repetidas atrapalhando a manutenção do código e causando dualidades. Assim, é possível "empacotar" trechos de códigos e utilizá-los em diversos lugares diferentes.

Uma função pode ser criada no Python utilizando a palavra reservada "def" seguida do nome da função e parâmetros desejados (caso existam). Após isso, deve ser criado o corpo da função que pode dispor de um retorno de valor ou não. A Figura 1.23 demonstra como criar e executar uma função sem retorno. Já a Figura 1.24 demonstra como criar e utilizar uma função genérica que retorna o salário semanal de um indivíduo e que pode ser utilizada em diversas outras partes de um programa.

```
def oi(nome):
   print('Olá',nome)
oi("Maria")
```

Figura 1.23. Código Python para exemplificar uma função simples

```
# Esta função calcula se a carga horária semanal passou de 40 horas e calcula o valor das horas extras. O resultado é o retorno do quanto o indivíduo deve receber.
def calcular_salario_semana(qtd_horas, valor_hora):
  horas = float(qtd_horas)
  taxa = float(valor_hora)
  if horas <= 40:
    salario=horas*taxa
  else:
    h_excd = horas - 40
    salario = 40*taxa+(h_excd*(1.5*taxa))
  return salario
mó
# 176 horas trabalhadas
salario = calcular_salario_semana(176, 10)
print("O valor que a pessoa precisa receber é: {}".format(salario))
```

Figura 1.24. Código Python para exemplificar uma função com retorno

As funções sem retorno, como a da Figura 1.23, não são muito usuais. Normalmente é esperado que, após o processamento realizado por uma função, seja retornado um resultado para que o fluxo continue a partir daquele valor (realizando outras operações ou exibindo algo ao usuário), assim como a função da Figura 1.24. É recomendado que as funções sejam mais genéricas possíveis para que o código possa ser reaproveitado ao máximo de vezes sem exigir alterações particulares para cada caso. Nesse contexto, é preciso criar algo genérico o suficiente para ser útil em várias partes do código, mas ao mesmo tempo com um grau de complexidade acessível que dê para realizar manutenção sem problemas maiores. Muitas funções possibilitam atender a diversos cenários, mas são extremamente complexas para serem entendidas. O mundo ideal é ter funções genéricas e simples de serem entendidas para futuras alterações e manutenções.

A linguagem Python dispõe ainda de diversas funções próprias prontas ao uso do programador. É possível, por exemplo, medir o tamanho de uma lista utilizando a função "len", além de indicar que todos os caracteres de uma *String* são minúsculos utilizando a função "lower". Existe uma gama de funções prontas para cada tipo de trabalho que se deseje realizar partindo desde formatação de texto até a análise de dados matemáticos complexos. A linguagem Python atualmente é uma das mais bem preparadas para atender tanto os programadores com demandas mercadológicas (criar sistemas, telas de autenticação, integrar com bancos de dados etc.) até pesquisadores que necessitam realizar cálculos em terabytes de dados.

1.7. *Hands-on*: Implementando um Algoritmo de Recomendação

A plataforma do Google Colab[3] foi escolhida como ambiente para a implementação de um algoritmo de recomendação simples, visando demonstrar de forma prática os conhecimentos anteriormente abordados nesse capítulo. Tal plataforma foi adotada por ser gratuita, on-line e por dispor de ambiente Python preparado com o necessário para a execução do código que é demonstrado nessa seção.

O Google Colab também conta com outros pontos positivos, como o fato do usuário poder salvar o *notebook* de códigos criados para utilizar e/ou estudar posteriormente. Além disso, também é possível acessá-lo de qualquer dispositivo que tenha acesso com a internet e navegador compatível. Por ser on-line, o ambiente do Google Colab proporciona que não seja necessário criar e configurar ambientes Python complexos para a implementação de algoritmos, simplificando, portanto, a empregabilidade do conhecimento abordado nessa seção e não exigindo a necessidade de conhecimentos sobre infraestrutura e sistemas operacionais para instalar e configurar o ambiente Python.

A presente seção aborda um exemplo de algoritmo de recomendação de filmes. É utilizada a base do MovieLens para exemplificar como um algoritmo de recomendação funciona. O objetivo dessa exemplificação é recomendar filmes que ainda não foram assistidos por um usuário de acordo com as avaliações da base de dados. Nessa base é possível selecionar avaliações de vários usuários u para diversos filmes f. Dessa forma, temos uma matriz u X f. Os filmes são avaliados considerando

[3] colab.research.google.com

uma escala de 1 até 5 (“avaliação 1 estrela” até “avaliação 5 estrelas”). Caso um usuário u não tenha avaliado um filme f, o valor dessa associação é 0.

Considerando que estamos utilizando a filtragem colaborativa como abordagem de recomendação, uma solução para prover recomendações de filmes f para um usuário u é analisar um conjunto de usuários S que são similares ao usuário u e que já assistiram ao filme f. A importação dessa base e a execução desses passos iniciais pode ser feita utilizando a biblioteca Python, Surprise. Através da utilização do Surprise, é possível carregar o dataset do MoviesLens, conforme demonstrado na Figura 1.25.

A partir do “ml-100k” é possível carregar um dataset com 100.000 registros de interação de usuários (linhas) com filmes (colunas) do MovieLens. O método “load_builtin()”, proveniente da biblioteca Surprise, permite a importação de 943 linhas e 1682 colunas deste *dataset*.

```
#instalação e importação do Surprise
#a exclamação antes do comando “pip” deve ser utilizada no caso de uso do Google Colab

!pip install surprise
from surprise import Dataset
from surprise import KNNBasic
#para carregar o dataset do MovieLens
dados = Dataset.load_builtin("ml-100k")
```

Figura 1.25. Código Python para importação da biblioteca Surprise e do *dataset* do MovieLens

A partir dessa matriz, é possível calcular a similaridade entre os filmes, ou seja, a abordagem de filtragem colaborativa chamada de baseado em item (*item-item*). Ao contrário da baseada em usuário (*user-based*), essa abordagem cria para cada filme um vetor de avaliações. Existem algumas implementações Python possíveis para calcular a similaridade entre dois vetores. Uma delas é a “cosine”. A similaridade “cosine” entre dois vetores pode ser facilmente calculada utilizando o “KNNBasic”. Dessa forma, é preciso indicar como propriedade da similaridade que não se trata de um “*user-based*” e de que se deseja implementar a “cosine”, conforme Figura 1.26.

```
opcoes_sim = {
    'name': 'cosine',
    'user_based': False
}
knn = KNNBasic(sim_options=opcoes_sim)
```

Figura 1.26. Código Python para declarar a similaridade cosine utilizando o KNNBasic

É possível definir também um conjunto de dados para treinamento utilizando toda a base carregada na variável “dados”. O “KNNBasic” representado no exemplo da Figura 1.26 pela variável “knn”, pode ser treinado utilizando o método “fit( )”, assim como demonstrado no código Python na Figura 1.27.

```
dados_treinamento = dados.build_full_trainset()
knn.fit(dados_treinamento)
```

Figura 1.27. Código Python para treinar o *dataset* utilizado

O principal beneficiado com as recomendações desse exemplo são aqueles usuários que ainda não assistiram algum filme da base. Para selecionar os pares desses usuários (usuário-filme) no conjunto treinado “trainingSet”, é possível utilizar o método “build_anti_testset()”. Com todos esses registros selecionados, é possível aplicar o método “test()” para se obter predições de avaliações.

Essas predições indicam (com base no conjunto treinado) que um usuário u que tenha avaliado positivamente alguns filmes de animação pode avaliar outros filmes semelhantes de animação com 5 estrelas, pode exemplo. Nesse contexto, há uma grande possibilidade que filmes de animação ainda não assistidos por esse usuário obtenham um índice maior de similaridade. Essa etapa está representada na Figura 1.28.

```
dados_teste = dados_treinamento.build_anti_testset()
predicoes = knn.test(dados_teste)
```

Figura 1.28. Código Python para gerar as predições da recomendação com base nos dados de treinamento

Com essas predições definidas, é possível ranquear e selecionar apenas as 4 predições mais bem pontuadas. Para isso, o método “get_top4_recomendacoes” é criado, conforme Figura 1.29.

```
from collections import defaultdict
 def get_top4_recomendacoes(predicoes, topN = 4):


   top_recomendacoes = defaultdict(list)
   for uid, iid, true_r, est, _ in predicoes:
      top_recomendacoes[uid].append((iid, est))
   for uid, avaliacoes_usuario in top_recomendacoes.items():
      avaliacoes_usuario.sort(key = lambda x: x[1], reverse = True)
      top_recomendacoes[uid] = avaliacoes_usuario[:topN]


   return top_recomendacoes
```

Figura 1.29. Código Python para declarar função de seleção apenas das quatro predições mais bem ranqueadas

Para isso, é preciso indicar que a variável "top_rcomendacoes" (a variável que armazena os resultados do método criado) corresponde a uma "defaultdict(list)", uma maneira mais simples da biblioteca "collections" de inicializar um dicionário. Nesse caso, o dicionário está sendo inicializado com uma coleção (uma lista). O retorno das predições é indicado pelos identificadores (*ids*) dos filmes. Para que seja possível compreender bem qual filme está no top-4 de determinado usuário, é possível utilizar um método[4] da biblioteca Surprise que cria uma relação (dicionário) dos nomes dos filmes e respectivos *ids*, conforme o código da Figura 1.30.

```
import os, io
def read_item_names():
    file_name = (os.path.expanduser('~') +
            '/.surprise_data/ml-100k/ml-100k/u.item')
   rid_to_name = {}
   with io.open(file_name, 'r', encoding='ISO-8859-1') as f:
      for line in f:
         line = line.split('|')
         rid_to_name[line[0]] = line[1]
    return rid_to_name
```

Figura 1.30. Código Python para declarar função de seleção de nome dos filmes do MovieLens

Fonte: Documentação da biblioteca Surprise[5]

Por fim, o código da Figura 1.31 demonstra como realizar a chamada dessas funções criadas para visualizar os resultados. O primeiro passo é utilizar a função "get_top4_recomendacoes" para selecionar apenas as 4 primeiras das predições geradas com o KNNBasic para cada aluno. Em seguida, é preciso selecionar os nomes desses filmes utilizando o método "read_item_names()". Como o retorno do método "get_top4_recomendacoes" é uma lista, é possível utilizar o método "items()" para iterar os valores da lista. Nesse caso, o objetivo é o de apenas associar o nome do filme para cada um dos "iid" (identificador do item, ou seja, o filme retornado no método).

```
top4_recomendacoes = get_top4_recomendacoes(predicoes)
rid_to_name = read_item_names()
for uid, avaliacoes_usuario in top4_recomendacoes.items():
   print(uid, [rid_to_name[iid] for (iid, _) in avaliacoes_usuario])
```

Figura 1.31. Código Python para exibir recomendação para os usuários sem classificações do *dataset*

[4] Disponível em https://surprise.readthedocs.io/en/stable/FAQ.html

[5] Disponível em https://surprise.readthedocs.io/en/stable/FAQ.html

Após execução do código Python da Figura 1.31 no Python, é retornado imediatamente na tela uma lista com o identificador do usuário (id), seguido pelo nome dos quatro filmes mais similares de acordo com o algoritmo anteriormente implementado. Uma parte desse retorno pode ser visualizado na Figura 1.32, que exibe os resultados para os usuários de *ids* 196,186, 22, 166, 298 e 115. É possível perceber que cada nome de filme vem acompanhado também do ano de lançamento entre parênteses.

```
196 ['Very Natural Thing, A (1974)', 'Walk in the Sun, A (1945)', 'War at Home, The (1996)', 'Sunchaser, The (1996)']

186 ['Mamma Roma (1962)', 'Conspiracy Theory (1997)', 'Toy Story (1995)', 'MatchMaker, The (1997)']

22 ['Entertaining Angels: The Dorothy Day Story (1996)', 'King of New York (1990)', 'Usual Suspects, The (1995)', "One Flew Over the Cuckoo's Nest (1975)"]

244 ['Other Voices, Other Rooms (1997)', 'Big Bang Theory, The (1994)', 'Godfather, The (1972)', 'Casablanca (1942)']

166 ['Mamma Roma (1962)', 'Delta of Venus (1994)', 'Carmen Miranda: Bananas Is My Business (1994)', 'Marlene Dietrich: Shadow and Light (1996) ']

298 ['North by Northwest (1959)', 'Pinocchio (1940)', 'Amadeus (1984)', 'When Harry Met Sally... (1989)']

115 ['2001: A Space Odyssey (1968)', 'Clockwork Orange, A (1971)', 'Three Colors: White (1994)', 'Leaving Las Vegas (1995)']
```

Figura 1.32. Extrato de resultados da recomendação de filmes utilizando biblioteca Surprise

1.8. Principais Desafios e Oportunidades de Pesquisa

Apesar dos algoritmos de recomendação serem amplamente utilizados nos últimos anos, ainda existem diversos desafios em aberto. A privacidade de dados abordada na Seção 1.5 é um aspecto que deve ser considerado em qualquer solução que manipule dados de usuários. Além dessas questões abordadas, pode-se mencionar alguns desafios:

- Falta de dados: não é sempre que usuários possuem dados a serem extraídos das plataformas para que seja feita a recomendação. Em uma aplicação de venda de imóveis, por exemplo, podemos considerar que seria aceitável se a aplicação não dispusesse de muitos dados de muitos clientes, uma vez que a compra de uma casa é algo pouco frequente. Para solucionar esse desafio, os sistemas de recomendação podem implementar técnicas baseadas em conhecimento que incluem uma base de conhecimento do usuário para apoiar na recomendação [Aggarwal 2016].
- Dificuldade para mostrar novos dados: em uma recomendação colaborativa de filmes, como a do nosso exemplo, é preciso considerar que pode existir uma dificuldade de incluir no resultado final da recomendação determinados filmes que são pouco classificados por usuários. Isso ocorre devido ao fato de que a classificação de usuários é um critério importante no cálculo de similaridade

entre itens. Se há ausência de classificações, pode existir consequentemente um baixo índice de recomendação.

- *Cold-start*: ainda há uma dificuldade em recomendar itens para usuários que não possuem muito tempo de utilização das plataformas. Logo, esses usuários possuem poucos dados registrados, como o de vizinhos (no caso da filtragem colaborativa) ou de históricos (no caso da baseada em conteúdo) para que a recomendação tenha mais chances de acerto.
- Itens imprevisíveis: alguns itens não possuem uma classificação totalmente em comum para todos os usuários. Um exemplo seria em recomendação de músicas. Algumas músicas são adoradas por muitos usuários e odiadas por diversos outros. Isso faz com que as soluções de recomendação tenham dificuldade de identificar se essa música pode ser um item relevante a ser considerado no processo de recomendação.

Os problemas mencionados são amplamente envolvidos em tópicos de pesquisa sobre sistemas de recomendação. Surgem possibilidades de inclusão de técnicas de mineração de dados (como em [Campos et al. 2019]), Linked Open Data (como em [Noia e Ostuni 2015]), redes sociais [Logesh et al. 2018], bem como estudos que abordem a perspectiva de Ecossistemas de Software (ECOS) para melhor compressão das interações do software, facilitando a reutilização (conforme adotado em [Campos et al. 2018] e/ou possibilitando o compartilhamento de soluções de recomendação, permitindo a integração com outros sistemas de informação (a exemplo de [Abdalla et al. 2018]).

1.9. Conclusão

Com o avanço da tecnologia, os algoritmos de recomendação têm sido utilizados cada vez mais, e em diversos cenários. Seja utilizando dados de diversos usuários ou apenas dados históricos de um usuário interessado, esses algoritmos permitem solucionar problemas, otimizar os processos de busca e recuperação da informação e consequentemente atribuem valor para as aplicações que os utilizam. Apesar da fácil compreensão do objetivo final de um algoritmo de recomendação, o processo em si pode ser complexo e pode utilizar ainda outras técnicas, como a de mineração de dados, para auxiliar no resultado final.

Uma preocupação existente na utilização desses algoritmos atualmente é a privacidade de dados dos usuários. Isso demanda uma necessidade em reavaliar quais dados são utilizados para o processo de recomendação de um usuário, quais dados de um usuário são utilizados para os processos de recomendação de outros usuários, mas também políticas públicas de privacidade de dados que amparem usuários de diversos websites ou qualquer solução que possua extração e uso de dados de usuários.

Esse capítulo apresentou os principais conceitos dos algoritmos de recomendação, ressaltando teorias da literatura que demonstram o funcionamento e o comportamento dos diversos tipos de soluções de recomendação. Entendemos ser essencial que os conhecimentos de linguagem de programação sejam introduzidos para que um interessado em implementar um algoritmo de recomendação possa compreender seu funcionamento em sua totalidade. Nesse contexto, adotamos o Python como linguagem de programação e demostramos as principais operações através do Google Colab.

A implementação do algoritmo de recomendação demonstrada nesse capítulo pode ser totalmente adotada através de outra linguagem de programação, desde que respeitados as particularidades de cada linguagem. Os resultados da recomendação exibem como essa técnica pode auxiliar na recuperação de uma informação próxima daquilo que determinado usuário busca, sendo assim, portanto, um recurso importante para as técnicas de busca e recuperação da informação.

Referências

Abdalla, A., Ströele, V., Campos, F., David, J. M. N. and Braga, R. (2018). Plataforma de Ecossistema de Software para Sistemas de Recomendação. In *Anais do XIV Simpósio Brasileiro de Sistemas de Informação*. . SBC.

Aggarwal, C. C. (2016). *Recommender systems*. Cham: Springer International Publishing.

Armknecht, F. and Strufe, T. (2011). An Efficient Distributed Privacy-preserving Recommendation System. [IEEE, Ed.]In *2011 The 10th IFIP Annual Mediterranean Ad Hoc Networking Workshop.* . https://researcher.ibm.com/researcher/view.

Baeza-Yates, R. and Ribeiro-Neto, B. (2013). *Recuperação de Informação - 2ed: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora.

Balabanović, M. and Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, v. 40, n. 3, p. 66–72.

Campos, R., Santos, R. P. and Oliveira, J. (2018). Web-Based Recommendation System Architecture for Knowledge Reuse in MOOCs Ecosystems. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. . IEEE. https://ieeexplore.ieee.org/document/8424707/.

Campos, R., Santos, R. P. and Oliveira, J. (2019). A Recommendation System Enhanced by Topic Modeling for Knowledge Reuse in MOOCs Ecosystems. *Reuse in Intelligent Systems*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing (in press). .

Cazella, S. C., Chagas, I. C. Das, Barbosa, J. L. V. and Reategui, E. B. (2008). Um modelo para recomendação de artigos acadêmicos baseado em filtragem colaborativa aplicado à ambientes móveis. *RENOTE: revista novas tecnologias na educação [recurso eletrônico]. Porto Alegre, RS*,

Cazella, S. C., Drumm, J. V. and Barbosa, J. L. V (2010). Um serviço para recomendação de artigos científicos baseado em filtragem de conteúdo aplicado a dispositivos móveis. *RENOTE*, v. 8, n. 3.

Cazella, S. C., Nunes, M. and Reategui, E. (2010). A Ciência da Opinião: Estado da arte em Sistemas de Recomendação. *André Ponce de Leon F. de Carvalho; Tomasz Kowaltowski..(Org.). Jornada de Atualização de Informática-JAI*, p. 161–216.

Costa, E., Aguiar, J. and Magalhães, J. (2013). Sistemas de Recomendação de Recursos Educacionais: conceitos, técnicas e aplicações. *Jornada de Atualização em Informática na Educação*, v. 1, n. 1.

Desrosiers, C. and Karypis, G. (2011). A Comprehensive Survey of Neighborhood-based Recommendation Methods. *Recommender Systems Handbook*. Boston, MA:

Springer US. p. 107–144.

Do, M.-P. T., Nguyen, D. V. and Nguyen, L. (2010). Model-based Approach for Collaborative Filtering. *Proceedings of The 6th International Conference on Information Technology for Education (IT@EDU2010)*, n. August 2010, p. 217–225.

Feng, P., Zhu, H., Liu, Y., Chen, Y. and Zheng, Q. (2018). Differential Privacy Protection Recommendation Algorithm Based on Student Learning Behavior. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*.

Grossman, D. A. and Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics*. Springer.

Herlocker, J. L., Konstan, J. A. and Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*.

Jiang, J.-Y., Li, C.-T. and Lin, S.-D. (2019). Towards a More Reliable Privacy-preserving Recommender System. *Journal of Information Sciences*, v. 482, p. 248--265.

Jones, K. S. (1997). *Readings in Information Retrieval*. Morgan Kaufman.

Konstan, J. A., Miller, B. N., Maltz, D., et al. (1997). GroupLens: applying collaborative filtering to Usenet news. *Communications of the ACM*, v. 40, n. 3, p. 77--87.

Koren, Y. (2008). Factorization meets the neighborhood. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. . ACM Press. http://doi.acm.org/10.1145/.

Logesh, R., Subramaniyaswamy, V. and Vijayakumar, V. (2018). A personalised travel recommender system utilising social network profile and accurate GPS data. *Electronic Government, an International Journal*, v. 14, n. 1, p. 90–113.

Mcsherry, F. and Mironov, I. (2009). Differentially Private Recommender Systems: Building Privacy into the Netflix Prize Contenders. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Motta, C. L. R. Da, Garcia, A. C. B., Vivacqua, A. S., Santoro, F. M. and Sampaio, J. O. (2011). Sistemas de recomendação. *Pimentel, M.; Fuks, H."Sistemas colaborativos". Rio de Janeiro: Elsevier*,

Noia, T. Di and Ostuni, V. C. (2015). Recommender Systems and Linked Open Data. *Reasoning Web International Summer School*. Springer. http://www.pandora.com/, [accessed on Apr 13].

Qi, L., Xiang, H., Dou, W., et al. (7 sep 2017). Privacy-Preserving Distributed Service Recommendation Based on Locality-Sensitive Hashing. In *Proceedings - 2017 IEEE 24th International Conference on Web Services, ICWS 2017*. . Institute of Electrical and Electronics Engineers Inc.

Ricci, F., Rokach, L., Shapira, B. and Kantor, P. B. (2015). *Recommender systems handbook*. Springer.

Santos, L. C. de M. Dos and Bräscher, M. (31 dec 2017). Uso de ontologia na recuperação da informação em acervos digitais de jornais. *Informação & Informação*, v. 22, n. 3, p. 346.

Shardanand, U. and Maes, P. (1995). Social Information Filtering: Algorithms for Automating \Word of Mouth". *Chi*. Citeseer. v. 95p. 210--217.

Wang, C., Zheng, Y., Jiang, J. and Ren, K. (1 feb 2018). Toward Privacy-Preserving Personalized Recommendation Services. *Engineering*, v. 4, n. 1, p. 21–28.

Sobre os Autores

Leonardo Herdy Marinho

Mestrando no Programa de Pós-Graduação em Informática (PPGI) da Universidade Federal do Rio de Janeiro (UFRJ) na área de Sistemas de Informação, com um tema de pesquisa focado em investigar como os sistemas de recomendação podem apoiar os ecossistemas de startups na geração de parcerias. Profissional com 5 anos de experiência como analista e desenvolvedor de sistemas. Graduado em Análise e Desenvolvimento de Sistemas, possui experiência em projetos com PHP, Laravel, Javascript, NodeJs, Angular, Ionic, Dart, Flutter, Aqueduct, Django e Python.

Rodrigo Campos

Mestrando no Programa de Pós-Graduação em Informática (PPGI) da Universidade Federal do Rio de Janeiro (UFRJ) na área de Sistemas de Informação, com um tema de pesquisa focado em investigar como os sistemas de recomendação podem apoiar os ecossistemas MOOCs na reutilização do conhecimento. Profissional com 4 anos de experiência como desenvolvedor de sistemas em empresas governamentais brasileiras, como a Marinha do Brasil e atualmente o Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro (IFRJ). Graduado em Análise e Desenvolvimento de Sistemas, possui experiência em projetos com JavaEE, JSP, Hibernate e SpringMVC.

Rodrigo Pereira dos Santos

Professor Adjunto do Departamento de Informática Aplicada (DIA) e membro efetivo do Programa de Pós-Graduação em Informática (PPGI) da Universidade Federal do Estado do Rio de Janeiro (UNIRIO), onde atualmente é Coordenador do Curso de Mestrado. Atuou como pesquisador visitante na University College London (BEX/CAPES, 2014-2015). Atuou como consultor em projetos de pesquisa e desenvolvimento em engenharia de sistemas na indústria nacional pela Fundação Coppetec (2008-2017). É editor-chefe da iSys: Revista Brasileira de Sistemas de Informação. É membro da Sociedade Brasileira de Computação (SBC) desde 2006 e Coordenador do Comitê Gestor da Comissão Especial de Sistemas de Informação (CE-SI) da SBC. Tem experiência na área de Ciência da Computação, com ênfase em Engenharia de Software e Sistemas de Informação. Seus principais campos de atuação são Engenharia de Sistemas Complexos (especialmente ecossistemas de software e sistemas-de-sistemas) e Educação em Engenharia de Software. Foi coordenador científico de mais de 20 eventos (simpósios, trilhas e workshops) no Brasil e no exterior e proferiu comunicações (palestras, minicursos e tutoriais) em mais de 20 eventos nacionais.

Mônica Ferreira da Silva

Professora do Programa de Pós-graduação em Informática (PPGI) da Universidade Federal do Rio de Janeiro (UFRJ). Doutora em Administração pelo COPPEAD/UFRJ (2006). Mestre em Engenharia de Sistemas e Computação pela COPPE/UFRJ (1998). Especialização em Gerência de Projetos pelo NCE/UFRJ (2001). Graduação em Informática pela Universidade Federal do Rio de Janeiro (UFRJ) concluído com grau Cum Laude de dignidade acadêmica (1988). Tem experiência nas áreas de Ciência da Computação e Administração, com ênfase em Gestão da Tecnologia da Informação. Atuando principalmente nos seguintes temas: estratégia e sistemas de informação, metodologia de pesquisa científica e adoção de tecnologia.

Jonice Oliveira

Short Bio: Profa. Jonice Oliveira obteve o seu doutorado em 2007 na área de Engenharia de Sistemas e Computação, ênfase em Banco de Dados, pela COPPE/UFRJ. Durante o seu doutorado recebeu o prêmio IBM Ph.D. Fellowship Award. Na mesma instituição realizou o seu Pós-Doutorado, concluindo-o em 2008. Desde 2009 é professora do Departamento de Ciência da Computação da UFRJ e atualmente é coordenadora do Programa de Pós-Graduação em Informática (PPGI-UFRJ). Coordena o Laboratório CORES (Laboratório de Computação Social e Análise de Redes Sociais), que conduz pesquisas multidisciplinares para o entendimento, simulação e fomento às interações sociais. Sua principal área de pesquisa é Computação Social, mais especificamente nos temas de Análise de Redes Sociais, Big Social Data, Suporte à Decisão e Recomendação. Possui uma larga experiência em tais áreas, com mais de centenas de artigos publicados, dezenas de orientações e envolvimento (como membro ou como líder) em projetos de pesquisas nacionais e internacionais. Maiores detalhes em: http://www.joniceoliveira.net/.

Capítulo 2

Introdução à Análise de Sentimentos com Word Clouds

André Viana Tardelli, Angélica Fonseca da Silva Dias, Juliana Baptista dos Santos França

Abstract

The goal of this short course is to allow an introductory discussion about Data Science, specially the Sentiment Analysis area, in order to present its results graphically and statistically. The course introduces the basic concepts of natural language processing via text and argues for strategies for analyzing large amounts of data in order to detect general opinions about a given number of assessments provided. The construction of this analysis is made through the execution of structured stages, which determine each analysis step separately, such as Training, Optimization and Graph Generation.

Resumo

O objetivo deste capítulo é permitir uma discussão introdutória sobre a área de conhecimento de Ciência de Dados, e abordando a Análise de Sentimento, de maneira a apresentar seus resultados de forma gráfica e estatística. O curso introduz os conceitos básicos de processamento da linguagem natural via texto, e argumenta sobre estratégias de análise de grandes quantidades de dados a fim de detectar opiniões gerais sobre determinado número de avaliações fornecidas. A construção desta análise é feita via a execução de estágios estruturados, que determinam cada etapa da análise separadamente, como Treinamento, Otimização e Geração dos Gráficos.

2.1. Introdução

Devido à grande quantidade de produtos e serviços disponibilizados via meios digitais na sociedade atual, a opinião das pessoas que os utilizam é um fator cada vez mais valioso para o discernimento de sua qualidade [Zhang et al., 2012]. Dessa forma, diversos profissionais em empresas são responsáveis por processar e analisar o *feedback* dado pelos usuários em canais digitais para prover sugestões e ideias para melhorar os produtos disponibilizados pelas empresas.

Todavia, muitas vezes esse processamento e análise do *feedback* dos usuários por humanos acabam tornando-se ineficaz, visto que existe um número demasiadamente grande de comentários e avaliações a serem lidos. Dessa forma, diversos *feedbacks* acabam sendo perdidos, visto que é muito difícil gerar um compilado eficaz de informações capaz de mostrar a opinião geral que determinados usuários estão sentindo em relação a algum produto. Assim, notou-se a escassez de algum método que fosse capaz de destacar os *feedbacks* mais relevantes, de acordo com o tipo de opinião fornecida [Agarwal et al., 2015].

Ao longo dos últimos anos foi introduzido o conceito de análise de sentimento, que possibilitou explicitar opiniões de maneira mais direta, categorizando os elementos principais de uma massa de texto de acordo com a sua relevância tanto positivamente quanto negativamente. A identificação de sentimentos em artefatos textuais é uma das áreas de pesquisa mais destacadas em Processamento de Linguagem Natural desde o início dos anos 2000 [Liu, 2010]. Dessa maneira, o principal objetivo da análise de sentimentos é definir técnicas automáticas capazes de extrair informações subjetivas de textos em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado [Benevenuto et al., 2015]. Empresas relevantes no mercado podem analisar seus produtos e serviços de maneira geral e assim, obter um *feedback* mais preciso sobre a massa de comentários disponibilizados por seus usuários.

Este capítulo tem como foco mostrar os conceitos básicos para gerar uma análise de sentimento com base em uma massa de dados (grande volume de comentários), de modo a conseguir representar de maneira gráfica e interativa os resultados obtidos. Para isto, serão mostrados os conceitos de aprendizado de máquina e processamento de linguagem natural através da linguagem *Python*, em um ambiente colaborativo e fácil de ser preparado. Ao final, espera-se que os participantes tenham internalizado conceitos relacionados à: i) Aprendizado de Máquina ii) Processamento de Linguagem Natural iii) Análise de Sentimento e iv) Manipulação de *datasets* e geração de gráficos na linguagem *Python*. É importante destacar que será apresentado no texto a mineração (análise de sentimentos) usando métodos de classificação por aprendizado de máquina supervisionado. No entanto, esta mineração poderia ser conduzida por outras técnicas como as não supervisionadas (dicionários léxicos), ou semi-supervisionados.

O capítulo 2 está organizado de maneira estruturada, sendo constituído de uma breve compreensão das abordagens teóricas e práticas. A seção 2.2 contém a base do referencial teórico a ser discutido, a seção 2.3 possui as especificações de ambiente e bibliotecas a serem utilizadas, a seção 2.4 discute todos os passos a serem realizados durante a parte prática do capítulo, e a seção 2.5 contém uma breve discussão dos resultados alcançados.

2.2. Referencial Teórico

Antes de realizar uma análise de sentimento, é necessário rever os conceitos base que compõem a sua estrutura. Nesta seção, serão apresentadas as bases teóricas da área de Ciência de Dados que serão aplicadas em diversas partes da construção da análise a ser realizada neste capítulo.

2.2.1. Aprendizado de Máquina

Com o avanço da tecnologia, tornou-se cada vez mais necessário a criação de sistemas inteligentes capazes de simular ações realizadas por seres humanos. Essas simulações ocorrem através da aplicação de algoritmos de Aprendizado de Máquina, com o objetivo de extrair informações de dados fornecidos e consequentemente desenvolver um modelo geral que seja capaz de representar o problema estudado [Horta, 2015]. Dessa forma, diferentes técnicas na área de Ciência de Dados foram implementadas de maneira a induzir a tomada de alguma ação, seja baseada em experiências anteriores ou a partir de medidas de qualidade de respostas obtidas.

Segundo Monard e Baranauskas (2003), o conceito de Aprendizado de Máquina possui como objetivo desenvolver técnicas computacionais para adquirir conhecimento de maneira automática, de forma a possibilitar que a máquina possa tomar decisões autônomas. Esse paradigma tornou-se cada vez mais popular ao longo dos anos, visto que a implementação de processos indutivos possibilitou a criação de classificadores automáticos dado um conjunto pré-determinado de dados.

Dessa forma, a utilização de aprendizado de máquina trouxe diversas vantagens na área de processamento de texto, visto que as acurácias das classificações são comparáveis a de um ser humano, evitando assim a necessidade da intervenção de especialistas da área para analisar e gerar os classificadores principais de diferentes tipos de categorias [Sebastiani, 2002].

Na área de Aprendizado de Máquina, é possível treinar o seu algoritmo através de uma maneira supervisionada, não supervisionada ou semi supervisionada. Com o foco na análise de sentimentos, e de acordo com Benevenuto, Ribeiro e Araújo (2015), a supervisionada é embasada nos conceitos de aprendizagem de máquina partindo da definição de características que permitam distinguir entre sentenças com diferentes sentimentos, treinamento de um modelo com sentenças previamente rotuladas e utilização do modelo de forma que ele seja capaz de identificar o sentimento em sentenças até então desconhecidas. Já a não supervisionada não conta com treinamento de modelos de aprendizado de máquina e, em geral, são baseadas em tratamentos léxicos de sentimentos que envolvem o cálculo da polaridade de um texto a partir de orientação semântica das suas palavras. Por fim a semi supervisionada trata-se de uma grande oportunidade para quem não pode bancar o preço de treinar todos os seus dados. Este método permite-nos melhorar significativamente a acurácia, pois permite utilizar dados não treinados com uma pequena quantidade de dados treinados. Neste capítulo, será utilizado o aprendizado supervisionado para realizar a categorização dos elementos, sendo discutido com mais detalhes na seção 2.2.1.1.

2.2.2. Aprendizado Supervisionado

Uma abordagem muito comum para gerar um aprendizado supervisionado é a separação da massa de dados em uma massa de treino e uma massa de teste, contendo as informações principais a serem analisadas. Dessa forma, a massa de treino é responsável por possuir todas as informações a serem induzidas pelo algoritmo, enquanto a massa de teste será utilizada para averiguar se as instâncias principais daquele conjunto de dados conseguem ser corretamente previstos.

De acordo com Benevenuto et. al., (2015) o termo supervisionado é apresentado justamente pelo fato de exigir uma etapa de treinamento de um modelo com amostras previamente classificadas. O procedimento para realizar a aprendizagem de máquina compreende as seguintes etapas principais: I. obtenção de dados rotulados para uso em treino e para teste; II. Definição das *features* ou características que permitam a distinção entre os dados; III. Treinamento de um modelo computacional com um algoritmo de aprendizagem; e IV. Aplicação do modelo. Essas etapas serão conduzidas neste capítulo através da biblioteca SKLearn apresentada nas sessões abaixo deste capítulo.

2.2.3. Processamento de Linguagem Natural

Uma das grandes dificuldades na área de Ciência de Dados é a capacidade de fazer com que máquinas consigam processar e interpretar textos. Para conseguir simular essa interpretação, foram elaborados conjuntos de técnicas para representar elementos textuais de maneira computacional de modo a alcançar um nível de interpretação linguística em uma máquina a nível de um ser humano.

Segundo Liddy (2001), o conceito de Processamento de Linguagem Natural é definido por um conjunto de técnicas computacionais para analisar e representar naturalmente textos em um ou mais níveis linguísticos de análise com o propósito de alcançar um nível humano de interpretação para diversos tipos de tarefas ou aplicações. Este nível de interpretação pode ser definido em diversas categorias, se baseando em estruturas morfológicas, fonéticas, lexicais ou semânticas. Desta forma, diferentes abordagens podem ser feitas para alcançar um nível de interpretação desejada, de acordo com as necessidades de cada análise.

Como o objetivo do capítulo visa categorizar palavras-chave baseadas no sentimento geral, uma análise léxica será utilizada para processar estes dados de maneira a gerar uma análise individual de cada termo. Todavia, os textos a serem analisados se encontram muito simples, dificultando a relevância de termos de maior importância. Dessa forma, será necessário realizar um pré-processamento no corpus textual a ser analisado de maneira a otimizar a interpretação de todos os termos presentes no mesmo. Estes métodos de pré-processamento serão discutidos mais à frente na seção 2.4.

2.2.4. Análise de Sentimentos

A área de Análise de Sentimento tem como foco principal explicitar termos principais referentes a uma opinião em forma de texto [Serranoguerrero, 2015]. Dessa forma, é possível analisar uma extensa variedade de dados referentes a algum produto ou serviço,

possibilitando uma visão generalizada da opinião dos clientes da mesma através de relatórios ou recursos gráficos.

Todavia, antes de adentrar num detalhamento sobre seus métodos, é necessário especificar o escopo a ser trabalhado ao citar a temática de sentimento. Sendo um conceito estudado em diversas áreas como psicologia, computação e biologia, o sentimento é representado em diversos estudos da literatura de maneiras distintas [Ceci et. al., 2017].

Sentimento ou emoção indica uma carga de sentido específico presente em uma mensagem, que pode ser: raiva, surpresa, felicidade, alegria, tristeza, etc. Alguns métodos apresentam abordagens capazes de identificar qual sentimento uma sentença representa. Um exemplo clássico trata-se da abordagem léxica Emolex [Mohammad and Turney, 2013], a qual é baseada a partir da avaliação de milhares de sentenças em inglês para 9 sentimentos diferentes: *joy, sadness, anger, fear, trust, disgust, surprise, anticipation, positive, negative*. A força do sentimento representa a sua intensidade. Normalmente, este resultado é flutuante entre (-1 e 1). Há trabalhos que por exemplo medem a força de sentimentos nos títulos das notícias como o Magnet News[Reis et al., 2014] [Reis et al., 2015b], capaz de separar eficientemente para o usuário notícias boas de notícias ruins.

Partindo de uma avaliação cognitiva do assunto, Jung (2003) qualifica os sentimentos através de um sinal positivo ou negativo. Dessa forma, é possível obter a categorização de um sentimento em um viés computacional através da polarização fornecida em um termo ou conteúdo. Assim, é muito comum ver a área de análise de sentimento associada com temas como Processamento de Linguagem Natural, por exemplo.

A área de Análise de Sentimento muitas vezes é igualada a área de Mineração de Opinião, pois ambas surgiram com o intuito de realizar tarefas de identificação, classificação, análise de opiniões e sentimentos. Dessa forma, ambas podem ser classificadas como a área da computação que estuda todos os sentimentos, opiniões e emoções expressas em um texto [Ceci et. al., 2017].

2.3. Ferramentas Utilizadas

Nesta etapa, serão discutidas as especificações do ambiente, bibliotecas e ferramentas a serem importadas e utilizadas para gerar a análise. As ferramentas foram escolhidas tendo em mente a utilização da linguagem *Python*, devido a vasta quantidade de bibliotecas e *plugins* disponíveis para a realização da análise de dados e por ser uma linguagem relativamente simples para a realização dos trechos de código a serem executados. Todavia, não é necessário um extenso conhecimento da linguagem em si, visto que todas as estruturas e nomenclaturas básicas de função serão discutidas passo a passo ao longo do capítulo.

O ambiente a ser utilizado para a execução dos *scripts* em *Python* será a plataforma *Google Colaboratory*, que proporciona um ambiente online sem necessidade

de configurações complexas e fornece um processamento de dados sobre uma máquina virtual disponibilizada gratuitamente através de uma conta *Google.* Dessa forma, o único recurso necessário para todas as máquinas a serem utilizadas será a disponibilidade de computadores com acesso à internet.

Além disso, alguns *scripts* necessitaram de funcionalidades que requerem a importação de alguma biblioteca, que serão explicitadas conforme o avanço da dinâmica. As bibliotecas utilizadas podem ser consultadas através da Tabela 1.1.

Tabela 1 - Bibliotecas a serem importadas e utilizadas durante a execução do capítulo

BIBLIOTECA	DESCRIÇÃO DA FUNCIONALIDADE
PANDAS	Possibilita a leitura e manipulação de estruturas de dados através de arquivos contendo agrupamentos de dados.
MATPLOTLIB	Biblioteca de Dados 2d capaz de produzir imagens de alta qualidade através da análise dos dados fornecidos. Será utilizada para plotar gráficos estatísticos e os *word clouds* gerados na análise
NLTK	Plataforma voltada para trabalhar com dados gerados pela linguagem humana. Será utilizada para fazer o processamento de linguagem natural dos textos fornecidos.
SKLEARN	Biblioteca de aprendizado de máquina, contendo algoritmos de regressão, agrupamento e seleção para gerar diferentes tipos de aprendizado.

2.3.1. Descrição do Dataset

De maneira a categorizar o sentimento associado a um conjunto de palavras, é necessária alguma base de dados contendo diversos comentários e opiniões associados a algum tema. Uma das maneiras mais simples de conseguir montar estas bases é coletando dados de alguma página contendo avaliações de clientes, onde normalmente existe alguma nota associada a eles.

A base de dados a ser utilizada será a base do IMDb (*Internet Movie Database*), que possui *reviews* de filmes, músicas e programas de televisão em geral. Devido a grande massa de dados disponível em conjunto das notas associadas ao comentário do mesmo, a massa é muito conveniente e propícia para realizar uma mineração de opinião.

De modo a fazer a análise do texto utilizando o idioma em português, a comunidade de desenvolvedores na plataforma Kaggle traduziu o conteúdo dos comentários em inglês e adaptou para o português brasileiro. Para facilitar a manipulação de dados e o tamanho do arquivo a ser enviado durante a realização do capítulo, o *dataset* fornecido também foi adaptado, possuindo as seguintes colunas:

Tabela 2 - Colunas principais do dataset a ser fornecido para análise

COLUNA	DESCRIÇÃO
TEXT_PT	Texto do comentário traduzido para o português brasileiro de um programa ou filme assistido
SENTIMENTO	Representa a opinião geral, onde notas iguais ou maiores a 7 = “pos” e abaixo de 7 = “neg”
CLASSIFICACAO	Classificação do sentimento convertido para um valor binário, de modo que recebe 0 para sentimento positivo e 1 para sentimento negativo.

2.4. Execução dos Estágios

Após explicar os objetivos do curso e ajudar os alunos no preparo do ambiente de desenvolvimento, a parte prática será dividida em quatro etapas (estágios), detalhadas a seguir (Figura 2.1):

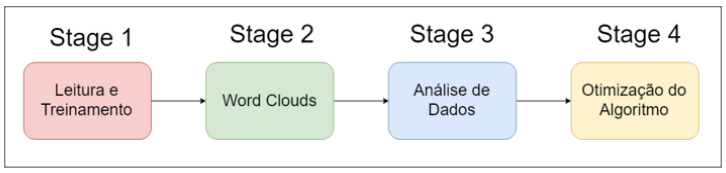


Figura 2.1 - Estrutura dos estágios a serem realizados no capítulo.

Cada etapa representará um dos passos realizados para a execução de uma análise de sentimento, de modo a facilitar o entendimento dos métodos de execução dividindo-os em partes. Além disso, um gabarito será disponibilizado contendo o conteúdo de cada parte ao término de sua explicação, de maneira a facilitar a identificação de possíveis erros durante a execução dos scripts por cada aluno. Cada etapa será detalhada sobre seus métodos e resultados a partir das próximas seções, sendo explicadas a seguir.

2.4.1. Stage 1

Nesse primeiro estágio, conforme a Figura 2.2**,** será iniciado o preparo do ambiente de desenvolvimento dos alunos através da leitura de uma base de dados previamente preparada para leitura e manipulação dos dados da mesma. Essa base de dados foi fornecida em um link para *download* no minicurso, possuindo uma base pré-processada e pronta para realizar a análise.

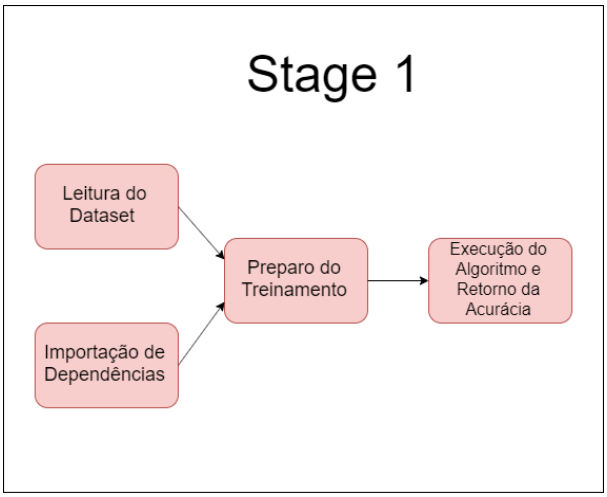


Figura 2.2 - Estrutura do primeiro estágio a ser realizado.

Após baixar a base de dados, será pedido aos participantes para criarem um novo notebook na plataforma *Google Colaboratory*, através de suas contas *Google* (Figura 2.3). Após criarem, os mesmos deverão fazer o upload da massa de dados baixada diretamente para o notebook criado, conforme a Figura 2.4, possibilitando a sua manipulação de dados.

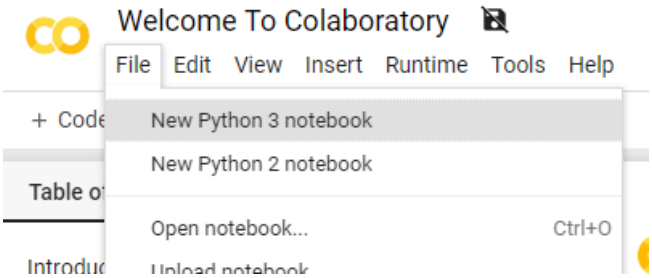


Figura 2.3 - Criação do arquivo a ser gerada a análise.

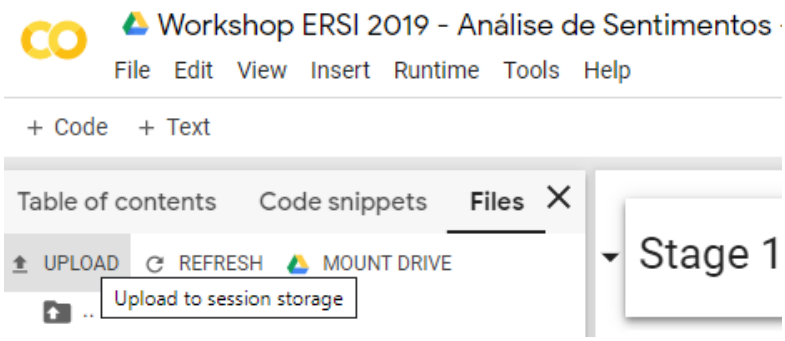


Figura 2.4 - Instruções de *upload* do *dataset* para possibilitar a manipulação de dados.

2.4.1.1. Utilizando a biblioteca Pandas para manipular dados

Com o ambiente e a massa preparados, é hora de começar a manipular os dados contidos na base. Realizando um comando para importar a biblioteca Pandas do *Python*, será possível ler os dados extraídos de um arquivo csv e convertê-lo para uma estrutura de dados propícia para habilitar a sua manipulação. Após converter esses dados em um *dataset*, é possível mostrar os resultados com a atribuição deste em uma variável, conforme mostrado na Figura 2.5.

	text_pt	sentiment	classificacao
0	Mais uma vez, o Sr. Costner arrumou um filme p...	neg	0
1	Este é um exemplo do motivo pelo qual a maiori...	neg	0
2	Primeiro de tudo eu odeio esses raps imbecis, ...	neg	0
3	Nem mesmo os Beatles puderam escrever músicas ...	neg	0
4	Filmes de fotos de latão não é uma palavra apr...	neg	0
5	Uma coisa engraçada aconteceu comigo enquanto ...	neg	0

Figura 2.5 - Primeira visualização dos dados do *dataset* na biblioteca Pandas.

2.4.1.2. *Bag-of-words*

Podendo manipular os dados, o próximo passo é converter todo o *corpus textual*, constituído por todos os comentários presentes no *dataset*, de maneira a criar uma representação para que o computador consiga interpretar estes dados. Uma das formas mais famosas para gerar essa visualização é através de uma abordagem onde todo o conteúdo do documento será representado como um vetor de palavras de acordo com suas ocorrências no mesmo [Matsubara et al., 2003]. Esse modelo de simplificação representativa é muito utilizado na área de Processamento de Linguagem Natural, sendo denominada *bag-of-words*.

A abordagem *bag-of-words* possibilita a representação de documentos textuais no formato de uma tabela atributo-valor, composta pelo número total de termos totais em cada uma de suas iterações. Dessa forma, todos os termos contidos no corpus textual tornam-se as colunas de um vetor, enquanto as linhas representam a frequência de cada uma das palavras contidas em cada iteração do mesmo. A presença de cada termo é categorizada de maneira binária, ou seja, se o termo está presente no documento, o valor

de uma posição Aij é acrescentado em 1, e caso contrário 0 [Matsubara et al., 2003] A Figura 2.6 ilustra a definição de um vetor utilizando a abordagem *bag-of-words*, de maneira a categorizar a frequência de todas as palavras presentes em cada frase do corpus textual de maneira binária.

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

Figura 2.6 - Exemplo de como os elementos são armazenados em uma *bag-of-words.*

2.4.1.3. Treinamento do algoritmo

Com o *array* montado, agora é possível analisar a frequência de todas as palavras em relação a todos os comentários. Dessa maneira, resta associar a frequência em conjunto ao sentimento associado, de maneira a verificar se as palavras associadas conseguem ser classificadas corretamente de acordo com o sentimento predominante na frase.

Para isso, será utilizado o conceito de aprendizado supervisionado previamente mencionado na seção 2.2.2, onde será separado 75% como massa de treino e 25% como massa de teste, de acordo com a função *train_test_split* da biblioteca SKLearn. Dessa forma, o *array* contendo a *bag-of-words*, em conjunto com a coluna de classificação do sentimento associado, irão compor as massas de treino e teste de acordo com esta proporção (Figura 2.7).

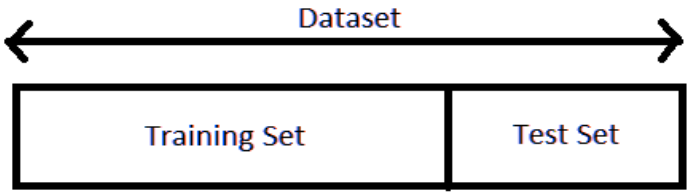


Figura 2.7 - Estrutura de uma base de treino e teste.

O último passo para finalizar este estágio é verificar a acurácia do modelo em si, de maneira a averiguar se os exemplos na massa de teste conseguem ser previstos corretamente baseados nos valores fornecidos pela massa de treino. Um dos métodos mais utilizados para conseguir prever os valores de uma resposta é o método da Regressão Linear, de modo que a resposta encontrada é baseada em um ou mais preditores [Hilbe, 2019]. Todavia, a utilização deste modelo proporciona uma resposta em uma variável contínua somente sendo baseada em valores numéricos, sendo assim incompatível com a análise dos termos no corpus textual por estarem em formato de texto.

Um outro modelo, denominado Regressão Logística, consegue realizar a previsão de variáveis discretas através de um valor binário fornecido, de maneira a

fornecer as chances de acerto dada determinada categoria no conjunto de dados. Devido à classificação binária do vetor de *bag-of-words*, é possível definir a acurácia do aprendizado supervisionado gerado utilizando este modelo, através da verificação dos valores na massa de teste com base na massa de treino (Figura 2.8).

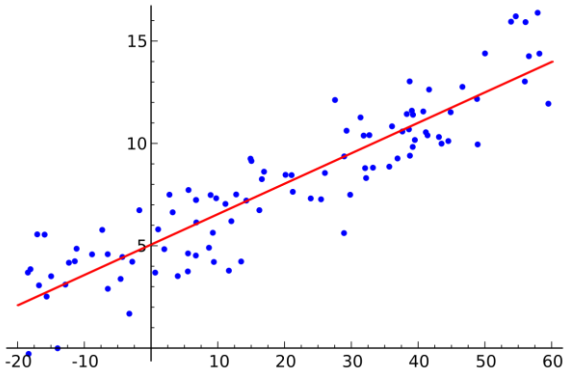


Figura 2.8 - Representação gráfica de uma Regressão Logística.

Desta forma, os valores binarizados no vetor *bag-of-words* são interpretados como variáveis discretas, sendo consequentemente preditores para a execução do algoritmo de Regressão Logística. Dessa forma, a porcentagem de acerto da regressão logística pode ser acessada através do método *score* da biblioteca SKLearn, baseada nos valores dados na coluna **"classificacao"** do *dataset* e nos valores separados na *bag-of-words*. A Figura 2.9 mostra a taxa de acerto obtida após a execução do algoritmo de Regressão Logística.

```
0.6664
```

Figura 2.9 - Primeira acurácia a ser obtida após a execução da Regressão Logística.

É importante destacar que neste capítulo está sendo aplicado o algoritmo de regressão já relatado devido sua simplicidade e objetivado para o objeto de estudo. No entanto, não é difícil encontrar trabalhos usando outros algoritmos como o Multinomial Naive Bayes (MNB) ou Support Vector Machine (SVM).

2.4.2. Stage 2

Nesse estágio, serão mostradas formas de representação dos dados obtidos através de imagens, de maneira a mostrar novas formas de interação e compreensão dos mesmos. A Figura 2.10 mostra um *overview* dos passos a serem realizados durante estes estágios.

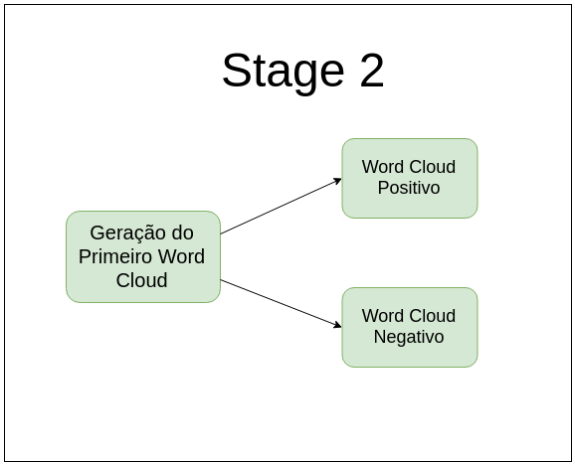


Figura 2.10 - Passos a serem realizados durante a execução do segundo estágio do *capítulo.*

2.4.2.1. *Word Cloud*

Ao gerar uma análise do corpus textual, é possível construir novas formas de visualização de dados para destacar a frequência das palavras geradas. Uma forma muito comum para desenvolver esse tipo de visualização é através de recursos gráficos, facilitando o entendimento dos resultados em geral. Segundo Lima (2008), o estudo da cultura visual em termos econômicos e tecnológicos pode proporcionar uma compreensão mais crítica em relação ao seu papel na contemporaneidade, facilitando o entendimento do tema e transcendendo o simples prazer visual que estas podem proporcionar.

Uma maneira para destacar a frequência de um termo é através de uma sumarização de texto, de forma a gerar um *overview* simples e intuitivo de um texto através do destaque de palavras que aparecem mais durante o mesmo [Heimerl et, al., 2014]. Isso normalmente é alcançado com o uso de recursos gráficos, como a mudança do tamanho da fonte de um texto que possui maior relevância ou possui um número maior de repetições presentes no mesmo.

Desta forma, pode-se ser introduzido o conceito de *word clouds*, que utiliza formas de percepção visual para facilitar a compreensão dos termos mais relevantes de maneira generalizada presentes em um corpus textual demasiadamente grande. Assim, um *word cloud* é representado por uma imagem contendo diversas palavras, onde a importância das mesmas é definida pelo tamanho de seu texto no canvas. A representação de um *word cloud* pode ser observada por meio da Figura 2.11.

Figura 2.11 - Um *word cloud* genérico (exemplo).

O objetivo principal nesse estágio é armazenar todas as palavras contidas na coluna de comentários e realizar uma contagem das mesmas, de modo a conseguir contabilizar o número de repetições feitas por uma delas. Ao contrário da representação no método *bag-of-words*, não será necessário armazenar as palavras de maneira binária, visto que o texto contido na palavra será essencial para representá-la no canvas da imagem a ser gerada.

Após a realização de um *loop* para armazenar todas as palavras em um vetor unidimensional, estes valores serão passados para a biblioteca *Python Wordclouds*, que será responsável por montar um objeto contendo uma figura com todas as palavras mais relevantes no mesmo. Essa imagem será plotada através da biblioteca Matplotlib, que irá definir as propriedades de tamanho e métodos de plotagem da figura. O primeiro resultado obtido pode ser visto na Figura 2.12:

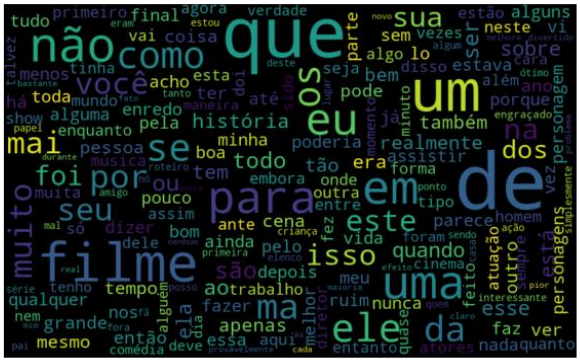


Figura 2.12 - Primeiro *word cloud* a ser gerado no segundo estágio.

Após gerar o primeiro *word cloud*, é proposto um desafio para aperfeiçoar as formas de visualização, transformando esses conhecimentos obtidos em uma função, e separando em dois tipos diferentes: um tipo para gerar *word clouds* somente para as palavras associadas a um sentimento positivo e outra para o sentimento negativo. O resultado pode ser observado nas Figuras 2.13 e 2.14.

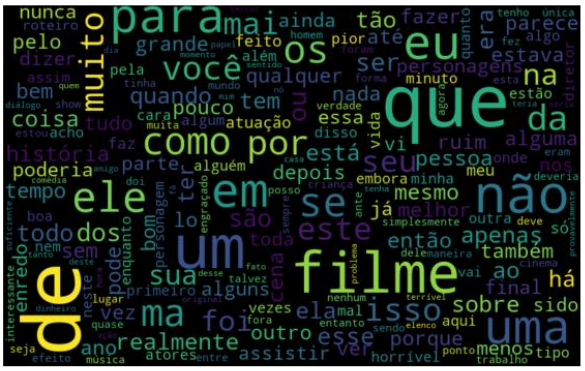


Figura 2.13 - *Word cloud* gerado contendo somente palavras associadas a um sentimento positivo.

Figura 2.14 - *Word cloud* gerado contendo somente palavras associadas a um sentimento *negative*.

2.4.3. Stage 3

Até agora, foram obtidas acurácia e uma forma de visualização em imagem, mas os resultados ainda não foram satisfatórios. A acurácia ainda possui uma precisão razoavelmente baixa, enquanto os *word clouds* mostraram valores semelhantes e contendo palavras em destaque como “de, que, um, os”.

Partindo desse ponto, é ideal que existam novas formas de visualização para analisar com maior precisão a frequência desses termos encontrados, de modo que venha a facilitar a geração de novas ideias para aperfeiçoar cada vez mais os algoritmos implementados. O objetivo desse estágio (Figura 2.15) é, então, proporcionar uma análise mais detalhada e estatística dos resultados obtidos.

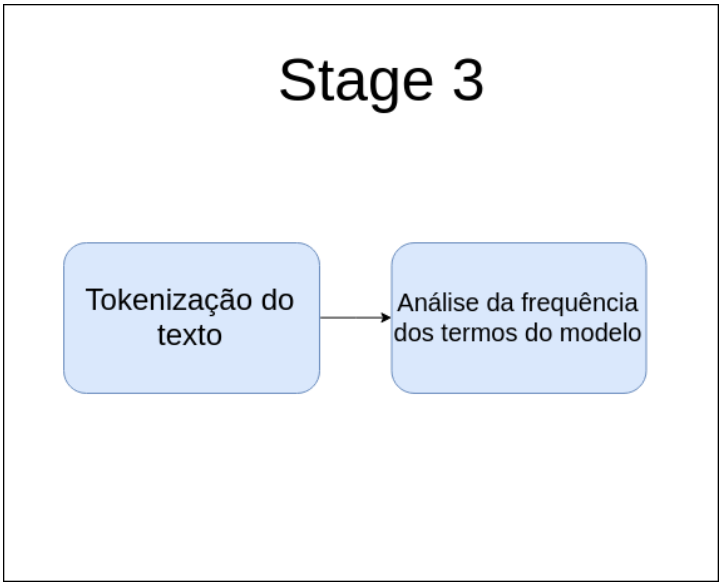


Figura 2.15 - Estrutura do terceiro estágio do capítulo.

2.4.3.1. Tokenização e análise estatística dos dados

Uma forma eficaz de ver a frequência dos termos é através da criação de outro *dataset*, contendo precisamente o número de repetições em cada comentário. Não será possível utilizar os valores armazenados de frequência armazenados dentro do objeto de *word cloud* gerado na seção 2.4.2.1, então será necessário utilizar outro método para gerar esse armazenamento.

Nesse caso, todas as palavras de uma frase precisam ser tokenizadas, ou seja, todo o corpus textual precisa ser convertido em lexemas. Um lexema é uma unidade de análise morfológica, sendo caracterizado como uma palavra, normalmente separada por um caractere de espaço dependendo do idioma [Chung e Gildea, 2009].

A partir desse método, todas as palavras podem ser caracterizadas, e assim possuírem seus valores agregados por suas frequências. Utilizando a biblioteca NLTK, esses valores podem ser somados e agregados no formato de um *array*, gerando uma nova coluna de frequência no novo *dataset*. Este resultado pode ser observado na Figura 2.16.

	Palavra	Frequencia
20	de	42441
14	que	32915
42	e	30045
3	o	24939
7	um	22420
102	a	21160
45	é	19511
200	em	13402
1	uma	13314
29	não	13003

Figura 2.16 - Primeira visualização da frequência dos dados obtidos em um novo *dataset.*

Esta visualização também pode ser feita através de um gráfico de barra. Reaproveitando o novo *dataset* obtido, é possível passar seus valores para serem plotados em uma figura através da biblioteca *Python Seaborn*, possibilitando uma visão estatística dos números obtidos de acordo com a quantidade desejada. Os resultados podem ser observados na Figura 2.17.

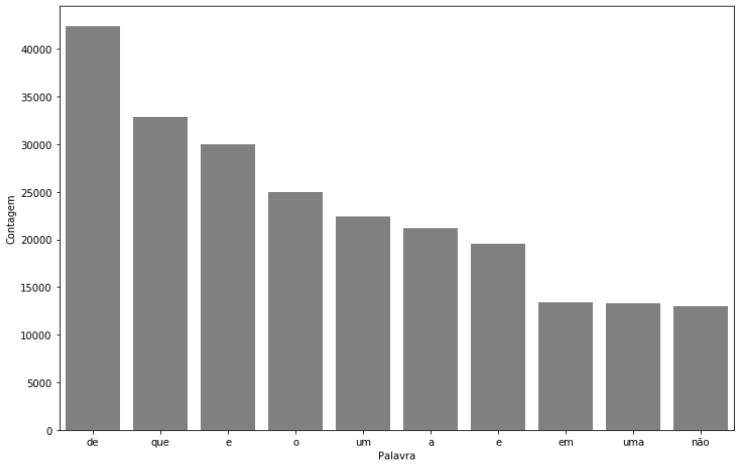


Figura 2.17 - Visualização da frequência dos dados em um gráfico de barra.

2.4.4. Stage 4

A etapa final consiste em otimizar a acurácia obtida na edição e remoção de trechos no texto do corpus textual que não agregam ao resultado desejado. Dessa forma, algumas das formas mais famosas para realizar o pré-processamento básico de um texto serão abordadas, conforme mostrado na Figura 2.18:

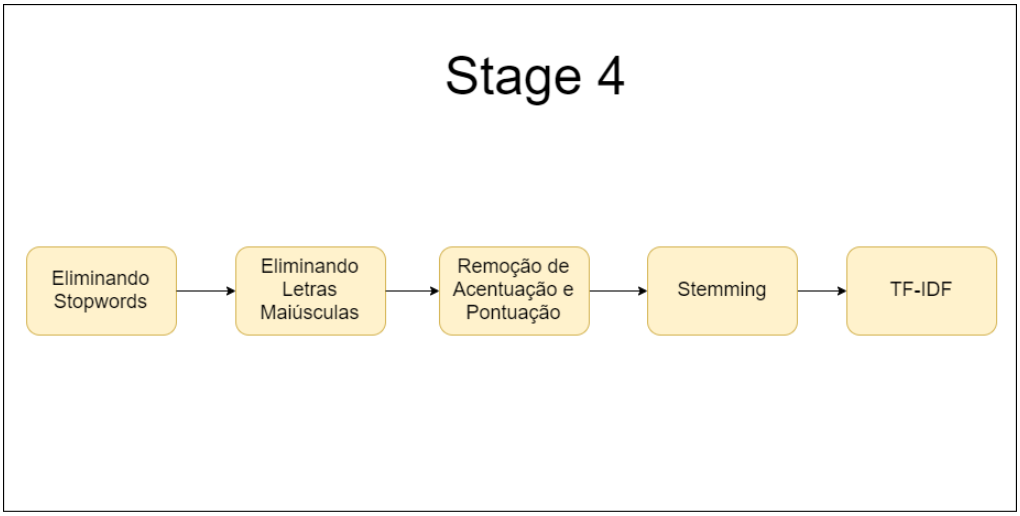


Figura 2.18 - Estrutura do quarto estágio do capítulo.

2.4.4.1. Remoção de *Stop words*

Após gerar os *word clouds* e os gráficos, foi possível ver que os termos de maior relevância associados a um sentimento são palavras que na língua portuguesa possuem uma frequência demasiadamente alta, como artigos, preposições e pronomes. Como esses termos não agregam ao objetivo a ser alcançado, não vale a pena que estes termos estejam contidos no modelo.

A solução é, então, realizar uma limpeza desses termos, sendo denominados *stop words*. Segundo E. Dragut et al. (2009), uma *stop word* é definida por palavras que não possuem um significado semântico relevante para um texto ou nas frases que aparecem. Dessa forma, a remoção das *stop words* do conjunto de palavras total irá facilitar e explicitar quais são os termos de principal relevância no modelo.

Utilizando a biblioteca *Stopwords* da biblioteca NLTK, é possível armazenar em uma variável todas as *stop words* associadas a determinado idioma, no caso desse curso, à língua portuguesa. Fazendo uma varredura das palavras e comparando se o termo presente é igual a uma *stop word*, é possível armazenar todos os termos não irrelevantes em um novo *array*, possibilitando a geração de outra coluna no *dataset* contendo somente os termos relevantes. Os resultados podem ser observados na Figura 2.19.

	text_pt	sentiment	classificacao	tratamento_1
0	Mais uma vez, o Sr. Costner arrumou um filme p...	neg	0	Mais vez, Sr. Costner arrumou filme tempo nece...
1	Este é um exemplo do motivo pelo qual a maiori...	neg	0	Este exemplo motivo maioria filmes ação mesmos...
2	Primeiro de tudo eu odeio esses raps imbecis, ...	neg	0	Primeiro tudo odeio raps imbecis, poderiam agi...
3	Nem mesmo os Beatles puderam escrever músicas ...	neg	0	Nem Beatles puderam escrever músicas todos gos...
4	Filmes de fotos de latão não é uma palavra apr...	neg	0	Filmes fotos latão palavra apropriada eles, ve...

Figura 2.19 - Nova coluna gerada no *dataset*, desta vez sem as *stopwords*.

2.4.4.2. Padronização do formato do texto

Ao gerar um gráfico de barra novamente (Figura 2.20), é possível ver que a palavra mais repetida na análise é “filme”, o que é intuitivo, devido a natureza desse *dataset*. Todavia, é possível ver que termos como “Eu, A, O” ainda estão sendo contabilizados e estão presentes na coluna nova, mesmo estas palavras sendo consideradas *stop words*.

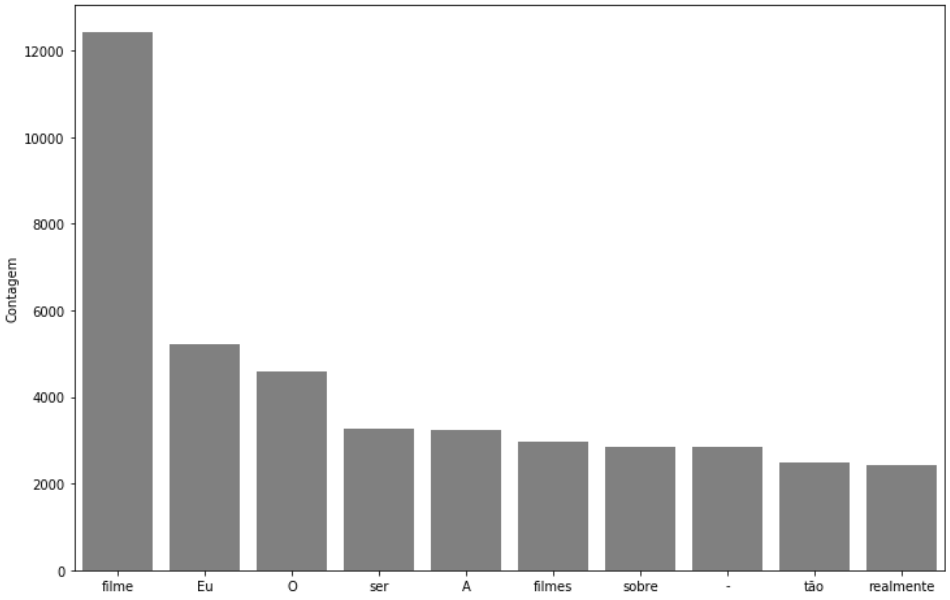


Figura 2.20 - Gráfico de barra contendo a nova frequência de palavras após a remoção dos *stop words.*

O que ocorre nesse caso é que o *array* contendo as *stop words* possui palavras em minúsculo, e a linguagem *Python* é uma linguagem *case sensitive*, que diferencia palavras com caracteres maiúsculos de minúsculos. Dessa forma, é necessário realizar outra varredura prévia antes de eliminar as *stop words,* convertendo todas as palavras em maiúsculo para minúsculo na coluna de comentário.

A segunda etapa de otimização visa então converter todo o texto, através da função nativa do *Python .lower().* Após essa transformação, a remoção das *stop words* pode ser feita novamente, dessa vez removendo todo o restante dos termos previamente em maiúsculo. Os resultados podem ser observados nas Figuras 2.21, 2.22 e 2.23.

tratamento_1	tratamento_2
Mais vez, Sr. Costner arrumou filme tempo nece...	vez, sr. costner arrumou filme tempo necessári...
Este exemplo motivo maioria filmes ação mesmos...	exemplo motivo maioria filmes ação mesmos. gen...
Primeiro tudo odeio raps imbecis, poderiam agi...	primeiro tudo odeio raps imbecis, poderiam agi...
Nem Beatles puderam escrever músicas todos gos...	beatles puderam escrever músicas todos gostass...
Filmes fotos latão palavra apropriada eles, ve...	filmes fotos latão palavra apropriada eles, ve...
Jma coisa engraçada aconteceu comigo enquanto ...	coisa engraçada aconteceu comigo enquanto assi...
Este filme terror alemão ser estranhos vi. Eu ...	filme terror alemão ser estranhos vi. ciente q...
Sendo fã longa data cinema japonês, esperava i...	sendo fã longa data cinema japonês, esperava i...
"Tokyo Eyes" fala menina japonesa 17 anos cai ...	"tokyo eyes" fala menina japonesa 17 anos cai ...
Fazendeiros ricos Buenos Aires têm longa polít...	fazendeiros ricos buenos aires têm longa polít...

Figura 2.21 - Nova coluna a ser inserida no *dataset*, após a transformação de todas as palavras para minúsculo.

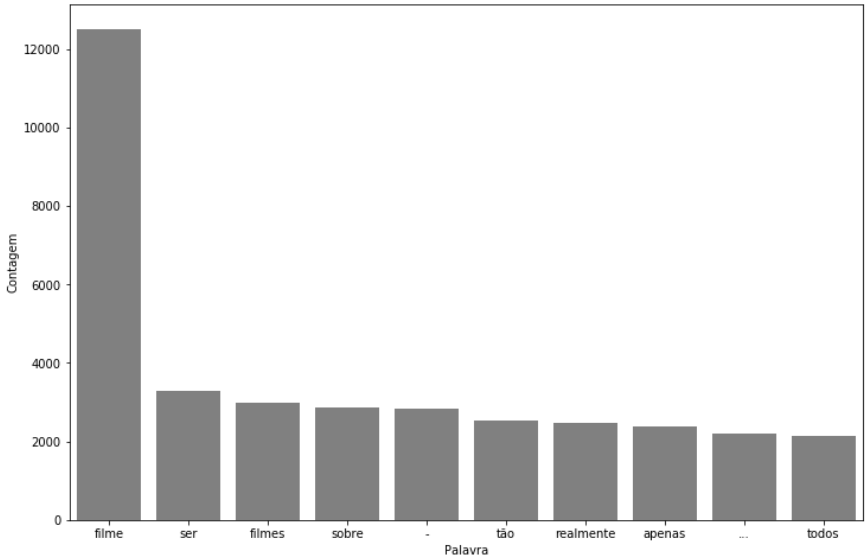


Figura 2.22 – Gráfico de barra gerado após a conversão das palavras no corpus textual para minúsculo.

Ao observar a figura 2.22, alguns problemas podem ser observados. Termos como "**-**" e "**...**" estão sendo sinalizados como os mais frequentes no corpus textual. Como estes termos não são palavras que agregam para o enriquecimento da análise, é conveniente remover todos os sinais de pontuação que permeiam o corpus textual, da mesma maneira que as *stop words* foram removidas.

Outra forma de otimização muito comum durante o pré-processamento do texto do corpus textual é a remoção da acentuação nas palavras fornecidas. Segundo Manning, et al, (2008), a remoção de acentos de palavras em idiomas como o inglês não realiza um grande impacto no significado da palavra, mantendo os seus significados originais e agrupando com sucesso todas as palavras no corpus textual. Todavia, em idiomas como o espanhol por exemplo, essa remoção pode comprometer significativamente no significado da palavra oferecida.

Apesar dos termos presentes na língua portuguesa possuírem palavras com significados diferentes devido à acentuação (como por exemplo as palavras "**secretaria**" e **secretária**", "**amém**" e "**amem**"), esse método é constantemente utilizado por conseguir agrupar palavras que possuem algum erro ortográfico referente a um erro de digitação ("**não**" por "**nao**", por exemplo), sendo assim consideradas válidas na frequência de somente um termo em si.

Dessa forma, ambas estas otimizações serão aplicadas no corpus textual, de maneira a conseguir remover os termos indesejados e agrupá-los com maior eficácia. Os resultados após a terceira otimização podem ser vistos na Figura 2.23.

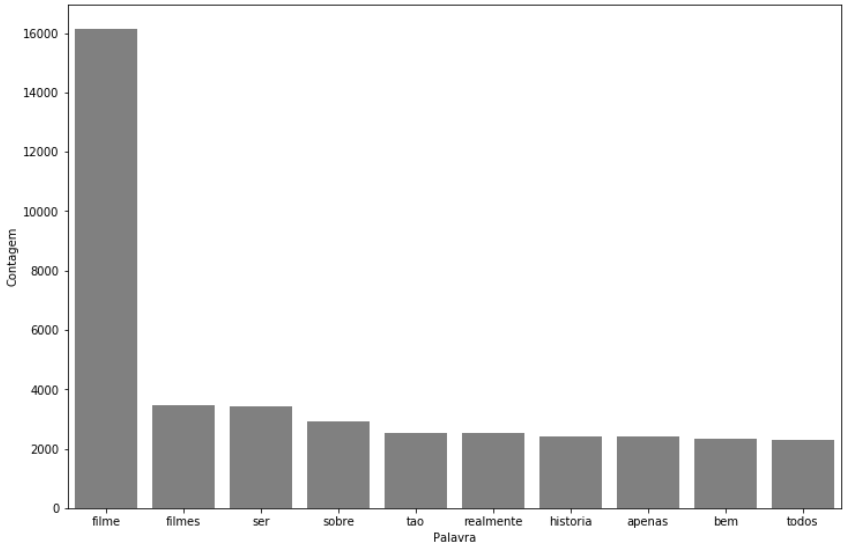


Figura 2.23 – Gráfico de barra gerado após a remoção de termos de pontuação e acentuação

tratamento_1	tratamento_2	tratamento_3
Mais vez, Sr. Costner arrumou filme tempo nece...	vez, sr. costner arrumou filme tempo necessári...	vez sr costner arrumou filme tempo necessário ...
Este exemplo motivo maioria filmes ação mesmos...	exemplo motivo maioria filmes ação mesmos. gen...	exemplo motivo maioria filmes ação mesmos géné...
Primeiro tudo odeio raps imbecis, poderiam agi...	primeiro tudo odeio raps imbecis, poderiam agi...	primeiro tudo odeio raps imbecis poderiam agir...
Nem Beatles puderam escrever músicas todos gos...	beatles puderam escrever músicas todos gostass...	beatles puderam escrever músicas todos gostass...
Filmes fotos latão palavra apropriada eles, ve...	filmes fotos latão palavra apropriada eles, ve...	filmes fotos latão palavra apropriada verdade ...
...	...	...
Como média votos baixa, fato funcionário locad...	média votos baixa, fato funcionário locadora a...	média votos baixa fato funcionário locadora ac...
O enredo algumas reviravoltas infelizes inacre...	enredo algumas reviravoltas infelizes inacredi...	enredo algumas reviravoltas infelizes inacredi...
Estou espantado forma filme maioria outros méd...	espantado forma filme maioria outros média 5 e...	espantado forma filme maioria outros média 5 e...
A Christmas Together realmente veio antes temp...	christmas together realmente veio antes tempo,...	christmas together realmente veio antes tempo ...
O drama romântico classe trabalhadora diretor ...	drama romântico classe trabalhadora diretor ma...	drama romântico classe trabalhadora diretor ma...

Figura 2.24 – Nova coluna no *dataset* gerada após a padronização de texto.

Figura 2.25 - *Word cloud* associado somente a palavras com sentimento positivo gerado após as otimizações

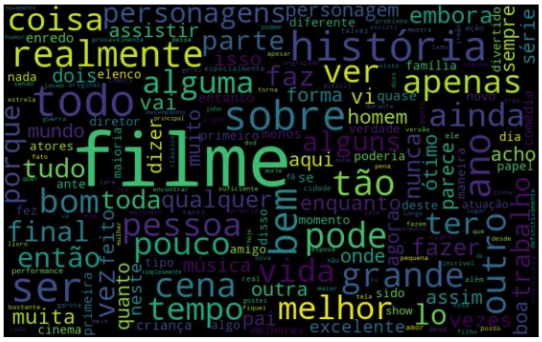


Figura 2.26 - *Word cloud* associado somente a palavras com sentimento negativo gerado após as otimizações.

2.4.4.3. Stemming

Ao verificar os resultados na Figura 2.23, é possível notar um caso peculiar. Após a padronização do corpus textual, a palavra “filme” e “filmes” são os dois termos com maior frequência. Todavia, ambas as palavras possuem o mesmo significado para a análise, de maneira que a segunda é somente uma derivação da primeira no plural. Dessa forma, pode-se dizer que o mesmo significado acaba ocupando duas posições no gráfico de termos mais relevantes, ocupando o espaço de termos com significado único que poderiam estar sendo representados no mesmo.

Ao analisar a estrutura morfológica de uma palavra, é possível ver casos em que a mesma é composta por uma estrutura similar, porém sendo derivada através de diferentes prefixos e sufixos. Estes muitas vezes não são de interesse imediato para realizar a contagem da frequência dos termos, visto que diferentes derivações resultam na contagem de termos distintos. Desta forma, seria conveniente à análise remover os sufixos de todas as palavras contidas no corpus, de maneira a aglutinar todas os termos com a mesma estrutura e significado.

O processo de remoção de todas as flexões das palavras de maneira a reduzi-las à mesma raiz de maneira computacional é denominado *Stemming*. Segundo Lovins (1968), a aplicação de *Stemming* ajuda a maximizar a utilidade dos

termos contabilizados, de maneira a otimizar as palavras agrupadas de acordo com um mesmo significado.

Dessa forma, a próxima etapa de otimização visa transformar todas as palavras do corpus textual e reduzi-las a uma mesma raíz (ou *stem*). Para isso, será utilizada um pacote da biblioteca NLTK denominada RSLPStemmer. O RSLP (Removedor de Sufixos na Língua Portuguesa) será responsável por converter todas as palavras no corpus textual em português brasileiro para seus respectivos radicais morfológicos, de maneira a otimizar a aglutinação dos termos.

Após realizar novamente a contagem dos termos no gráfico de barra (Figura 2.26), é possível notar uma mudança no total dos termos presentes. O novo termo "**film**" possui aproximadamente 20000 repetições, agrupando a frequência dos termos anteriores "**filme**" e "**filmes**". Outros termos, como '**tod**", subiram de posição, agrupando palavras com sufixos como "**todo**", "**todos**", "**todas**", etc. Dessa forma, é possível ter uma análise mais precisa de acordo com o significado real das palavras disponibilizadas no corpus.

tratamento_1	tratamento_2	tratamento_3	tratamento_4
Mais vez, Sr. Costner arrumou filme tempo nece...	vez, sr. costner arrumou filme tempo necessári...	vez sr costner arrumou filme tempo necessario ...	vez sr costn arrum film temp necessari alem te...
Este exemplo motivo maioria filmes ação mesmos...	exemplo motivo maioria filmes ação mesmos. gen...	exemplo motivo maioria filmes acao mesmos gene...	exempl motiv maior film aca mesm gener chat na...
Primeiro tudo odeio raps imbecis, poderiam agi...	primeiro tudo odeio raps imbecis, poderiam agi...	primeiro tudo odeio raps imbecis poderiam agir...	prim tud odei rap imbecil pod agir arm pressio...
Nem Beatles puderam escrever músicas todos gos...	beatles puderam escrever músicas todos gostass...	beatles puderam escrever musicas todos gostass...	beatl pud escrev music tod gost emb walt hill ...
Filmes fotos latão palavra apropriada eles, ve...	filmes fotos latão palavra apropriada eles, ve...	filmes fotos latao palavra apropriada verdade ...	film fot lata palavr apropri verdad tant ous q...
...	...	...	...
Como média votos baixa, fato funcionário locad...	média votos baixa, fato funcionário locadora a...	media votos baixa fato funcionario locadora ac...	med vot baix fat funcionari loc ach tud bem ",...
O enredo algumas reviravoltas infelizes inacre...	enredo algumas reviravoltas infelizes inacredi...	enredo algumas reviravoltas infelizes inacredi...	enred algum reviravolt infeliz inacredita enta...
Estou espantado forma filme maioria outros méd...	espantado forma filme maioria outros média 5 e...	espantado forma filme maioria outros media 5 e...	espant form film maior outr med 5 estrel men f...
A Christmas Together realmente veio antes temp...	christmas together realmente veio antes tempo,...	christmas together realmente veio antes tempo ...	christm togeth real vei ant temp cri john denv...
O drama romântico classe trabalhadora diretor ...	drama romântico classe trabalhadora diretor ma...	drama romantico classe trabalhadora diretor ma...	dram roman cl trabalh dire martin ritt tao ina...

Figura 2.27 – Nova coluna no *dataset* gerada após a remoção de sufixos.

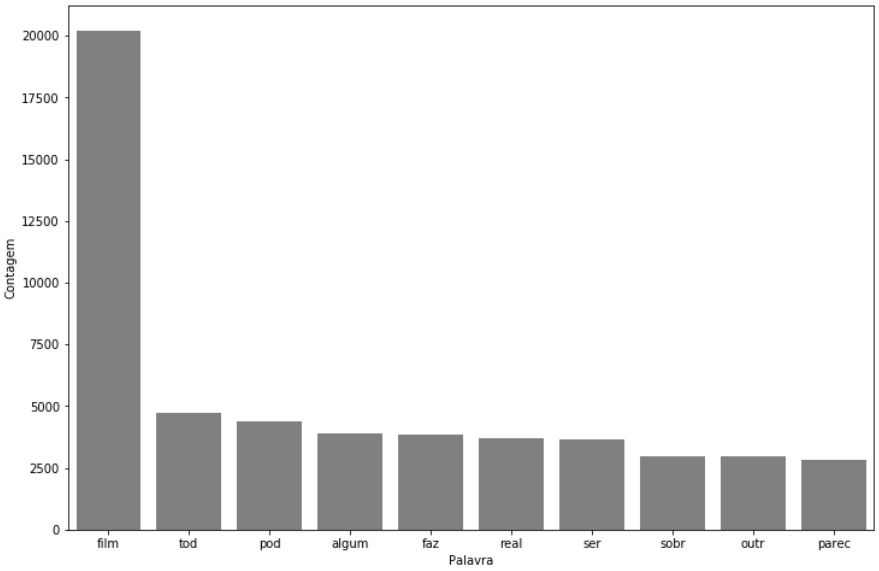


Figura 2.28 – Nova coluna no *dataset* gerada após a remoção de sufixos.

2.4.4.4. TF-IDF

Ao gerar a análise, é possível notar que palavras com maior frequência não são suficientes para demonstrar palavras de maneira polarizada, de modo que sejam diretamente associadas a algum sentimento. Analisando os dois *Word Clouds* gerados nas Figuras 2.25 e 2.26, nota-se que diversas palavras se repetem tanto associadas ao sentimento negativo quanto ao positivo. Dessa forma, é necessário gerar uma nova forma de análise, de modo a observar os termos que mais definem o sentimento na frase.

Uma forma de alcançar esse resultado é partindo do caminho oposto, ou seja, ao invés de analisar os termos mais frequentes, deve-se analisar os termos mais raros. Em frases como **"Eu achei o carro bonito"** e **"Eu achei o carro feio"**, as palavras **"bonito"** e **"feio"** são críticas para definir o sentimento da frase, apesar de estarem em uma frequência muito menor do que as outras palavras no corpus textual. Além disso, estas palavras têm muito mais chance de estarem presentes em frases associadas a somente um sentimento, aumentando assim a acurácia do modelo.

Dessa forma, a última etapa do capítulo será realizar a implementação de uma nova medida estatística que possibilite a análise de acordo com valores equilibrados pela frequência inversa dos termos, mais conhecida como TF-IDF. De acordo com Hiemstra (2000), o valor TF-IDF define que o peso dos termos presentes nos documentos deve ser proporcional à frequência dos termos em todo o corpus textual, e inversamente proporcional à sua frequência total em um documento. Dessa forma, é possível equilibrar a relevância de palavras comuns e evitar algumas palavras de serem mais importantes que outras.

Através da biblioteca SKLearn, é possível implementar essa nova ponderação, que irá equilibrar a frequência de todas as palavras no corpus textual de acordo com a quantidade de aparições. Dessa forma, palavras mais raras possuirão um peso maior,

enquanto palavras mais comuns receberão um peso menor, uniformizando o modelo de aprendizado.

Por fim, será gerado um aprendizado supervisionado da mesma forma feita no Stage 1, de modo a verificar se houve alguma mudança significativa na acurácia gerada. É possível também gerar um novo *dataset*, contendo o peso das palavras de acordo com a sua nova relevância de acordo com os seus valores TF-IDF (Figuras 2.29, 2.30).

	0
otim	3.180694
excel	2.400139
maravilh	1.743315
divert	1.621445
am	1.620834
incri	1.594425
gost	1.503948
favorit	1.487855
perfeit	1.452242
final	1.304025
vid	1.288748
muit	1.287829
mund	1.273158

Figura 2.29 - *Score* obtido de palavras positivas após uma iteração do algoritmo TF-IDF.

	0
ruim	-4.096869
pi	-3.652940
horri	-3.422551
terri	-2.879176
parec	-2.313847
nenhum	-2.205929
nad	-2.121887
chat	-1.995251
estup	-1.793116
minut	-1.743705
mal	-1.684672
dialog	-1.642053

Figura 2.30 - *Score* obtido de palavras positivas após uma iteração do algoritmo TF-IDF.

2.5. Discussão dos Resultados Obtidos

Na **Tabela 2.3** é possível ver todas as acurácias obtidas a partir das execuções dos algoritmos executados durante os estágios do capítulo.

Tabela 2.3 – Acurácias obtidas após os passos realizados durante o capítulo.

DESCRIÇÃO	ACURÁCIA
APÓS REGRESSÃO LOGÍSTICA COM CORPUS TEXTUAL INALTERADO (STAGE 1)	0.6664
APÓS REGRESSÃO LOGÍSTICA COM CORPUS TEXTUAL SEM *STOPWORDS* (STAGE 4)	0.6808
APÓS REGRESSÃO LOGÍSTICA COM CORPUS TEXTUAL SEM *STOPWORDS* E EM LOWERCASE (STAGE 4)	0.664
APÓS REGRESSÃO LOGÍSTICA COM CORPUS TEXTUAL SEM *STOPWORDS*, EM LOWERCASE E COM PONTUAÇÃO E ACENTUAÇÃO REMOVIDA (STAGE 4)	0.6792
APÓS REGRESSÃO LOGÍSTICA COM CORPUS TEXTUAL SEM *STOPWORDS*, EM LOWERCASE, COM PONTUAÇÃO E ACENTUAÇÃO REMOVIDAS E STEMMING APLICADO (STAGE 4)	0.6936
APÓS REGRESSÃO LOGÍSTICA PADRONIZADA COM VALORES TF-IDF, COM CORPUS TEXTUAL SEM *STOPWORDS*, EM LOWERCASE, COM PONTUAÇÃO E ACENTUAÇÃO REMOVIDAS E STEMMING APLICADO (STAGE 4)	0.884

Após realizar algumas otimizações, é possível ver que em alguns casos a acurácia acabou diminuindo. Esses casos são comuns, visto que ao aproximar o corpus textual de padronização mais homogênea muitas vezes pode acarretar numa maior taxa de erro, dependendo do tamanho e da forma do cálculo da acurácia do seu *dataset*.

É importante ressaltar o aumento significativo do aumento da acurácia após a implementação dos valores TF-IDF. A análise dos termos mais raros é muitas vezes uma das formas mais eficazes para obter informações pertinentes referentes ao seu conjunto de informações. Ainda assim, é importante verificar sob diferentes perspectivas, de maneira a expandir o número de conclusões a serem tiradas a partir do seu conjunto de dados.

A aplicação de *Stemming* nas palavras também é um fator importante a ser discutido. Embora o agrupamento dos termos derivados tenha contribuído para o aumento da acurácia em si, a representação das novas palavras definidas somente por sua raiz muitas vezes gera uma considerável perda semântica, tornando o entendimento

das mesmas mais difícil. Dessa forma, a visualização desse novo corpus textual acaba não sendo totalmente intuitiva para a representação gráfica em um *Word cloud*.

Por fim, é importante ressaltar que os resultados são afetados diretamente pelo tamanho de registros no *dataset* e da *seed* fornecida para separar os dados. No caso do capítulo, o número de dados foi consideravelmente menor, de maneira a facilitar o processamento de dados realizando-os de uma maneira mais ágil. De acordo com o número de informações do *dataset*, é possível obter uma classificação mais abrangente, consequentemente aumentando a precisão a ser obtida.

2.6. Conclusão

Este capítulo apresentou e discutiu os conceitos base na área de ciência de dados, de modo a possibilitar um primeiro contato no tema de Análise de Sentimento. Desta forma, foi necessário implementar alguns algoritmos e métodos léxicos de maneira a assegurar uma acurácia significativa para o modelo proposto.

É importante ressaltar que o *dataset* utilizado foi previamente montado e adaptado para facilitar a manipulação dos dados, e assegurar o foco somente no ensino dos conceitos base. Outras bases de dados possuem um número de registros muito maior, necessitando de mais tempo de processamento para a execução dos comandos. Além disso, estes necessitam de um tratamento prévio para possibilitar o uso de algoritmos como o de regressão logística e afins.

Além disso, existem outras formas de conseguir alcançar o mesmo resultado obtido, porém através de abordagens diferentes. Neste capítulo, foi abordado a técnica de Aprendizado Supervisionado, porém é possível obter resultados similares utilizando Aprendizados Semi ou Não Supervisionados. Dessa forma, é essencial aprender sobre as necessidades e recursos que seu projeto irá possuir, de modo a implementar o método que melhor se encaixe para o sucesso do mesmo.

Referências

Agarwal, B., Mittal, N., Bansal, P., & Garg, S. (2015). *Sentiment Analysis Using Common-Sense and Context Information. Computational Intelligence and Neuroscience, 2015, 1–9.*

Benevenuto, F., Ribeiro, F. and Araújo, M. (2015) *Métodos para Análise de Sentimentos em Mídias Sociais.*, Short course in the Brazilian Symposium on Multimedia and the Web (Webmedia).

Bing, Liu (2010) Sentiment Analysis and Subjectivity , 2nd edn., Handbook of Natural Language Processing.

Ceci, F., Alvarez, G., Gonçalves, A. (2017) *Análise de sentimento e mineração de opinião: Uma revisão bibliométrica da literatura*, Revista Espacios, Vol. 38 (Nº 14).

Chung, T. and Gildea, D. (2009) *Unsupervised tokenization for machine translation*, EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009: .

Dragut, E., Fang, F., Sistla, P., Yu, C. and Meng, W. (2009) *Stop word and related problems in web interface integration*, Proceedings of the VLDB Endowment.

Dumais, S., Platt, J., Hecherman, D. & Sahami, M. (1998) *Inductive Learning Algorithms and Representations for Text Categorization*, Proceedings of the seventh international conference on Information and knowledge management: .

Franz, M., Hillman, J. (2016) *A Tipologia de Jung - Ensaios sobre Psicologia Analítica*, 2nd edn., Cultrix.

Heimerl, F., Lohmann, S., Lange S. and Ertl, T. (2014) *Word cloud explorer: Text analytics based on word clouds*, Institute for Visualization and Interactive Systems (VIS)

Hiemstra, D. (2000). *A probabilistic justification for using tf×idf term weighting in information retrieval. International Journal on Digital Libraries, 3(2), 131–139.*

Hilbe, J. M. (2009). *Logistic regression models*. Chapman and hall/CRC.

Horta, E. G. (2015) *Aplicação de Máquinas de Aprendizado Extremo ao Problema de Aprendizado Ativo.* Programa de PósGraduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, UFMG.

Liddy, E.D. (2001) *Natural Language Processing*, 2nd edn., Encyclopedia of Library and Information Science: Marcel Decker, Inc.

Lima, C. (2008) *O uso da leitura de imagens como instrumento para a alfabetização visual.* Cadernos PDE, Vol. II. Curitiba.

Lovins, J. (1968) *Developing of a Stemming Algorithm* - Mechanical Translation and Computational Linguistics

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Matsubara, E., Martins, C., Monard, M. (2003) *PreTexT: uma ferramenta para pre-processamento de textos utilizando a abordagem bag-of-words*., Relatórios Técnicos do ICMC - São Carlos.

Mohammad, S. and Turney, P. (2013) *Crowdsourcing a Word–Emotion Association Lexicon*, Institute for Information Technology, National Research Council Canada: .

Monard, M. and Baranauskas, J. (2003) *Conceitos sobre aprendizado de máquina*, 1.1 edn., Sistemas inteligentes-Fundamentos e aplicações.

Reis, J.; Benevenuto, F.; de Melo, P. V.; Prates, R.; Kwak,H.; and An, J. (2015). *Breaking the news: First impressionsmatter on online news*. In Proceedings of the ICWSM.

Reis, J.; Gonçalves, P.; Vaz de Melo, P. O.; Prates, R.; and Benevenuto, F. 2014. *Magnet news: You choose the polarity of what you read*. Proceedings of ICWSM

Sebastiani, F (2002) *Machine learning in automated text categorization*, ACM Computing Surveys.

Serrano-Guerrero, J., Olivas, J., Romero, F. Herrera-Viedma, E. (2015) *Sentiment analysis: A review and comparative analysis of web services*, 2nd edn., Information Sciences, v. 311, p. 1838.

Zhang, K., Cheng, Y., Liao, W., & Choudhary, A. (2012). *Mining millions of reviews. Proceedings of the 13th International Conference on Electronic Commerce - ICEC '11*.

Biografia Resumida dos Autores

André Viana Tardelli

André é graduando do curso de Ciência de Computação na Universidade Federal do Rio de Janeiro. Atualmente atua como instrutor e desenvolvedor no Grupo Caelum, ministrando cursos presenciais nos temas de Front End e Data Science. Seu foco de pesquisa, atualmente contendo artigos apresentados em conferências nacionais e internacionais, envolve implementar conceitos de psicologia aplicados à tecnologia, buscando novas formas de humanizar as interações realizadas de maneira digital.

Lattes: http://lattes.cnpq.br/8627393261758849

Angélica Fonseca da Silva Dias

É doutora em Informática pela Universidade Federal do Rio de Janeiro (PPGI - 2018) com ênfase em Gestão de Sistemas Complexos. Mestre em Sistemas de Informação pela UFRJ. MBA em Gestão Executiva e E-Business pela COPPEAD e Inteligência e Database Marketing, Aperfeiçoamento em Gerência Avançada de Projetos/NCE/UFRJ. Graduação em Processamento de Dados/UNESA. Foi Diretora da Área de Extensão do Instituto Tércio Pacitti/UFRJ, Coordenadora Acadêmica dos Cursos de pós-graduação na UFRJ e atua como Professor Convidado no programa de pós-graduação em informática, Instituto de Economia e HCTE da UFRJ com a orientação e coorientação de alunos de graduação, pós-graduação e mestrado. Tem experiência nas áreas de Administração Pública, Gerência de Projetos, Ciência da Computação e Educação. Temas de interesse: Gestão de Conhecimento, Social Computing, Economia Circular Computacional, Data Science, Data Literacy, Trabalho e Aprendizagem Cooperativa apoiada por computador (CSCW e CSCL), Tecnologia Assistiva, Gestão Estratégica da Informação e Educação a distância. Lattes: http://lattes.cnpq.br/8795875378897586

Juliana Baptista dos Santos França

É doutora em Informática pela Universidade Federal do Rio de Janeiro (PPGI/UFRJ - 2018) com ênfase em Gestão de Sistemas Complexos. Finalizou seu Pós doutorado na UFRJ na área de CSCW (PPGI/UFRJ - 2018). Mestre em Informática pelo Programa de Pós-Graduação em Informática (PPGI/UNIRIO - 2012) da Universidade Federal do Estado do Rio de Janeiro, com ênfase na linha de pesquisa Sistemas de Apoio a Negócios. Possui graduação em Sistemas de Informação pela Universidade Federal do Estado do Rio de Janeiro (UNIRIO), e também graduação em Engenharia Cartográfica pela Universidade do Estado do Rio de Janeiro (UERJ). Atualmente é Professora Adjunto na Universidade Federal Rural do Rio de Janeiro (UFRRJ) junto ao Departamento de Computação (DECOMP) em Banco de Dados e atua como colaboradora no programa de pós-graduação em informática da UFRJ com a coorientação de alunos de mestrado. Atuou na organização de eventos científicos nacionais e internacionais como ISCRAM 2016, Summer School em IHC/CSCW 2019, e SBSC 2019. Tem experiência na área de Sistemas de Informação e atua principalmente nos seguintes temas: Colaboração (CSCW), Gestão de Conhecimento, Processos de Negócio, Aprendizagem Colaborativa, Suporte à Decisão e Modelagem Conceitual e Ontológica. Lattes: http://lattes.cnpq.br/9341068095520817

Capítulo 3

LGPD em Ambientes de Bancos de Dados nas Organizações

Ana Carolina Brito de Almeida, Letícia Dias Verona, Maria Luiza Machado Campos e Fernanda Araujo Baião

Abstract

Information Security has been an especially relevant topic for the development of Information Systems (IS) in organizations. In 2018, it became even more crucial as the Law 13.709 (LGPD) was passed in Brazil, giving organizations more responsibility for the collection, processing and protection of personal data. It is well known that much of the corporate information is stored in repositories under the management of Database Management Systems (DBMS). In this context, this course addresses the theme of LGPD in database systems (DB) environments within organizations. The course includes a discussion of the two crucial concepts of Data Security and Data Privacy; presents an overview of LGPD, exemplifying its ten principles; and discusses the operational support of some DBs to the principles advocated by such Law.

Resumo

A Segurança da Informação é um tema especialmente relevante para desenvolvimento de Sistemas de Informação (SI) nas organizações. Em 2018, tornou-se ainda mais crucial, pois foi sancionada a Lei 13.709 (LGPD), no Brasil, que atribui mais responsabilidade às organizações quanto à coleta, ao tratamento e à proteção dos dados pessoais. É notório que grande parte das informações corporativas estão armazenadas em repositórios sob a gestão de Sistemas Gerenciadores de Bancos de Dados (SGBD). Neste contexto, o presente minicurso aborda, o tema de LGPD em ambientes de bancos de dados (BDs) nas organizações. O minicurso discute os conceitos de Segurança e Privacidade; apresenta uma visão geral da LGPD, exemplificando seus dez princípios; e discute o suporte operacional em BDs aos princípios preconizados na tal Lei.

3.1. Introdução

A Segurança da Informação é uma preocupação constante durante o desenvolvimento de Sistemas de Informação. Tal área obteve ainda mais visibilidade com a sanção da LGPD,

a Lei Geral de Proteção de Dados [Brasil 2018]. Essa Lei é baseada no Regulamento Geral sobre a Proteção de Dados 2016/679 (RGPD, ou, como é mais conhecida em inglês, GDPR - General Data Protection Regulation), elaborado pela União Europeia. A Lei obriga organizações a seguirem uma série de itens quanto à coleta, ao tratamento e à proteção dos dados pessoais. A RGPD tem como foco proteger direitos fundamentais de liberdade e privacidade dos indivíduos, complementando regulamentações previamente existentes na Convenção Europeia de Direitos Humanos e impondo maiores obrigações aos agentes privados detentores de dados pessoais. A lei brasileira foi aprovada em 2018 e ainda necessita ser regulamentada para sua entrada em vigor em 2020. O seu espectro de atuação, bastante amplo e ainda muito subjetivo, contempla o direito ao cidadão de impedir a divulgação ou posse de qualquer dado pessoal a seu respeito, isolado ou agregado estatisticamente. Em ambas, é destacada a necessidade de consentimento explícito para coletar os dados pessoais e transparência total sobre o que será feito a partir deles.

Em tempos de internet das coisas (IoT), redes sociais e aplicativos móveis, é importante que haja atenção tanto por parte dos usuários quanto ao cadastro e à transferência de seus dados quanto pelas empresas que detém tais dados, de forma a protegê-los de vazamentos. Por exemplo, quando um usuário se cadastra em um aplicativo móvel de corrida, que registra o tempo e a distância que ele percorre, a empresa desenvolvedora do aplicativo não pode enviar os dados coletados para uma empresa de plano de saúde, suplemento alimentar ou de marcas esportivas sem o consentimento do usuário.

Usualmente, os dados de usuários são mantidos em repositórios corporativos nas organizações. Além disso, a maioria das organizações armazena os dados pessoais coletados dos clientes em Sistemas de Gerenciamento de Banco de Dados (SGBD). Dessa forma, é necessário saber como e o quanto as empresas detentoras dos principais SGBDs comerciais estão preparadas para dar suporte à implantação de estratégias da LGPD de forma eficaz e eficiente.

Tradicionalmente, os SGBDs dos principais fornecedores de mercado dispõem de recursos para prover segurança da informação em seus produtos. Mais recentemente, no entanto, este aspecto vem se ampliando para tratar da privacidade, também motivado por este cenário recente em que a RGPD se insere. Neste sentido, a Oracle disponibiliza uma série de pacotes para aumentar a segurança dos dados armazenados[1], tais como: *Oracle Advanced Security, Oracle Key Vault, Oracle Data Masking and Subsetting* etc. Tais pacotes oferecem suporte à: criptografia de dados transparente; gerenciamento de chave de criptografia; controle de acesso multifatores e usuários com privilégios; classificação e descoberta de dados; monitoramento e bloqueio de atividades de banco de dados (BD); auditoria e relatórios consolidados; e mascaramento de dados. Já a Microsoft disponibilizou um *guia*[2] de conformidade com o RGPD, mencionando como ferramentas do SGBD SQL Server 2017 podem auxiliar neste sentido, visando auxiliar estratégias de implantação da LGPD. Alguns exemplos de ferramentas são o *Microsoft Compliance Manager*, uma solução que permite aos clientes que trabalham com nuvem gerenciar sua

[1] https://www.oracle.com/database/security/

[2] https://info.microsoft.com/sql-server-gdpr-ebook-registration.html

própria conformidade e o *Data Discovery and Classification*, uma ferramenta para descobrir, classificar, rotular e relatar os dados sensíveis nos BDs dos usuários.

O objetivo do minicurso é possibilitar uma visão geral da LGPD tanto com a perspectiva de usuário quanto com a perspectiva de um administrador de BD, exemplificando recursos de alguns dos principais SGBDs de mercado. O conteúdo está estruturado em tópicos principais, descritos a seguir:

- ✓ **Segurança e Privacidade:** apresentação e discussão dos conceitos de segurança e de privacidade.

- ✓ **Introdução e visão geral da Lei Geral de Proteção de Dados Pessoais (LGPD):** apresentação dos principais conceitos envolvidos na Lei, as boas práticas indicadas e exemplos de aplicação dos dez princípios sobre o tratamento de dados pessoais com o uso de aplicativos de celular, incluindo redes sociais.

- ✓ **LGPD e mecanismos de segurança nos SGBDs:** apresentação das funcionalidades encontradas nas principais empresas desenvolvedoras de SGBD do mercado para contemplar os tópicos da LGPD: anonimização de dados e criptografia, notificação de vazamento de dados, recursos de auditoria. Apresentação de exemplos práticos, ilustrando diversas situações reais e frequentemente encontradas no dia a dia de uma empresa, explorando os mecanismos de proteção apresentados e como eles atendem essas situações. Discussão do estado da arte, oportunidades e desafios.

3.2. Segurança e Privacidade

É importante destacar que, embora estejam muito relacionados e exista uma sobreposição considerável entre questões relacionadas ao acesso a recursos (segurança) e questões relacionadas ao uso de informações (privacidade), existem diferenças importantes entre os conceitos de segurança e privacidade.

Elmasri e Navathe (2019) diferenciam bem esses conceitos, definindo que Segurança na Tecnologia da Informação refere-se a muitos aspectos da proteção de um sistema do uso não autorizado, incluindo autenticação de usuários, criptografia de informações, controle de acesso, políticas de *firewall* e detecção de intrusões. Para o propósito do minicurso, limitaremos nosso tratamento de Segurança aos conceitos associados a quão bem um sistema pode proteger o acesso às informações que ele contém, incluindo a integridade e disponibilidade dessas informações.

O conceito de privacidade, por sua vez, transcende o aspecto de segurança e de fato vem sendo tratado como parte de uma discussão mais ampla sobre Transparência em cenários recentes da 4a Revolução Industrial (*Fourth Industrial Revolution*, ou 4IR), caracterizados por uma fusão de tecnologias e eliminando fronteiras entre os meios físico, digital e biológico [Teixeira et al, 2019]. A privacidade examina até que ponto o uso de informações pessoais que o sistema detém sobre um usuário está em conformidade com as suposições explícitas ou implícitas em relação a esse uso. Do ponto de vista do usuário final, a privacidade pode ser considerada a partir de duas perspectivas diferentes: impedir o armazenamento de informações pessoais e garantir o uso apropriado de informações

pessoais. Para o presente minicurso, a ideia básica é discutir os mecanismos que os Sistemas de Gerenciamento de Banco de Dados oferecem para garantir o uso apropriado de informações pessoais por eles armazenadas.

Na indústria, a privacidade realmente se concentra nos seguintes conceitos [Dean 2017]:

- ✔ Quais dados devem ser coletados?
- ✔ Quais são os usos permitidos?
- ✔ Com quem isso pode ser compartilhado?
- ✔ Por quanto tempo os dados devem ser retidos?
- ✔ Qual modelo de controle de acesso granular é apropriado?

De uma forma resumida, os controles de segurança são criados para controlar quem pode acessar as informações, enquanto a privacidade é mais granular, controlando quais dados específicos eles podem acessar, e quando. Dean [2017] exemplifica os dois conceitos em um cenário: Se você deposita em uma instituição financeira nacional, todos os caixas do país podem ser provisionados (ou seja, acesso de segurança concedido) para acessar os detalhes da sua conta. Isso fornece a flexibilidade para um cliente visitar uma filial em sua cidade natal, uma filial na costa oeste durante uma viagem de negócios ou uma filial da Flórida durante as férias. Mas a privacidade é outra camada. Embora o caixa possa ser provisionado para exibir todos os detalhes da conta dos clientes, a privacidade só permite acesso quando existe uma necessidade comercial; como um cliente entrando em uma filial em outra cidade para acessar suas contas. Mas a privacidade não permite que o mesmo caixa veja o saldo da conta de seus vizinhos ou talvez o saldo de uma pessoa famosa, apenas porque eles estão interessados - apesar de seus privilégios de acesso lhes concederem acesso.

Portanto, a aplicação comercial dos termos privacidade e segurança é muito diferente, mas com uma certa sobreposição. Segundo Dean (2017), "Você não pode ter privacidade sem segurança, mas pode ter segurança sem privacidade".

3.3. Lei Geral de Proteção de Dados Pessoais (LGPD)

A Lei Geral de Proteção de Dados (Lei 13.709/18), alterada pela Lei 13.853/19, é uma lei brasileira fortemente inspirada pelo RGPD europeu, e que dispõe sobre o tratamento de dados pessoais com objetivo de proteger a liberdade e a privacidade dos cidadãos [Brasil 2018]. A lei abrange toda atividade que envolve coleta, tratamento e armazenamento de dados pessoais, seja ela praticada por entes privados ou públicos no Brasil, inclusive empresas internacionais com atividade no país.

3.3.1. Principais conceitos e princípios

O conceito de dado pessoal, conforme descrito na lei, significa qualquer informação, que individualmente ou combinada com outras, possa identificar uma pessoa ou submetê-la a

um tratamento específico. Como exemplos de dados pessoais podem ser citados nome, CPF, endereço, *cookies* gravados em computadores pessoais, informações partilhadas em redes sociais, dados financeiros ou qualquer informação que permita a identificação de um indivíduo. A lei estabelece ainda o conceito de dado pessoal sensível como sendo dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural.

A LGPD determina que o titular do dado, ou seja, a pessoa natural a quem se referem os dados objeto da coleta, tratamento e armazenamento, tem direito a saber como seus dados estão sendo tratados, a ter acesso aos mesmos, corrigi-los, pedir a sua exclusão, correção e revogar o consentimento ao seu uso. Pode ainda solicitar a informação de quem teve acesso aos seus dados através de atividades compartilhadas com o controlador, que vem a ser a entidade pública ou privada a cujos interesses o processamento dos dados é submetido.

Segundo a LGPD, toda a atividade de tratamento de dados deve obedecer aos seguintes princípios:

- ✓ Princípio 1 - finalidade: realização do tratamento para propósitos legítimos, específicos, explícitos e informados ao titular, sem possibilidade de tratamento posterior de forma incompatível com essas finalidades;
- ✓ Princípio 2 - adequação: compatibilidade do tratamento com as finalidades informadas ao titular, de acordo com o contexto do tratamento;
- ✓ Princípio 3 - necessidade: limitação do tratamento ao mínimo necessário para a realização de suas finalidades, com abrangência dos dados pertinentes, proporcionais e não excessivos em relação às finalidades do tratamento de dados;
- ✓ Princípio 4 - livre acesso: garantia, aos titulares, de consulta facilitada e gratuita sobre a forma e a duração do tratamento, bem como sobre a integralidade de seus dados pessoais;
- ✓ Princípio 5 - qualidade dos dados: garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados, de acordo com a necessidade e para o cumprimento da finalidade de seu tratamento;
- ✓ Princípio 6 - transparência: garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial;
- ✓ Princípio 7 - segurança: utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão;
- ✓ Princípio 8 - prevenção: adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais;
- ✓ Princípio 9 - não discriminação: impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos;

- ✓ Princípio 10 - responsabilização e prestação de contas: demonstração, pelo agente, da adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas.

O acesso aos dados pessoais é normatizado e somente pode ser realizado se o titular dos dados der seu consentimento explícito para realização de atividades específicas, a não ser que o tratamento se enquadre em: cumprimento de leis, tanto da atividade privada do controlador quanto da administração pública; realização de estudos por órgão de pesquisa; exercício de direitos contratuais ou judiciais; quando necessário para a execução de contrato ou de procedimentos preliminares relacionados a contrato do qual seja parte o titular; proteção da vida ou integridade física do titular ou terceiros; tutela da saúde do titular; garantia de prevenção à fraude ou segurança do titular e para a proteção do crédito.

3.3.2. A genealogia da LGPD

Em 2010, foi realizada no Brasil a primeira consulta pública a respeito da proteção de dados. Em 2014, entra em vigor o decreto do Marco Civil da Internet que, dentre os seus objetos, inclui a privacidade do usuário da Internet no Brasil. De 2014 a 2018, diversos projetos de lei se referiram ao tema. Destes, a criação do cadastro positivo que previu a criação de um banco de dados para que as instituições financeiras facilitem o acesso ao crédito a bons pagadores englobou uma ampla discussão sobre privacidade e por fim permitiu que o titular dos dados solicite sua exclusão e esquecimento. Em agosto de 2018, estes projetos de lei foram consolidados na LGPD. No fim do mesmo ano, uma Medida Provisória (869/2018) vetou alguns artigos e criou a Autoridade Nacional de Proteção de Dados (ANPD) com o objetivo de fiscalizar e regulamentar a aplicação da LGPD. As funções e procedimentos desta agência e como se dará sua atuação ainda são desconhecidos para o grande público.

Os fatores decisivos para a concretização e publicação da lei foram o interesse do governo brasileiro de ingressar na OCDE. Um dos requisitos para a entrada no grupo é que o país possua uma lei geral de proteção de dados que permita discutir questões comerciais entre países e a publicação do RGPD europeu, em cujas bases a lei brasileira se assenta. Empresas brasileiras com subsidiárias fora do país, com clientes e fornecedores europeus, e mesmo as que poderiam ter dados de um cidadão europeu na sua base, passaram a se preocupar com o cumprimento do regulamento e com a necessidade de segurança jurídica interna.

Existem, entretanto, diferenças importantes entre a lei brasileira e o regulamento europeu. Em relação à aplicação da lei, as sanções europeias são ágeis e as multas substanciais. Empresas multinacionais de *software*, de serviços *on-line* e de transporte foram multadas por autoridades europeias em dezenas de milhões de dólares por expor indevidamente dados dos seus usuários. As multas brasileiras são limitadas a 2% do faturamento bruto da empresa, o que pode causar uma desproporcionalidade entre o lucro obtido pelo controlador dos dados e o dano causado ao seu titular.

Em termos conceituais, a RGPD afirma que o dado só deve ser usado para o propósito limitado para o qual foi coletado. Já a legislação brasileira admite o uso dos

dados para o legítimo interesse do controlador, o que pode ser conflitante com os direitos do titular e múltiplas circunstâncias. Além disto a LGPD considera a possibilidade da transferência dos dados para outro controlador, o que não é previsto no regulamento europeu sem o consentimento explícito do titular. A redação da lei brasileira permite um amplo espectro de interpretações e questionamentos e alguns são apontados na seção a seguir.

3.3.3. Questões importantes a serem respondidas

A anonimização de dados ou a solicitação de consentimento é uma prática comum na realização de pesquisas científicas e outros trabalhos baseados em dados. Ainda assim, os termos da lei, se não esclarecidos, podem causar uma insegurança jurídica em instituições de pesquisa e entidades governamentais. A definição de dado pessoal como todo dado que pode identificar unicamente uma pessoa natural faz com que, se considerarmos sua combinação com outro dado, possa abranger qualquer informação, ainda que para o controlador original e com interesses legítimos de pesquisa seja um dado anonimizado.

No campo da saúde, a questão é agravada pela definição da lei considerar dados biológicos, genéticos e relativos à saúde como dados pessoas sensíveis, que possuem um grau maior de severidade na aplicação da lei. As pesquisas relacionadas à saúde, em muitos estudos, envolvem a ampla discussão de um caso específico, os protocolos de tratamento adotados e os resultados obtidos. Esses dados podem alegadamente identificar unicamente uma pessoa natural e impedir o seu uso para fins de avanço da ciência.

Do outro lado do espectro de interesses, a inclusão de proteção ao crédito na área financeira, como uma das possibilidades de viabilização de coleta e tratamento de dados sem consentimento, cria uma vulnerabilidade ao titular, pois esses dados, a exemplo de adimplência e contratos de empréstimos, são itens de privacidade importantes para o cidadão.

Ainda sob a ótica das lacunas da lei que podem ser utilizadas em prol das empresas está o conceito de legítimo interesse. Em uma sociedade de livre mercado, o lucro é um interesse legítimo de uma empresa e esse argumento pode justificar a coleta sem consentimento - e sem conhecimento - de dados pessoais com o intuito de direcionar publicidade, o que em última instância implica também em manipulação de emoções, desejos e orientações políticas.

Por fim, a privacidade de agentes do poder público deve ser equilibrada com o interesse civil em fiscalizar suas ações e os limites entre a lei de transparência e a LGPD podem ser fluidos e, por essa razão, devem ser explicitados.

A LGPD, no momento da publicação deste capítulo, ainda carece de regulamentação e muitos conceitos e aplicações possuem lacunas de entendimento. A necessidade de adaptação de empresas e entes públicos, entretanto, é notória e urgente para que possam atender os requisitos mínimos da lei. Esta necessidade gera uma demanda de entendimento e conhecimento dos profissionais e cientistas da área da Ciência da Computação e áreas correlatas, bem como uma adaptação administrativa da maioria das empresas com negócios no país.

As seções a seguir objetivam fornecer um panorama das ferramentas e possibilidades existentes em alguns dos principais SGBDs existentes, com relação aos aspectos de segurança e seus impactos nas questões de privacidade.

3.4. LGPD e mecanismos de segurança nos SGBDs

A presente seção descreve como as principais empresas desenvolvedoras de ambientes de Bancos de Dados, e seus respectivos SGBDs, buscam auxiliar as organizações na adaptação à LGPD.

Com base em um estudo sobre a RGPD [Rajasekharan 2017], levantamos que os principais requisitos de segurança de dados da LGPD podem ser amplamente classificados em três categorias: **avaliação**, **prevenção** e **monitoramento/detecção**.

A **avaliação** está relacionada ao impacto na proteção de dados quando certos tipos de processamento de dados pessoais provavelmente apresentarão um "alto risco" para o titular dos dados. A avaliação deve incluir uma avaliação sistemática e abrangente dos processos, perfis da organização e como essas ferramentas salvaguardam os dados pessoais (LGPD - Capítulo VII - Seção II - Art. 50. § 2º - Letra d).

Em relação à **prevenção** de brechas de segurança, a própria LGPD recomenda algumas técnicas para prevenir os ataques. São elas: anonimização e pseudonimização, controle de acesso de usuário privilegiado, controle de acesso refinado e minimização de dados. A anonimização de dados é a técnica de embaralhar ou ofuscar completamente os dados e a pseudonimização refere-se à redução da vinculação de um conjunto de dados com a identidade original de um titular de dados. A LGPD afirma que as técnicas de anonimização e pseudonimização podem reduzir o risco de divulgação acidental ou intencional de dados, tornando as informações não identificáveis para um indivíduo ou entidade (LGPD - Capítulo II - Seção II - Art. 13.). O controle de acesso de usuário privilegiado que têm acesso aos dados pessoais deve impedir ataques de informações privilegiadas e contas de usuário comprometidas (LGPD - Capítulo VI - Seção I). Além do controle privilegiado do usuário, a LGPD recomenda a adoção de uma metodologia refinada de controle de acesso para garantir que os dados pessoais sejam acessados seletivamente e apenas para uma finalidade definida. Esse tipo de granulação fina do controle de acesso pode ajudar as organizações a minimizar o acesso não autorizado aos dados pessoais (LGPD - Capítulo II - Seção I - Art. 10 - § 1º). A minimização de dados diz respeito à recomendação de minimizar a coleta e retenção de dados pessoais para reduzir o limite de conformidade. Ao coletar, processar ou compartilhar dados pessoais, é necessário ser frugal e limitar a quantidade de informações às necessidades de uma atividade específica (Capítulo I - Art. 6º - III).

O **monitoramento/detecção** de brechas é necessário porque nenhuma organização, mesmo com a adoção de medidas preventivas de segurança, consegue eliminar totalmente a possibilidade de uma violação de dados. A LGPD recomenda esse monitoramento e o alerta para detectar tais violações através dos seguintes mecanismos: dados de auditoria e monitoramento e alerta oportuno. A LGPD exige não apenas o registro ou a auditoria das atividades nos dados pessoais, mas também recomenda que esses registros devem ser mantidos centralmente sob a responsabilidade do controlador (LGPD - Capítulo VI - Seção I - Art. 37.). Por fim, o monitoramento constante das atividades de dados pessoais é fundamental para detectar anomalias. A LGPD também

exige notificações oportunas em caso de violação (LGPD - Capítulo VII - Seção I - Art. 48.).

Além disso, a LGPD também exige conformidade com os princípios de proteção de dados para aprimorar a qualidade e o rigor da proteção dos dados. Entre os dez princípios da LGPD, destacam-se três deles relacionados aos ambientes de SGBDs:

- ✓ Princípio 7 – Segurança: proteger os dados armazenados;
- ✓ Princípio 8 – Prevenção: coibir vazamento de dados;
- ✓ Princípio 10 – Responsabilização: o agente de tratamento de dados pessoais deve demonstrar quais medidas foram adotadas para evitar o vazamento de dados (auditoria).

Diante da categorização ampla dos requisitos de segurança da LGPD e desses três princípios que impactam diretamente os ambientes de SGBDs, investigam-se soluções de diversas naturezas que possam diminuir a vulnerabilidade dos dados armazenados nas organizações. Essas soluções envolvem tanto funcionalidades diretas do SGBD e ferramentas associadas, quanto serviços, disponibilizados em plataformas na nuvem, que atuam como uma camada, abrangendo não só os SGBDs como também as demais aplicações executadas na mesma plataforma.

Baseado no quadrante mágico da Gartner de 2018 (Figura 3.1), decidiu-se investigar três principais empresas (Microsoft, Oracle, Amazon Web Services), além de um SGBD gratuito, o PostgreSQL.

Figura 3.1. Quadrante mágico para Sistemas Gerenciadores de Bancos de Dados Operacionais [Feinberg et al, 2018]

3.4.1. Microsoft

A Microsoft dispõe de diversos aspectos de segurança para auxiliar os usuários de SGBDs na adaptação à LGPD, controlando o acesso, prevenindo e detectando intrusos e vulnerabilidades e gerando relatórios de auditoria.

Seguindo a categorização ampla dos requisitos de segurança da LGPD, alguns aspectos propostos pela Microsoft são [Microsoft 2018a]:

- Avaliação: *data discovery and classification* e *sql vulnerability assessment*.
- Prevenção: *dynamic data masking (DDM), static data masking, sql server authentication, object-level permissions, role-based security, row-level security, transport layer security (TLS), transparent data encryption (TDE)* e *always encrypted*.
- Monitoramento/detecção: *sql server audit, sql server temporal tables* e *sql vulnerability assessment*.

3.4.1.1 Microsoft - Categoria de Avaliação

Toda organização possui um grande volume de dados, incluindo dados pessoais (e, particularmente, também dados sensíveis que devem ser protegidos segundo a LGPD). Diante do grande volume de dados, é importante que o controlador tenha o auxílio de uma ferramenta que o ajude a avaliar, identificar e categorizar os dados pessoais.

A *feature* ***data discovery and classification*** ajuda a organizar e classificar os dados para garantir o manuseio adequado e o melhor gerenciamento de informações pessoais [Microsoft 2019a]. Essa *feature* é acessada através da ferramenta *SQL Server Management Studio* (SSMS), ao selecionar um banco de dados, e já disponibiliza uma lista com as recomendações de classificação dos dados (Figura 3.2). Em seguida, o administrador do banco pode selecionar as recomendações com as quais concorda. Além disso, o administrador do banco de dados pode adicionar classificações de forma manual.

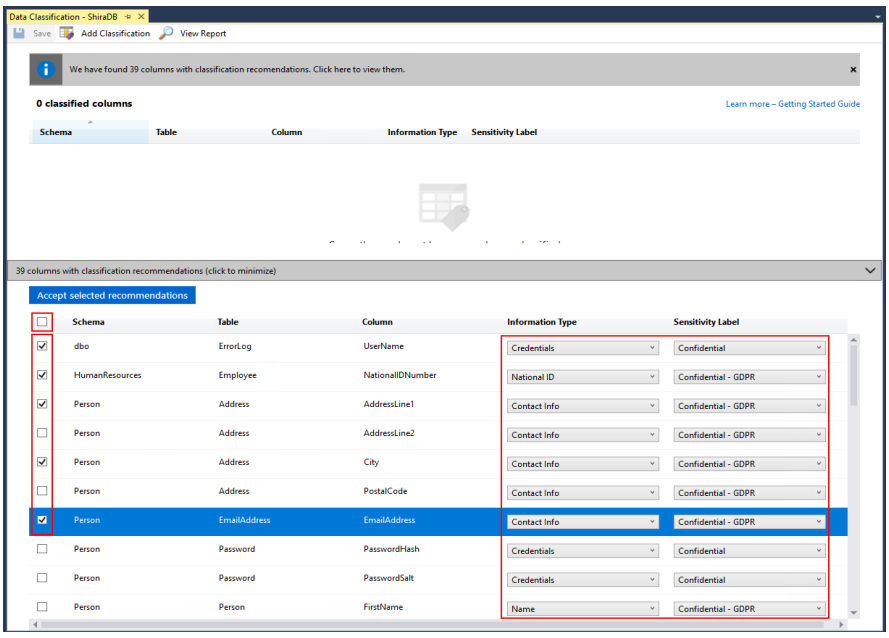


Figura 3.2. Lista de classificação dos dados proposta pelo Data Discovery and Classification [Microsoft 2019a]

Além da descoberta de dados pessoais, é importante avaliar as possíveis vulnerabilidades que existem no SGBD. A *feature* ***sql vulnerability assessment*** pode ajudar a detectar problemas de segurança e permissões. Quando um problema é detectado, pode-se fazer relatórios com uma busca detalhada no banco de dados para encontrar ações para resolução aos problemas através do SSMS [Microsoft 2017d]. Esse relatório (auditoria) também auxilia na categoria de monitoramento/detecção da LGPD. A Figura 3.3 apresenta um exemplo de relatório de verificação de vulnerabilidades. O relatório apresenta uma visão geral do seu estado de segurança, quantos problemas foram encontrados e suas respectivas gravidades.

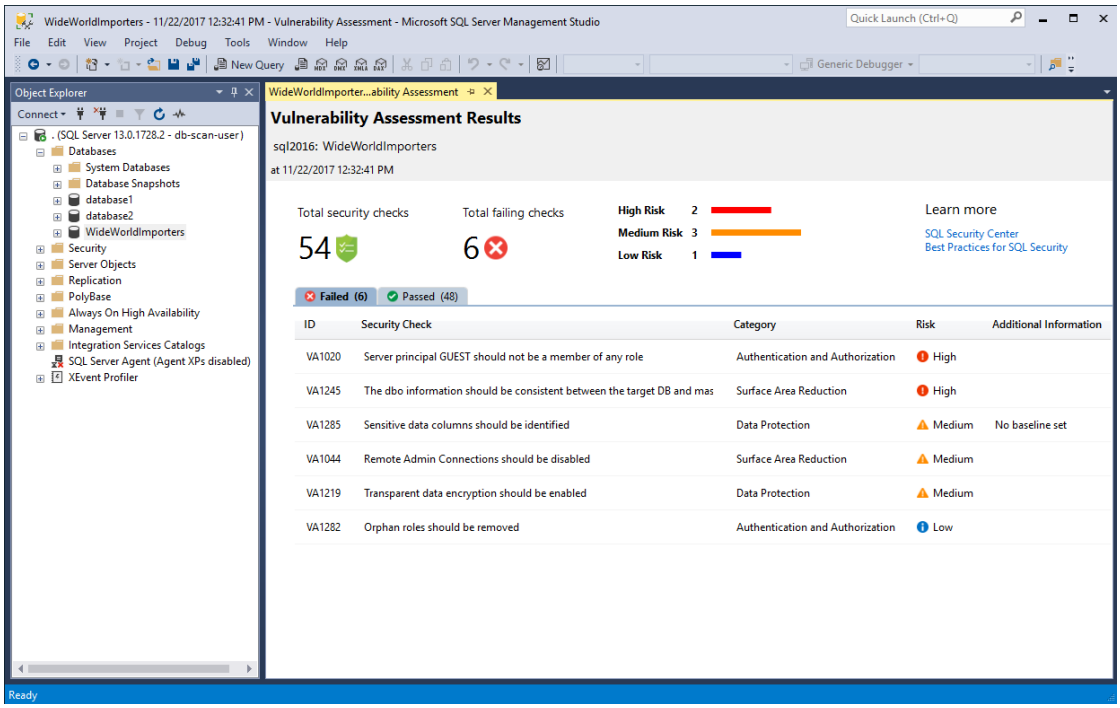


Figura 3.3. Exemplo de relatório de verificação de vulnerabilidades [Microsoft 2017d]

A Figura 3.4 apresenta uma solução possível para um problema de vulnerabilidade recomendada pela própria ferramenta de verificação. A recomendação é remover o membro `GUEST` de todos os papéis.

3.4.1.2 Microsoft - Categoria de Prevenção

A Microsoft disponibiliza *features*, de acordo com as recomendações da LGPD na categoria de prevenção, para a anonimização e pseudonimização através do mascaramento dinâmico e estáticos de dados, o controle de acesso de usuário privilegiado e a criptografia de dados, que apesar de não constar explicitamente na LGPD, é um meio de segurança para os dados, em caso de vazamento.

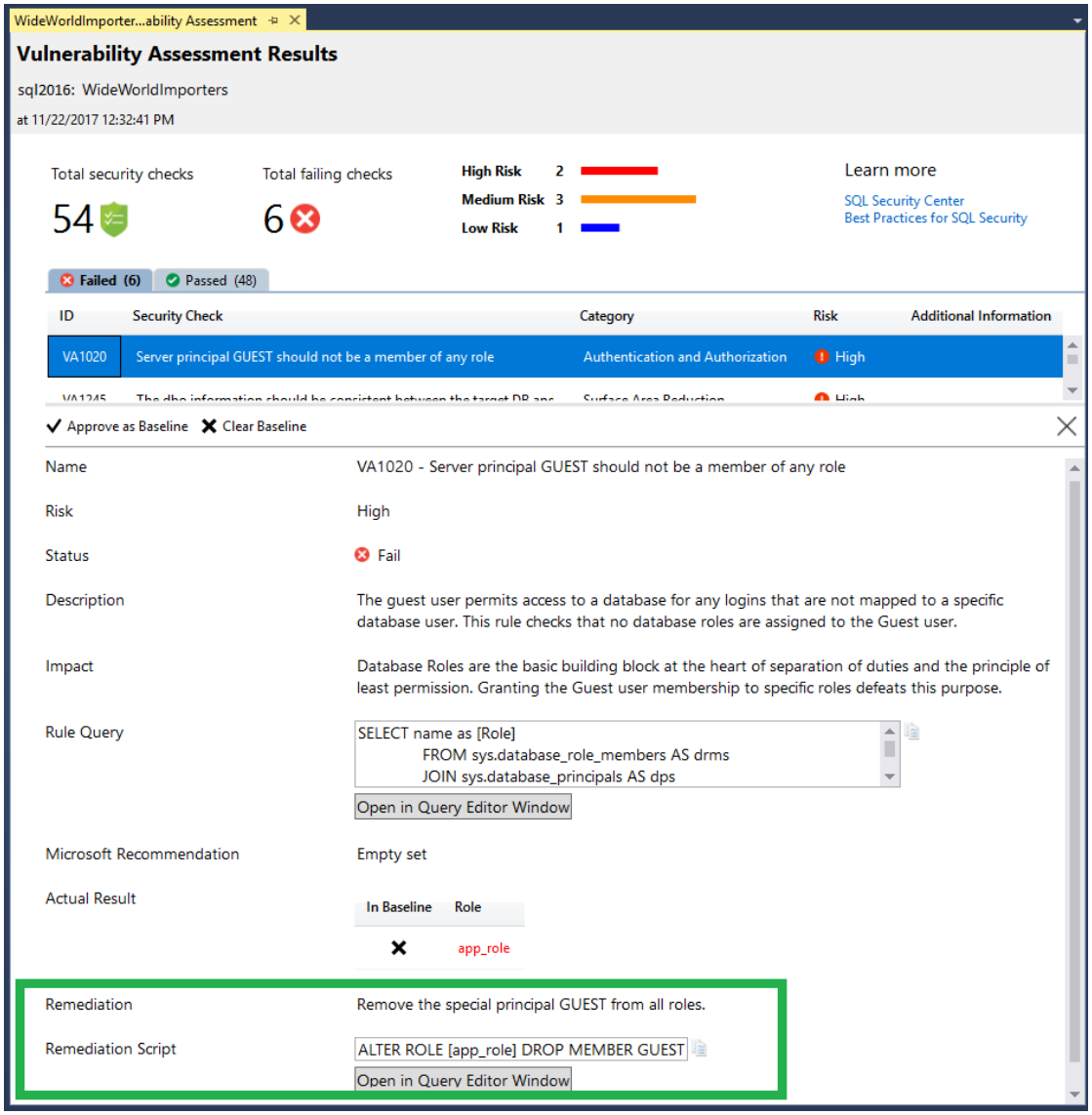


Figura 3.4. Exemplo de solução para vulnerabilidade detectada [Microsoft 2017d]

A *feature* ***dynamic data masking*** **(DDM)** limita a exposição aos dados pessoais através do mascaramento deles para usuários não privilegiados [Microsoft 2019c]. Ele mascara os dados em tempo de execução, facilitando a modelagem e o código de segurança nas aplicações. O mascaramento de dados pode ocorrer de forma total ou parcial e, no caso de dados numéricos, existe um tipo de máscara aleatória. Por exemplo, na Figura 3.5, tem-se um exemplo de comando que adiciona uma função de mascaramento (partial) para a coluna `LastName` da tabela `Membership`. O primeiro argumento da função é o prefixo, ou seja, a posição do primeiro caractere real a ser mostrado do dado, o segundo argumento possui os caracteres do meio e que irão mascarar o conteúdo no momento da exibição do resultado da consulta e o último argumento é o sufixo, ou seja, a posição do último caractere real a ser mostrado.

```
ALTER TABLE Membership
ALTER COLUMN LastName ADD MASKED WITH (FUNCTION = 'partial(2,"XXX",0)');
```

Figura 3.5. Comando de inclusão de mascaramento de dados a uma coluna existente [Microsoft 2019c]

A *feature* ***static data masking***, disponível no Sql Server Management Studio 18.0 preview 5 e posterior, possibilita a criação de uma cópia do banco de dados que tenha os dados pessoais mascarados para que o usuário possa compartilhar tal cópia sem compartilhar os dados pessoais contidos no banco [Granet 2018]. Diferente do DDM, o mascaramento ocorre em nível de armazenamento e todos os usuários da cópia do banco de dados tem o mesmo dado mascarado. Esse mascaramento ocorre em nível de coluna como pode ser visto na Figura 3.6, no passo 1 (step 1), onde as colunas `AddressLine`, `DateOfBirth`, `EmailId`, `FirstName`, `LastName` e `SSN`, da tabela `dbo.Customers`, são selecionadas para mascaramento. As funções escolhidas para o mascaramento de dados são: *shuffle, single value, null, group shuffle* e *string composite*. A função *shuffle* embaralha os dados (`AddressLine`) para as novas linhas e não introduz nenhum valor novo. A função *single value* substitui todos os conteúdos da coluna (`DateOfBirth`) pelo único valor inserido no momento da configuração. A função *null* substitui o conteúdo da coluna (`EmailId`) pelo valor null. Nesse caso, a coluna precisa ser opcional para poder usar essa função. A função *group shuffle* vincula mais de uma coluna (`FirstName` e `LastName`) no mascaramento aleatório, ou seja, usa o conteúdo de mais colunas para o embaralhamento. A função *string composite* permite o mascaramento da coluna inteira ou de parte dela. Por exemplo, o `SSN` pode ser parcialmente mascarado, mantendo-se apenas os seus quatro últimos dígitos. Ainda na Figura 3.6, no passo 2 (step 2), seleciona-se a localização do arquivo destino da cópia mascarada do banco de dados e no passo 3 (step 3) detalha-se o nome do banco de dados destino e o arquivo onde constará o log do mascaramento. Na Figura 3.7 tem-se um exemplo dos dados não mascarados (*Unmasked Data*) e dos dados após o mascaramento (*Masked Data*).

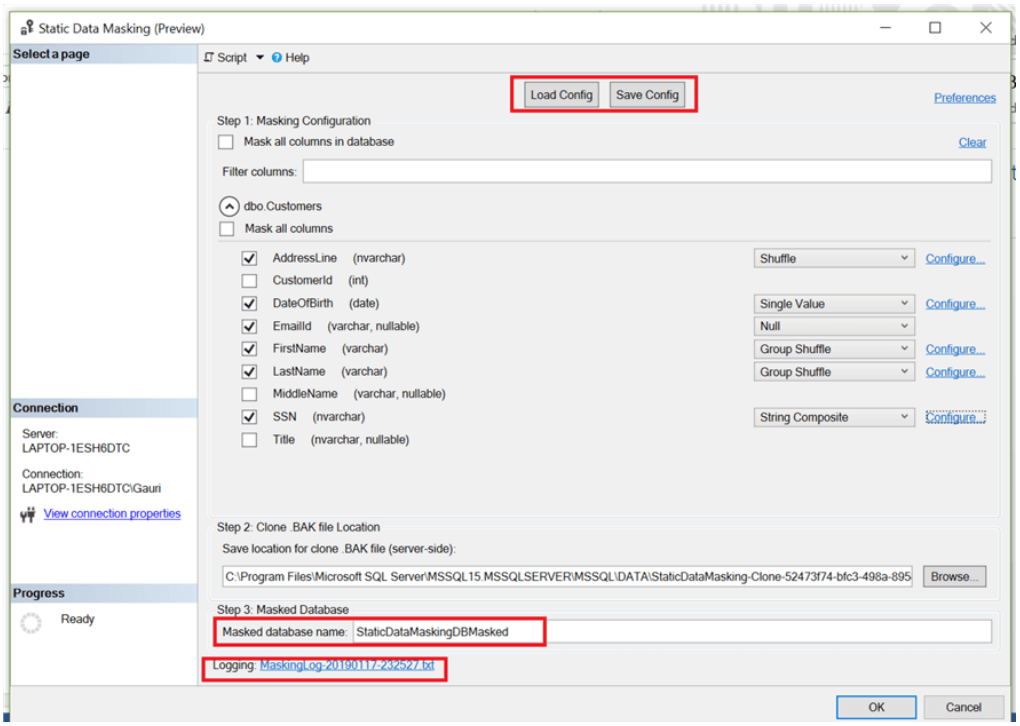


Figura 3.6. Seleção de colunas a serem mascadas fisicamente na cópia do banco de dados [Mahajan 2019]

Unmasked Data

FirstName	MiddleName	LastName	DateOfBirth	SSN	AddressLine	EmailId
Mihail	U	Frintu	1969-01-05	364-95-1699	1970 Napa Ct.	mihail@adventure-works.com
Ken	M	Ray	1996-05-29	150-85-5065	9833 Mt. Dias Blv.	terri0@adventure-works.com
Terri	A	Selikoff	1991-03-17	549-82-9234	7484 Roundtree Drive	roberto0@adventure-works.com
Roberto	N	Poland	1977-03-29	281-65-5849	9539 Glenside Dr	rob0@adventure-works.com
Rob	W	Rettig	1955-05-22	413-25-8713	1226 Shoe St.	gail0@adventure-works.com
Gail	V	Osada	1980-01-29	204-67-1050	1399 Firestone Drive	jossef0@adventure-works.com
Jossef	J	Philips	1991-03-17	764-92-9954	5672 Hale Dr.	dylan0@adventure-works.com
Dylan	C	Netz	1977-03-29	582-55-5002	6387 Scenic Avenue	diane1@adventure-works.com
Diane	M	Keyser	1955-05-22	798-17-7390	8713 Yosemite Ct.	gigi0@adventure-works.com
Gigi	M	Brown	1980-01-13	868-43-4288	250 Race Court	michael6@adventure-works.com
Michael	T	Kaliyath	1972-01-26	864-70-1391	1318 Lasalle Street	ovidiu0@adventure-works.com
Ovidiu	S	Frintu	1966-01-29	381-02-3744	5415 San Gabriel Dr.	thierry0@adventure-works.com
Thierry	N	Creasey	1980-01-29	755-88-1537	9265 La Paz	janice0@adventure-works.com
Janice	R	Cook	1972-01-14	732-84-0387	8157 W. Book	michael8@adventure-works.com
Michael	A	Martinez	1989-01-29	207-57-9704	4912 La Vuelta	sharon0@adventure-works.com
Sharon	Z	Goldstein	1969-01-29	045-14-7883	40 Ellis St.	david0@adventure-works.com
David	A	Cornelsen	1969-01-29	732-05-0120	6696 Anchor Drive	kevin0@adventure-works.com
Kevin	J	Petculescu	1976-12-18	126-77-2761	1873 Lion Circle	john5@adventure-works.com
John	R	Stadick	1969-01-29	822-08-3794	3148 Rose Street	mary2@adventure-works.com
Mary	R	Wedge	1973-01-29	586-80-9583	6872 Thornwood Dr.	wanida0@adventure-works.com

Masked Data

FirstName	MiddleName	LastName	DateOfBirth	SSN	AddressLine	EmailId
Patrick	U	Earls	2000-01-01	XXX-YY-1699	40 Ellis St.	NULL
Gigi	M	Brown	2000-01-01	XXX-YY-5065	2115 Passing	NULL
Mark	A	Yu	2000-01-01	XXX-YY-9234	9537 Ridgewood Drive	NULL
Ryan	N	Sacksteder	2000-01-01	XXX-YY-5849	4948 West 4th St	NULL
James	W	Scardelis	2000-01-01	XXX-YY-8713	7511 Cooper Dr.	NULL
Pete	V	Caron	2000-01-01	XXX-YY-1050	1285 Greenbrier Street	NULL
Danielle	J	Hay	2000-01-01	XXX-YY-9954	Pascalstr 951	NULL
Sandeep	C	Vanderhyde	2000-01-01	XXX-YY-5002	2354 Frame Ln.	NULL
Kevin	M	Koch	2000-01-01	XXX-YY-7390	7726 Driftwood Drive	NULL
Roberto	M	Poland	2000-01-01	XXX-YY-4288	2466 Clearland Circle	NULL
Bonnie	T	Ralls	2000-01-01	XXX-YY-1391	34 Waterloo Road	NULL
Mary	S	Martin	2000-01-01	XXX-YY-3744	10203 Acorn Avenue	NULL
Gail	N	Osada	2000-01-01	XXX-YY-1537	6387 Scenic Avenue	NULL
Denise	R	Bischoff	2000-01-01	XXX-YY-0387	2059 Clay Rd	NULL
Thomas	A	Barbariol	2000-01-01	XXX-YY-9704	5669 Ironwood Way	NULL
Houman	Z	Yalovsky	2000-01-01	XXX-YY-7883	5666 Hazelnut Lane	NULL
Sidney	A	Lertpiriyasuwat	2000-01-01	XXX-YY-0120	8463 Vista Avenue	NULL
Reuben	J	Watters	2000-01-01	XXX-YY-2761	1061 Buskirk Avenue	NULL
Peter	R	Keyser	2000-01-01	XXX-YY-3794	502 Alexander Pl.	NULL
Michael	R	Gubbels	2000-01-01	XXX-YY-9583	9784 Mt Etna Drive	NULL

Figura 3.7. Dados antes e após o mascaramento estático de dados [Mahajan 2019]

A partir daqui as *features* de prevenção lidam com o controle de acesso dos usuários. A *feature* ***sql server authentication*** ajuda a gerenciar as identidades dos usuários que acessam os bancos de dados e os servidores, impedindo o acesso não autorizado e pode ser configurado no SSMS [Microsoft 2018b]. Existem duas formas de autenticação no SGBD sql server: a autenticação do Windows e o modo misto. A autenticação do Windows é a forma padrão, onde as contas de usuário e grupos específicas do Windows são confiáveis para conectarem ao sql server. Já o modo misto suporta tanto a autenticação pelo Windows quanto pelo próprio sql server. Os pares de nome de usuário e senha são mantidos no sql server. A recomendação da Microsoft é utilizar a autenticação do Windows sempre que for possível, visto que a autenticação do Windows usa diversas mensagens criptografadas para autenticar os usuários no sql server, enquanto as credenciais do sql server trafegam pela rede, tornando-as menos seguras.

A *feature object-level permissions* permite a concessão de permissões em um nível excepcionalmente granular – até visualização de tabela, procedimento armazenado,

função escalar ou serviço de fila [Microsoft 2014]. Na Figura 3.8 tem-se um exemplo de consulta sobre as permissões que os usuários possuem sobre os objetos do banco de dados.

Results | Messages

	UserName	UserType	DatabaseUserName	Role	PermissionType	PermissionState	ObjectType	ObjectName	ColumnNa
1	NULL	Windows User	dbo	NULL	CONNECT	GRANT	DATABASE	NULL	NULL
2	NULL	SQL User	guest	NULL	NULL	NULL	NULL	NULL	NULL
3	Test	SQL User	Test	NULL	CONNECT	GRANT	DATABASE	NULL	NULL

Figura 3.8. Permissões de usuários aos objetos do banco de dados [Microsoft 2014]

A *feature* ***role-based security*** permite conceder permissões baseadas em papéis ou grupos de usuários ao invés de usuários individuais, reduzindo o ataque ao banco de dados e simplificando a administração de segurança [Microsoft 2017a]. O sql server disponibiliza papéis, em nível de servidor, para a administração do SGBD e as permissões atribuídas a eles não podem ser alteradas. O papel *sysadmin* abrange todos os outros papéis e tem escopo ilimitado, devendo ser atribuído somente a usuários altamente confiáveis. Além disso, existem papéis em nível de banco de dados, tendo um conjunto pré-definido de permissões. Os usuários do banco de dados podem ser adicionados aos papéis do banco de dados ou do servidor.

A *feature* ***row-level security*** restringe o acesso, de acordo com os direitos do usuário, limitando o acesso a linhas em uma tabela baseado no relacionamento entre o usuário e o dado [Microsoft 2019b]. Essa *feature* é implementada no sql server através da instrução `CREATE SECURITY POLICY` e predicados criados como funções com valores embutidos da tabela. Na Figura 3.9 tem-se um exemplo de criação de uma função que retorna o valor 1 (um) quando o conteúdo de uma linha da coluna do representante de vendas (@SalesRep) é o mesmo que o usuário que executa a consulta (@SalesRep = USER_NAME()) ou se o usuário que está executando a consulta for o gerente (USER_NAME() = 'MANAGER').

```
CREATE FUNCTION Security.fn_securitypredicate(@SalesRep AS sysname)
    RETURNS TABLE
WITH SCHEMABINDING
AS
    RETURN SELECT 1 AS fn_securitypredicate_result
WHERE @SalesRep = USER_NAME() OR USER_NAME() = 'Manager';
```

Figura 3.9. Função que realiza o cruzamento do usuário que realiza a consulta e a linha que está sendo consultada [Microsoft 2019b]

Na Figura 3.10 é apresentado o comando para a criação de uma política de segurança que adiciona a função anterior (Figura 3.9) como um predicado de filtro sobre a tabela de vendas (dbo.Sales). O estado (STATE=ON) precisa ser definido como ON para habilitar a política.

```
CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE Security.fn_securitypredicate(SalesRep)
ON dbo.Sales
WITH (STATE = ON);
```

Figura 3.10. Comando de criação de política de segurança [Microsoft 2019b]

Na Figura 3.11 tem-se as permissões de consulta (`GRANT SELECT ON`) na função para os usuários `Manager, Sales1` e `Sales2`. Dessa forma, quando o gerente consultar os dados, a função retornará 1 e o mesmo terá acesso a todas as linhas contidas

na tabela de vendas. Já os vendedores `Sales1` e `Sales2` só terão acesso às linhas de suas próprias vendas, pois, como ele não possui o papel de gerente, a função só retornará 1 quando o usuário que estiver consultando for igual ao conteúdo da linha da coluna representante de vendas.

```
GRANT SELECT ON security.fn_securitypredicate TO Manager;
GRANT SELECT ON security.fn_securitypredicate TO Sales1;
GRANT SELECT ON security.fn_securitypredicate TO Sales2;
```

Figura 3.11. Exemplo de permissões aos usuários [Microsoft 2019b]

Finalizando as *features* de prevenção, tem-se àquelas ligadas à criptografia. Na camada de transporte, a *feature* **TLS** é um protocolo de comunicação que garante comunicações altamente seguras, onde os dados são criptografados para ajudar a garantir que nenhum dado seja interceptado durante o tráfego entre o banco de dados e a aplicação cliente [Microsoft 2019d]. O TLS pode ser usado para validação do servidor quando uma conexão do cliente solicita criptografia. Se a instância do sql server estiver sendo executada em um computador ao qual foi atribuído um certificado de uma autoridade de certificação pública, a identidade do computador e a instância do SQL Server serão emitidas pela cadeia de certificados que leva à autoridade raiz confiável. Essa validação de servidor exige que o computador, no qual o aplicativo cliente está sendo executado, seja configurado para confiar na autoridade raiz do certificado que é usado pelo servidor.

A *feature* **TDE** protege os dados em repouso mesmo que a mídia física (cópias de segurança) seja perdida ou que os dados sejam descartados incorretamente [Microsoft 2019e]. Ele criptografa e descriptografa o banco de dados, as cópias de segurança e os logs de transações em tempo real, sem requerer qualquer mudança nas aplicações. A criptografia usa uma DEK (chave de criptografia do banco de dados), que é armazenada no registro de inicialização do banco de dados para disponibilidade durante a recuperação. A DEK é uma chave simétrica protegida por um certificado armazenado no banco de dados mestre do servidor ou uma chave assimétrica protegida por um módulo EKM (gerenciamento extensível de chaves). Na Figura 3.12 tem-se uma série de comandos para a utilização da TDE. Primeiro é necessário criar uma chave mestra (`CREATE MASTER KEY`), atribuindo-se uma senha. Em seguida, cria-se um certificado protegido pela chave mestra (`CREATE CERTIFICATE`). Após essa criação, cria-se uma chave de criptografia de banco de dados (`CREATE DATABASE ENCRYPTION KEY`), protegendo-a com o certificado anterior e por fim, define-se o banco de dados para usar a criptografia (`ALTER DATABASE ... SET ENCRYPTION ON`).

```
USE master;
GO
CREATE MASTER KEY ENCRYPTION BY PASSWORD = '<UseStrongPasswordHere>';
go
CREATE CERTIFICATE MyServerCert WITH SUBJECT = 'My DEK Certificate';
go
USE AdventureWorks2012;
GO
CREATE DATABASE ENCRYPTION KEY
WITH ALGORITHM = AES_128
ENCRYPTION BY SERVER CERTIFICATE MyServerCert;
GO
ALTER DATABASE AdventureWorks2012
SET ENCRYPTION ON;
GO
```

Figura 3.12. Comandos para uso do TDE [Microsoft 2019e]

A *feature* ***always encrypted*** é uma tecnologia que auxilia na proteção de dados pessoais enquanto eles estão em uso em nível de coluna [Microsoft 2017c]. Ela criptografa e descriptografa no computador cliente sem revelar a chave de criptografia para o servidor do banco de dados. Dessa forma, os dados ficam visíveis somente para as pessoas responsáveis por gerenciar tais dados e não para os administradores do banco de dados ou usuários altamente privilegiados que não tenham acesso. Como resultado, essa tecnologia fornece uma separação entre aqueles que possuem os dados (e podem exibi-lo) e aqueles que gerenciam os dados (mas que não devem ter acesso). Um driver é instalado no computador cliente e automaticamente, criptografa e descriptografa os dados confidenciais. O driver criptografa as colunas de dados confidenciais antes de passar os dados para o servidor de banco de dados e reconfigura automaticamente as consultas para que a semântica do aplicativo seja preservada. Um cenário em que esse tipo de tecnologia é útil é quando uma empresa deseja que um fornecedor externo administre o sql server. Dessa forma, eles não terão acesso aos dados confidenciais, pois os mesmos estarão criptografados no banco. Na Figura 3.13 apresenta-se um exemplo de comandos que cria os metadados de uma chave mestra (`CREATE COLUMN MASTER KEY`) de coluna, os metadados de chave de criptografia de coluna (`CREATE COLUMN ENCRYPTION KEY`) e uma tabela com colunas criptografadas (`CustName` e `SSN`). O valor de `ENCRYPTED_VALUE` foi cortado para não sobrecarregar a figura.

```
CREATE COLUMN MASTER KEY MyCMK
WITH (
     KEY_STORE_PROVIDER_NAME = 'MSSQL_CERTIFICATE_STORE',
     KEY_PATH = 'Current User/Personal/f2260f28d909d21c642a3d8e0b45a830e79a1420'
   );
-----------------------------------------------
CREATE COLUMN ENCRYPTION KEY MyCEK
WITH VALUES
(
    COLUMN_MASTER_KEY = MyCMK,
    ALGORITHM = 'RSA_OAEP',
    ENCRYPTED_VALUE = 0x01700000016C006F00630061006C006D0061006300680069006E0065002F006D0
);
-----------------------------------------------
CREATE TABLE Customers (
    CustName nvarchar(60)
        COLLATE  Latin1_General_BIN2 ENCRYPTED WITH (COLUMN_ENCRYPTION_KEY = MyCEK,
        ENCRYPTION_TYPE = RANDOMIZED,
        ALGORITHM = 'AEAD_AES_256_CBC_HMAC_SHA_256'),
    SSN varchar(11)
        COLLATE  Latin1_General_BIN2 ENCRYPTED WITH (COLUMN_ENCRYPTION_KEY = MyCEK,
        ENCRYPTION_TYPE = DETERMINISTIC ,
        ALGORITHM = 'AEAD_AES_256_CBC_HMAC_SHA_256'),
    Age int NULL
);
GO
```

Figura 3.13. Comandos para dados sempre criptografados [Microsoft 2017c]

3.4.1.3 Microsoft - Categoria de Monitoramento/detecção

Quando um vazamento acontece, a organização precisa detectá-lo o mais rapidamente possível para minimizar seu impacto, além de entender quais registros foram afetados. É importante que toda a organização seja alertada imediatamente quando alguma atividade

fora do comum for detectada e, a partir daí, monitore comportamentos suspeitos. Além disso, o controlador deve ter o auxílio de ferramentas de auditoria que possam gerar relatórios sobre as atividades que ocorreram no banco de dados.

A *feature* ***sql server audit*** rastreia as atividades do banco de dados para ajudar no entendimento e identificação de possíveis ameaças, abusos suspeitos ou violação de segurança [Microsoft 2016a]. Os eventos auditados (ações atômicas) podem ser gravados nos logs de eventos ou nos arquivos de auditoria e ocorrem em nível de servidor ou de banco de dados. Por padrão, o sql server não habilita a auditoria. Ela pode ser habilitada pelo SSMS ou via linha de comando sql. Utilizando a linha de comando, na Figura 3.14, são mostrados exemplos de criação de auditoria se servidor (`CREATE SERVER AUDIT`), a habilitação da auditoria do servidor (`ALTER SERVER ... STATE=ON`) e a criação de uma auditoria de banco de dados que audita as instruções `SELECT` e `INSERT` realizadas por qualquer usuário `dbo` para a tabela `HumanResources.EmployeePayHistory` no banco de dados `AdventureWorks2012` [Microsoft 2017b].

```
USE master ;
GO
-- Create the server audit.
CREATE SERVER AUDIT Payrole_Security_Audit
    TO FILE ( FILEPATH =
linux file path) ;
GO
-- Enable the server audit.
ALTER SERVER AUDIT Payrole_Security_Audit
WITH (STATE = ON) ;
USE AdventureWorks2012 ;
GO
-- Create the database audit specification.
CREATE DATABASE AUDIT SPECIFICATION Audit_Pay_Tables
FOR SERVER AUDIT Payrole_Security_Audit
ADD (SELECT , INSERT
     ON HumanResources.EmployeePayHistory BY dbo )
WITH (STATE = ON) ;
GO
```

Figura 3.14. Comandos para criar e habilitar auditorias de servidor e banco de dados [Microsoft 2017b]

A *feature* ***sql server temporal tables*** são tabelas com versão do sistema, sendo do tipo de tabela de usuário modeladas para manter um histórico completo da mudança dos dados em qualquer ponto no tempo e que podem ser usadas para gerar relatórios sobre os dados auditados [Microsoft 2016b]. A tabela temporal é dita com versão do sistema porque o período de validade para cada linha é gerenciado pelo sistema (SGBD). Cada tabela temporal possui duas colunas do tipo datetime2, que armazena o período de validade para cada linha sempre que uma linha é modificada. Além disso, a tabela temporal possui uma referência a outra tabela com um esquema espelhado (histórico), que é usada para armazenar a versão anterior da linha, automaticamente, sempre que uma linha na tabela temporal é atualizada ou excluída. Na Figura 3.15 é apresentado um exemplo de consulta à tabela de histórico (`FOR SYSTEM_TIME`), onde são retornadas as versões de linhas que satisfaçam a condição de `EmployeeID = 1000` e que estavam ativas durante o período de 01/01/2014 e 01/01/2015, inclusive.

```
SELECT * FROM Employee
    FOR SYSTEM_TIME
      BETWEEN '2014-01-01 00:00:00.0000000' AND '2015-01-01 00:00:00.0000000'
        WHERE EmployeeID = 1000 ORDER BY ValidFrom;
```

Figura 3.15. Consulta a tabela de histórico [Microsoft 2016b]

A *feature* ***sql vulnerability assessment***, conforme já descrito no item de avaliação, também auxilia com relatórios sobre as atividades ocorridas no banco de dados.

3.4.2. Oracle

A Oracle disponibiliza diversos aspectos de segurança para apoiar a privacidade dos dados. Para encolher a superfície de ataque e reduzir o número de maneiras pelas quais os invasores podem acessar os bancos de dados, é extremamente importante impor a segurança o mais próximo possível dos dados. Considerando as três categorias de requisitos da LGPD, descrevemos a seguir alguns dos diversos produtos que a Oracle disponibiliza para auxiliar na proteção dos dados armazenados e no controle de acesso a esses dados.

- Avaliação: *oracle data safe* (avaliação de segurança e risco) e *oracle database vault*.
- Prevenção: *oracle data safe* (mascaramento de dados), *oracle data masking and subsetting*, *oracle advanced security*, *oracle key vault*, *oracle database vault* e *oracle label security*.
- Monitoramento/detecção: *oracle data safe* (auditoria de atividades), *oracle audit vault* e *database firewall*.

3.4.2.1. Oracle - Categoria de Avaliação

Um dos desafios ao avaliar a natureza dos riscos é determinar o que avaliar, porque os aplicativos de banco de dados normalmente contêm vários pontos de entrada e têm dados pessoais espalhados por várias colunas e tabelas com controle de acesso vagamente definido. A Oracle auxilia nesse desafio da LGPD provendo ferramentas para ajudar na avaliação de segurança e risco.

Oracle data safe

O *oracle data safe* é um serviço na nuvem que garante segurança para os bancos de dados residentes na nuvem [Oracle 2019a]. Essa segurança é obtida através de avaliações de segurança e risco do usuário, auditoria de atividades, descoberta de dados confidenciais e mascaramento de dados.

A avaliação de segurança auxilia na identificação da existência de lacunas na estratégia de configuração do banco de dados e sugere maneiras de corrigir essas lacunas. Dessa forma, é possível identificar vulnerabilidades de segurança, por exemplo, quando senhas padronizadas estão sendo utilizadas ou quando os usuários possuem mais privilégios do que eles deveriam. Por exemplo, a Figura 3.16 apresenta um alerta do serviço para a quantidade de usuários que possuem o privilégio de DBA (DBA Role) no banco de dados, visto que esse tipo de privilégio deve ser restrito apenas aos administradores da base de dados e que possam ter acesso total aos dados, incluindo os dados sensíveis.

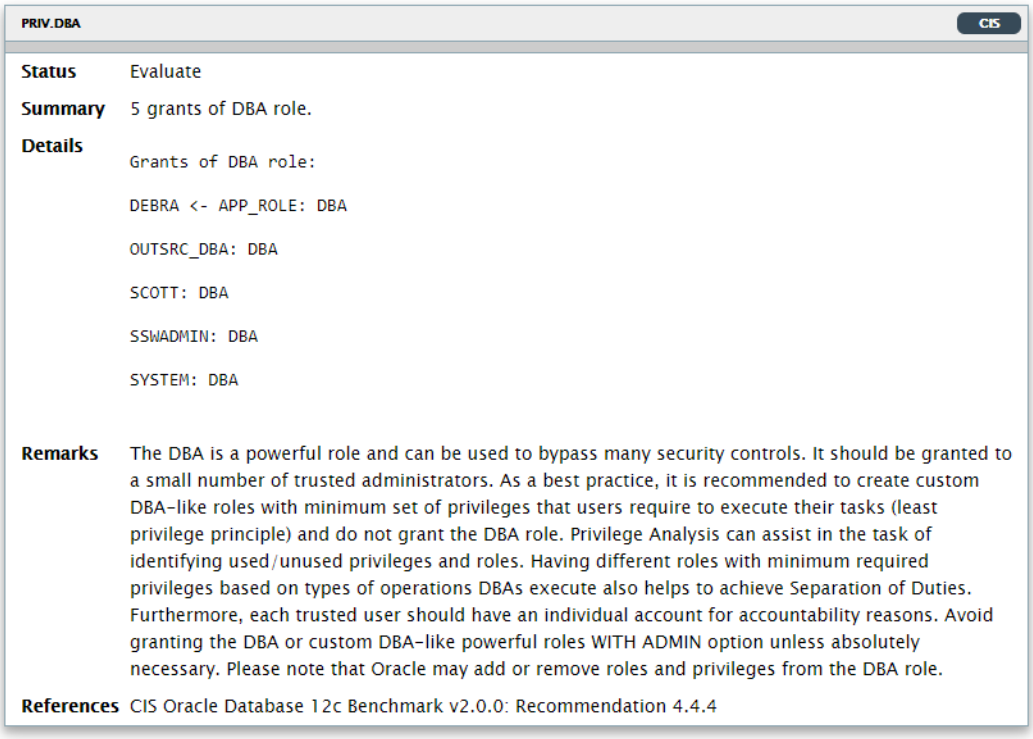


Figura 3.16. Exemplo de uso do serviço de avaliação de segurança[3]

A avaliação de risco do usuário permite avaliar e monitorar o usuário de forma a identificar possíveis riscos associados a contas privilegiadas. A Figura 3.17 apresenta um exemplo de análise realizada pelo serviço, onde algumas contas de usuários apresentam alto nível de exposição a riscos.

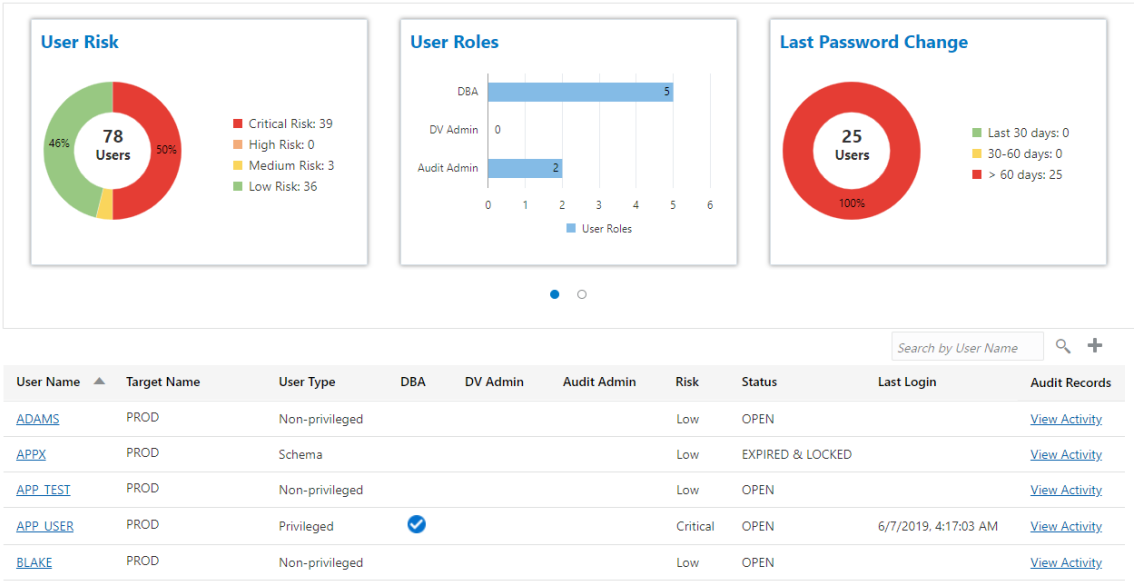


Figura 3.17. Exemplo de avaliação de riscos de usuários [Oracle 2019a]

[3] https://blog.bronto.com/bg/database/technologies/security/dbsat.html

As demais funcionalidades do *oracle data safe* são detalhadas nas suas respectivas categorias.

Oracle database vault

O *oracle database vault* trabalha em conjunto com o banco de dados oracle para evitar ameaças que exploram credenciais roubadas, usuários que usam contas privilegiadas para acessar dados confidenciais, criar novas contas, conceder privilégios adicionais, usuários que ignoram as políticas de uso dos dados das organizações, ameaças aos dados confidenciais durante a janela de manutenção das aplicações entre outras [Oracle 2019c]. Devido às suas diversas funções, o *oracle database vault* auxilia tanto na avaliação quanto na categoria de prevenção de ataques.

Após a identificação dos dados pessoais, torna-se importante identificar usuários (titulares de dados, terceiros, autoridades de supervisão e destinatários), incluindo usuários e administradores privilegiados (controladores, operadores), que não podem apenas acessar, mas também processar os dados pessoais [Rajasekharan 2017]. Durante a modelagem e a manutenção do sistema, privilégios adicionais podem ser concedidos inadvertidamente aos usuários. A análise de privilégios do *oracle database vault* ajuda a aumentar a segurança dos aplicativos, identificando os privilégios reais usados no tempo de execução. Os privilégios identificados como não utilizados, podem ser avaliados para uma possível revogação, ajudando a obter um modelo de privilégios mínimos.

3.4.2.2. Oracle - Categoria de Prevenção

Conforme já discutido, a LGPD recomenda diversas técnicas preventivas, tais como: pseudonimização, anonimização, controle de usuário privilegiado entre outras. Um dos desafios de qualquer controle de proteção de dados preventivo é a possível sobrecarga que ele cria nos sistemas e nas operações diárias de TI (tecnologia da informação). Esta sobrecarga pode vir em termos de mudança de processos; alterações necessárias no código-fonte do sistema, sobrecarga de desempenho e preocupações com escalabilidade. No entanto, a Oracle descreve que aborda tais desafios através de controles preventivos transparentes para a maioria dos sistemas e com um impacto mínimo no desempenho e nas operações contínuas de TI [Rajasekharan 2017].

Oracle data safe

O *oracle data safe* auxilia no item de anonimização dos dados, através do seu mascaramento, para que caso ocorra vazamento de dados pessoais, esses dados não sejam vinculados às pessoas reais. Para isso, é importante identificá-los. A descoberta de dados confidenciais ajuda a decidir quais dados devem ser protegidos. Esse serviço identifica e classifica mais de 125 tipos de dados sensíveis, tais como: dados de tecnologia da informação, dados financeiros, dados de saúde entre outros. Esse serviço é particularmente útil para empresas que possuem várias equipes de desenvolvimento e seus dados estejam distribuídos sobre vários bancos de dados, sendo difícil a identificação dos dados sensíveis e onde os mesmos estão localizados. A Figura 3.18 mostra algumas das categorias pré-definidas de dados sensíveis e a partir daí, o usuário pode selecionar a categoria que ele deseja descobrir quais seriam os dados sensíveis nos seus bancos de dados.

Figura 3.18. Exemplo de categorias de descoberta de dados sensíveis [Oracle 2019a]

O mascaramento de dados substitui os dados sensíveis do ambiente de produção por dados fictícios, mas realistas. Esse mascaramento pode ser usado para os ambientes de desenvolvimento e homologação das organizações, onde o desenvolvedor não precisa ter acesso, por exemplo, ao número do cartão de crédito de um usuário para realizar testes mais próximos da realidade. Dessa forma, o conjunto de dados de teste passa a ser realista, mas sem expor os dados sensíveis. A Figura 3.19 apresenta um exemplo de mascaramento do identificador do cliente (SSN) e do seu cartão de crédito (Credit Card).

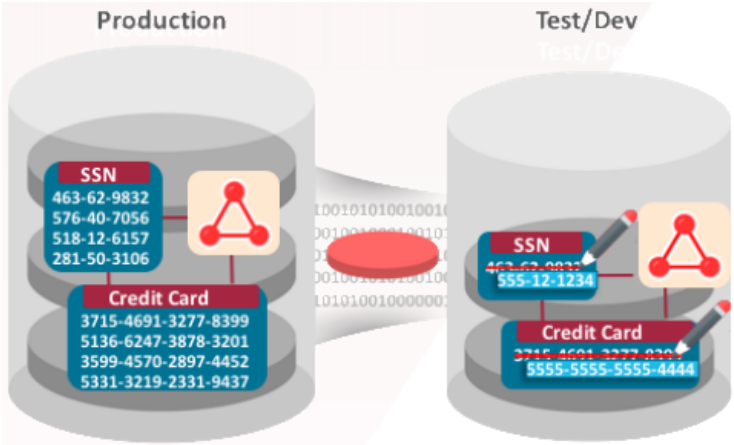


Figura 3.19. Exemplo de mascaramento de dados [Oracle 2019a]

Oracle data masking and subsetting

O *oracle data masking and subsetting* também auxilia na anonimização de dados e é um *plugin* acoplado ao banco de dados Oracle que cria um ambiente (desenvolvimento ou homologação) com um subconjunto dos dados de produção mascarados, mas realistas [Oracle 2017]. O mascaramento de dados segue um conjunto de regras pré-definidas para mapeamento. Esse plugin possui o objetivo similar ao mascaramento de dados contido no *oracle data safe*.

Oracle advanced security

Embora a LGPD não indique, explicitamente, a criptografia como uma forma de prevenção, a RGPD recomenda tal técnica, sendo ela extremamente importante para prevenção. O desafio para as organizações é a implementação da criptografia não só para os dados pessoais em tabelas criptografadas, mas também para *backups*, *data dumps* e

arquivos de *log*. O *oracle advanced security* é uma opção do banco de dados Oracle 19c e auxilia na criptografia de dados através do *Oracle advanced security transparent data encryption* (TDE), mas também na pseudonimização através do *Oracle advanced security data redaction* [Oracle 2019b].

O TDE criptografa os dados sensíveis na camada de banco de dados e, com isso, auxilia na prevenção de ataques que tentam ignorar o banco de dados e ler informações confidenciais de arquivos de dados no nível do sistema operacional, de *backups* de banco de dados ou de exportações de banco de dados. Os aplicativos e usuários autenticados no banco de dados continuam tendo acesso aos dados de forma transparente, enquanto usuários não autenticados que tentam burlar o banco de dados têm acesso negado para descriptografar os dados.

Na Figura 3.20 apresenta-se um exemplo de ataque que o banco de dados pode sofrer de pessoas que tenham acesso ao usuário do sistema operacional que tenha privilégios sobre os arquivos do banco de dados. No exemplo citado, o usuário do sistema operacional pode buscar por conteúdos com números, no arquivo que possui dados financeiros (*tablespace*) do banco de dados. Com isso, o usuário consegue obter a informação limpa com os números dos cartões de créditos registrados no banco de dados.

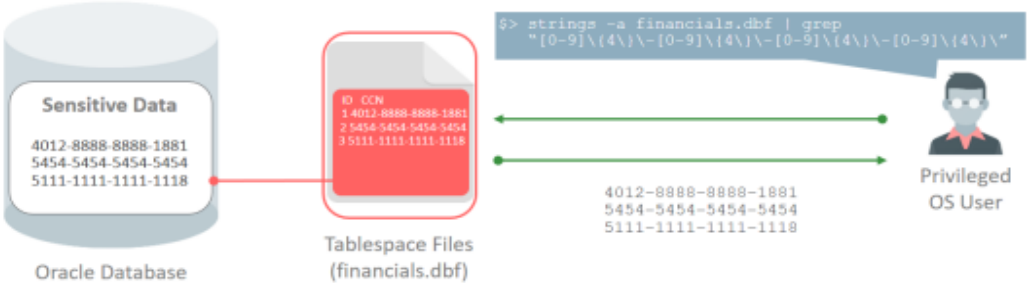


Figura 3.20. Exemplo de ataque ao banco de dados [Oracle 2019b]

Já na Figura 3.21 consegue-se ver o mesmo cenário com o TDE em ação. O TDE pode criptografar *tablespaces* ou até mesmo bancos de dados inteiros, incluindo as *tablespaces* SYSTEM, SYSAUX, TEMP e UNDO. Todo esse processo é transparente para as aplicações porque os processos de criptografia e descriptografia não requerem qualquer mudança na aplicação e os usuários das aplicações não conseguem lidar diretamente com os dados criptografados. Além dessas opções, o DTE também permite criptografar somente algumas colunas na tabela, desde que o usuário saiba quais seriam os dados sensíveis. Essa opção é relevante para bancos de dados enormes, tal como o *datawarehouse*.

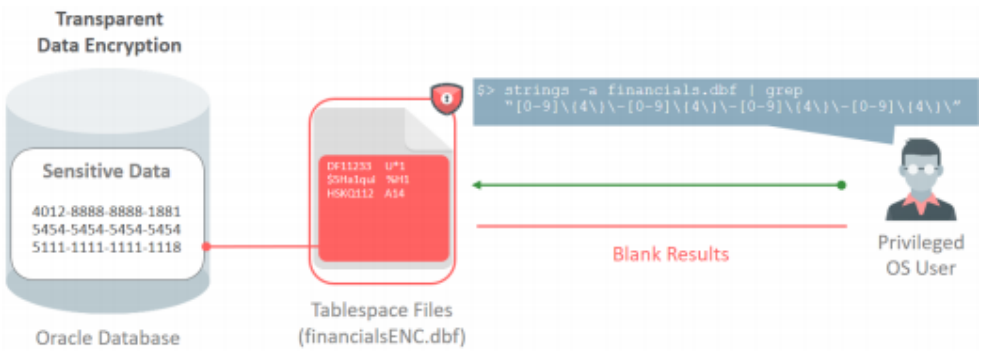


Figura 3.21. Exemplo de intervenção do TDE ao ataque no banco de dados [Oracle 2019b]

O *Oracle advanced security data redaction* fornece uma redação seletiva e dinâmica de dados sensíveis durante a apresentação dos resultados de uma consulta ao banco de dados, antes da exibição pelos aplicativos, para que os usuários não autorizados não possam visualizar tais dados. Os dados armazenados permanecem inalterados, enquanto os dados exibidos são transformados e editados antes de sair do banco de dados. O tipo de cenário em que essa redação pode ser relevante é o uso de aplicações de *call center*. O atendente não deve possuir acesso às informações confidenciais de clientes, como o número do cartão de crédito, visto que isso pode violar alguns regulamentos de privacidade (ex.: LGPD) e expor dados confidenciais sem necessidade.

A Figura 3.22 apresenta alguns exemplos de transformações que podem ser realizadas antes de serem exibidas nas aplicações.

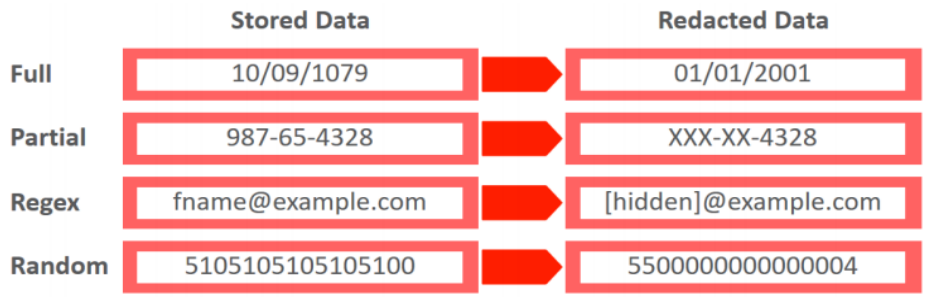


Figura 3.22. Exemplo de redação de dados para as aplicações [Oracle 2019b]

Oracle key vault

O *oracle key vault* também auxilia na prevenção através do controle centralizado sobre dados criptografados com o TDE. Ele possui a capacidade de suspender o acesso à chave mestra e renderizar os dados criptografados de forma ininteligível em caso de violação de dados ou atividade suspeita [Rajasekharan 2017]. O *oracle key vault* é um sistema de segurança para armazenar, centralizar e gerenciar chaves mestras do TDE (usadas para criptografia e descriptografia de dados) de vários bancos de dados Oracle e outros aplicativos de segurança [Oracle 2018a]. Esse sistema elimina alguns desafios operacionais de gerenciamento de chaves tais como: rotação periódica de senhas, realização de cópias de segurança e recuperação de senhas perdidas.

Além disso, conforme mostrado na Figura 3.23, as chaves de criptografia são armazenadas fisicamente e gerenciadas em um local separado de onde os dados criptografados residem, atendendo a uma regra frequente em regulamentos de segurança.

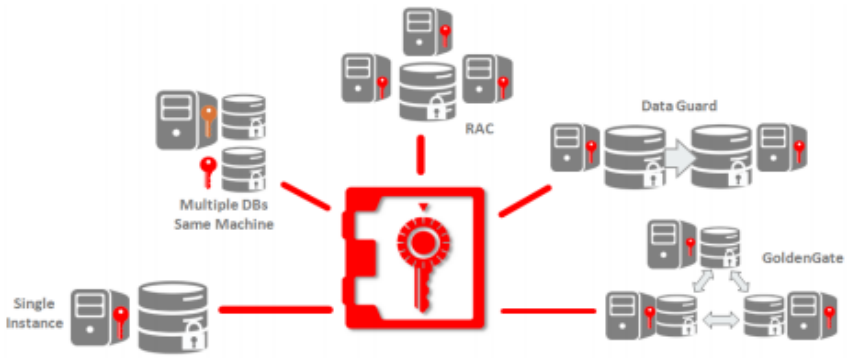


Figura 3.23. Cenário de uso do oracle key vault [Oracle 2018a]

Oracle database vault

O *oracle database vault* auxilia na prevenção através do controle de usuário privilegiado (papel de DBA). Esse tipo de conta, normalmente, possui acesso completo aos dados armazenados no banco de dados. No entanto, com o *oracle database vault*, é criado um ambiente de aplicação restrito ("Realm") dentro do banco de dados que previne o acesso aos dados da aplicação a partir de contas privilegiadas enquanto continua permitindo as atividades administrativas autorizadas regulares no banco de dados. Na Figura 3.24 tem-se um exemplo onde o DBA não consegue recuperar os dados de uma determinada tabela (`hr.emp`), ou seja, somente a pessoa autorizada para visualizar esses dados que consegue.

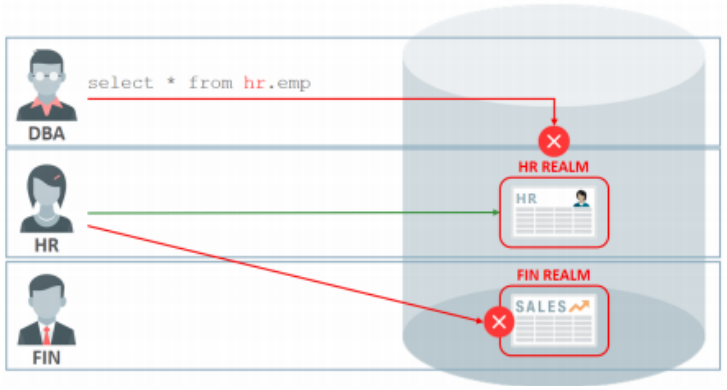


Figura 3.24. Oracle database vault atuando sobre contas privilegiadas [Oracle 2019c]

Outra função do *oracle database vault* é controlar a configuração do banco de dados de forma a impedir alterações no banco de dados que possam levar a configurações inseguras, desvios de configuração (alterações em estruturas de tabelas), reduzir a possibilidade de constatações de auditoria e melhorar a conformidade. Essa prevenção é adquirida através do controle do uso de comandos, tais como: ALTER SYSTEM, ALTER USER, CREATE USER, DROP USER entre outros. Por exemplo, na Figura 3.25, tem-se um cenário em que comandos do DBA são recusados, tais como: TRUNCATE TABLE e CONNECT de um IP desconhecido pelo banco de dados. Em resumo, é controlado o uso de comandos SQL que possam modificar o dicionário e a configuração do banco de dados e com isso, abrir o banco de dados para vulnerabilidades de segurança.

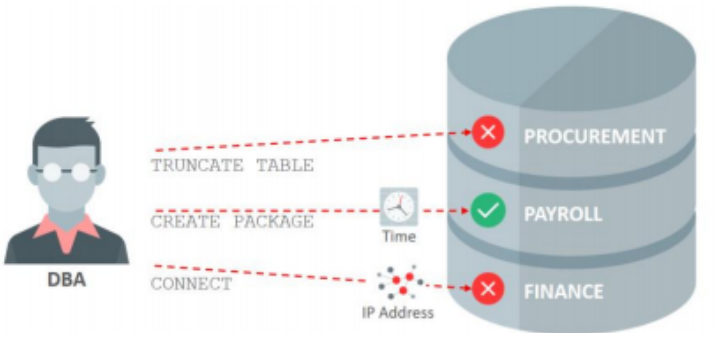


Figura 3.25. Oracle database vault atuando sobre a configuração do banco de dados [Oracle 2019c]

Oracle label security

O *oracle label security* também auxilia no controle de acesso, mas permitindo o controle de acessos multiníveis, requeridos por aplicações governamentais e militares [Oracle

2018b]. Ele é integrado ao *oracle enterprise manager* e está disponível junto com o banco de dados oracle – edição *enterprise*.

No controle de acesso multinível, tanto os dados quanto os usuários do banco de dados recebem uma classificação (*label* ou *classification*). A cada vez que um usuário tenta acessar um dado no banco, é verificado quais dados que o usuário possui acesso e o banco de dados só retorna aqueles dados que o usuário possui acesso. No exemplo da Figura 3.26, o usuário só possui acesso aos dados classificados como sensíveis (*sensitive*) do tipo *alpha* e *beta*. Sendo assim, somente duas linhas da tabela `locations` que são retornadas, mesmo a tabela possuindo cinco linhas (demais dados classificados como altamente sensíveis).

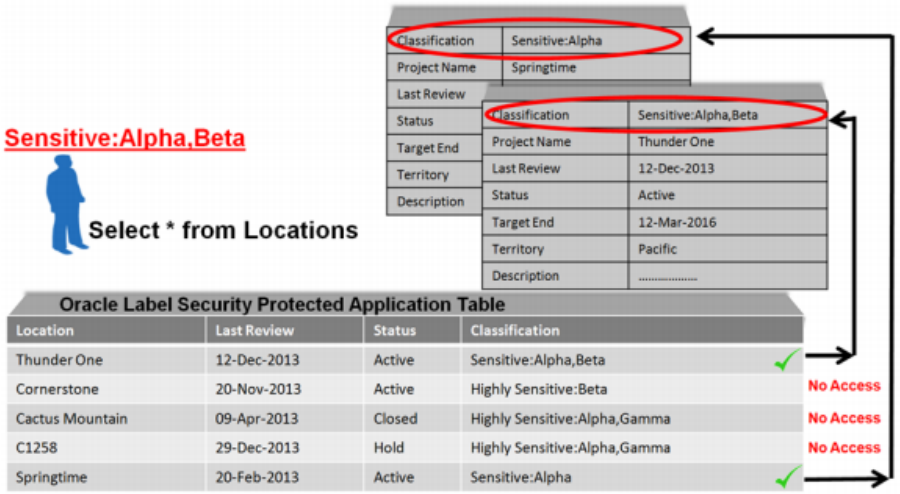


Figura 3.26. Oracle label security avaliando o acesso aos dados [Oracle 2018b]

3.4.2.3. Oracle - Categoria de Monitoramento/detecção

A LGPD determina que as organizações devem manter um registro de suas atividades de processamento. Esse registro só pode ser alcançado através do monitoramento e da auditoria constante das atividades sobre os dados pessoais. Esses dados de auditoria podem ser usados para notificar oportunamente as autoridades, em caso de violação. Além de exigir auditoria e alertas oportunos, a LGPD também exige que as organizações mantenham os registros de auditoria sob seu controle. Um controle centralizado dos registros de auditoria evita que invasores ou usuários mal-intencionados cubram os rastros de suas atividades suspeitas, excluindo registros da auditoria local [Rajasekharan 2017].

Oracle data safe

O *oracle data safe* auxilia na auditoria do banco de dados. A auditoria de atividades monitora as atividades dos usuários nos bancos de dados na nuvem, coletando e mantendo registros de auditoria por indústria e requisitos de conformidade regulamentar, acionando alertas para atividades não usuais. Por exemplo, a mudança em dados sensíveis pode ser auditada, caso ocorra falha no login de um administrador do banco de dados, pode ser gerado um alerta entre outros avisos possíveis. Na Figura 3.27 encontra-se um exemplo das atividades que podem ser monitoradas por esse serviço (ex.: quando o DBA desconecta do banco de dados (Event = LOGOFF); quando o DBA confirma uma transação (Event = COMMIT)).

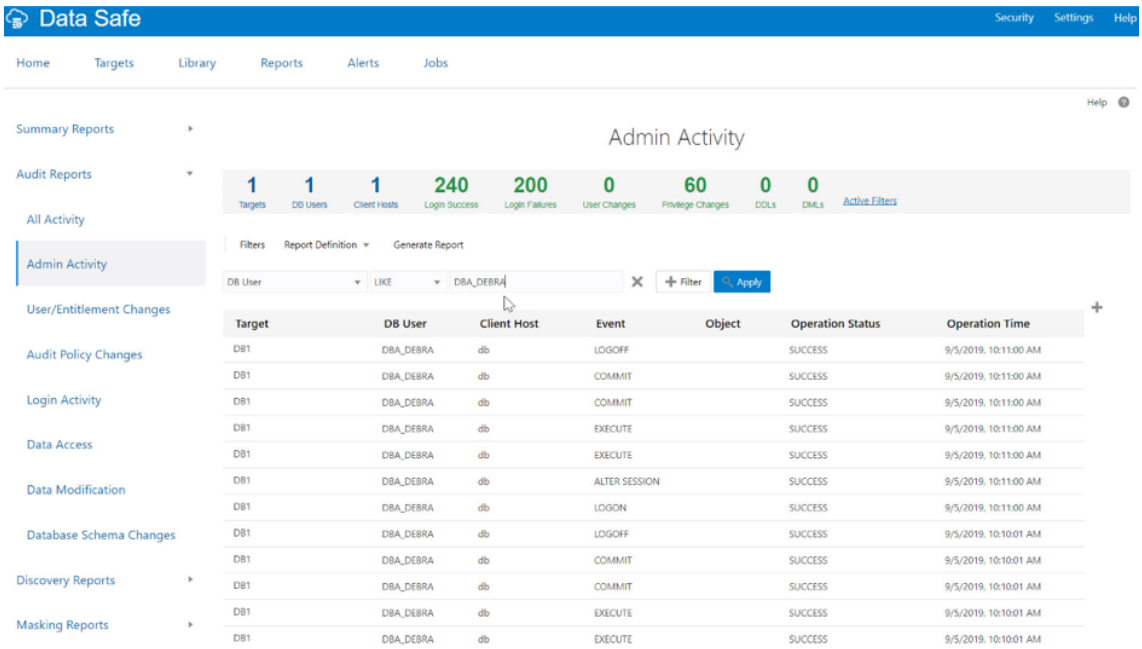


Figura 3.27. Exemplo de atividades monitoradas [Oracle 2019a]

Oracle audit vault e database firewall

O *oracle audit vault and database firewall* é uma plataforma de auditoria e proteção centrada em dados, que fornece monitoramento abrangente e flexível através de consolidação de dados de auditoria de bancos de dados Oracle e não Oracle, sistemas operacionais, sistemas de arquivos e aplicativos específicos [Rajasekharan 2017]. Ao mesmo tempo, o *oracle database firewall* pode atuar como a primeira linha de defesa na rede, impor o comportamento esperado do aplicativo, ajudando a impedir a injeção de SQL, o desvio do sistema e outras atividades que possam alcançar o banco de dados. O *oracle audit vault e o database firewall* podem consolidar os dados de auditoria de vários bancos de dados (Oracle, SQL Server, MySQL, DB2 entre outros) e monitora o tráfego SQL procurando, alertando e impedindo SQL não autorizado ou fora da política de segurança. Os responsáveis pela proteção de dados e os controladores podem especificar as condições sob as quais os alertas podem ser gerados em tempo real, tentando capturar os intrusos com as atividades anormais.

3.4.3. Amazon Web Services (AWS)

A AWS disponibiliza mais de 500 recursos e serviços com foco na segurança e compatibilidade. Segundo a categorização da LGPD, alguns dos serviços disponíveis são:

- Avaliação: *amazon macie* e *amazon inspector*.
- Prevenção: *AWS identity and access management (IAM)* e *aws key management service (KMS)*.
- Monitoramento/Detecção: *amazon guardduty e aws config*.

3.4.3.1. AWS - Categoria de Avaliação

O *amazon macie* é um serviço de segurança que usa aprendizado de máquina para descobrir, classificar e proteger, automaticamente, dados confidenciais na AWS [aws

2019a]. A Figura 3.28 apresenta um gráfico gerado a partir deste serviço sobre o comportamento de um usuário.

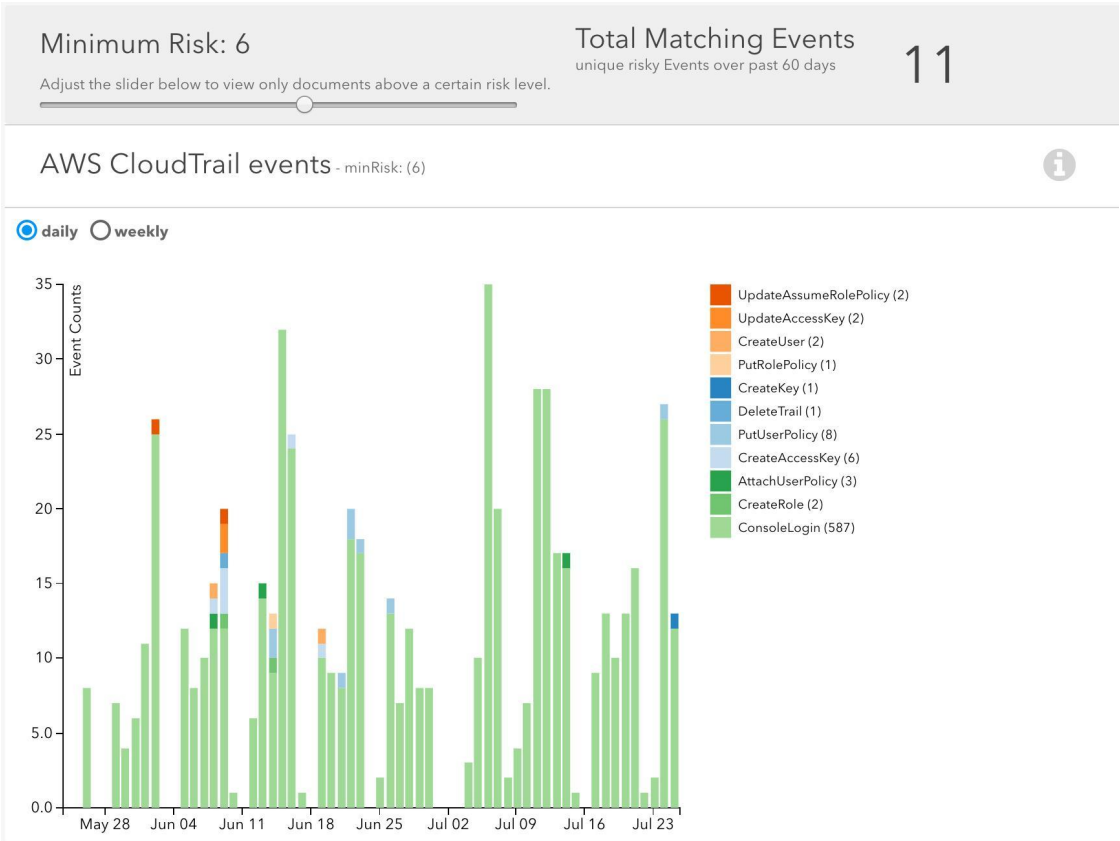


Figura 3.28. Análise do comportamento do usuário [aws 2019a]

O *amazon inspector* é um serviço de avaliação de segurança automático que ajuda a melhorar a segurança e a conformidade das aplicações implantadas na AWS [aws 2019b]. O *amazon inspector* avalia automaticamente as aplicações em busca de exposições, vulnerabilidades ou discrepâncias em relação às melhores práticas. Após realizar uma avaliação, o *amazon inspector* produz uma lista detalhada de descobertas de segurança priorizadas de acordo com o nível de severidade. A Figura 3.29 mostra um exemplo dessa lista. Caso o usuário selecione um dos itens que apresenta vulnerabilidade, é exibido um detalhamento sobre ele (Figura 3.30), com a descrição da vulnerabilidade (sujeito a ataques remotos por causa da função `bergetnext`) e a recomendação do serviço para o reparo da vulnerabilidade (atualizar o sistema operacional).

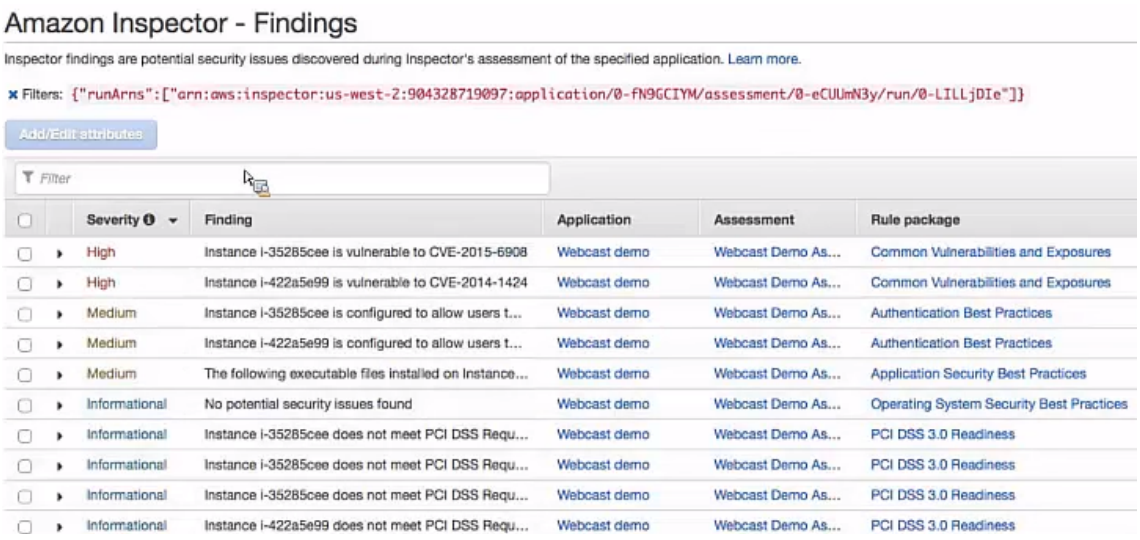


Figura 3.29. Lista de vulnerabilidades encontradas em uma aplicação [aws 2015]

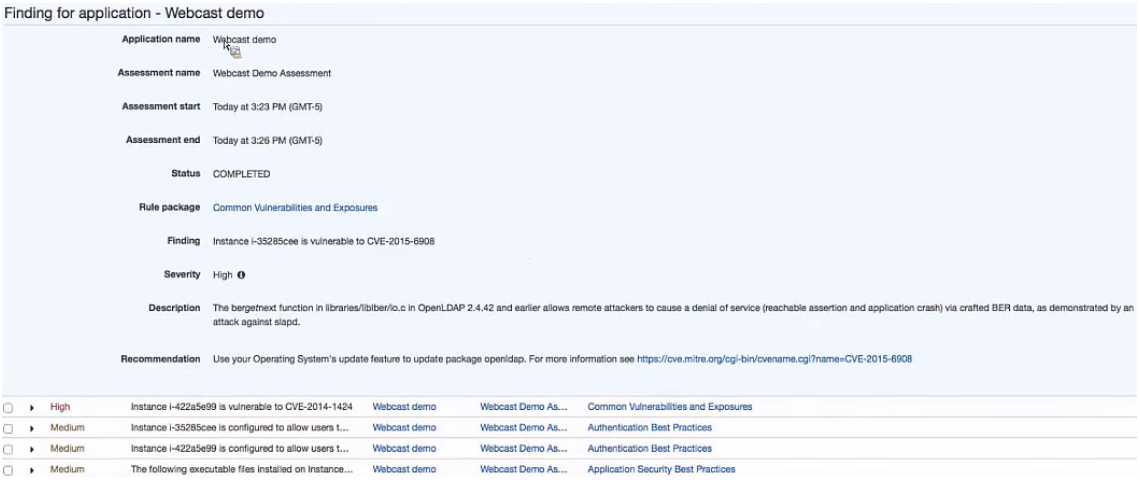


Figura 3.30. Detalhamento de uma das vulnerabilidades encontradas [aws 2015]

3.4.3.2. AWS - Categoria de Prevenção

O serviço *aws identity and access management (IAM)* permite a gerência, com segurança, do acesso aos serviços e recursos da AWS [aws 2019c]. Usando o IAM, pode-se criar e gerenciar usuários e grupos da AWS e usar permissões para conceder e negar acesso a recursos da AWS. A Figura 3.31 mostra um exemplo de atribuição de permissão ao usuário `testIAMuser`. O serviço permite adicionar o usuário a um grupo/papel pré-existente (ex.: `administrators`), copiar as permissões de um usuário existente ou atribuir políticas de segurança existentes de forma direta ao usuário.

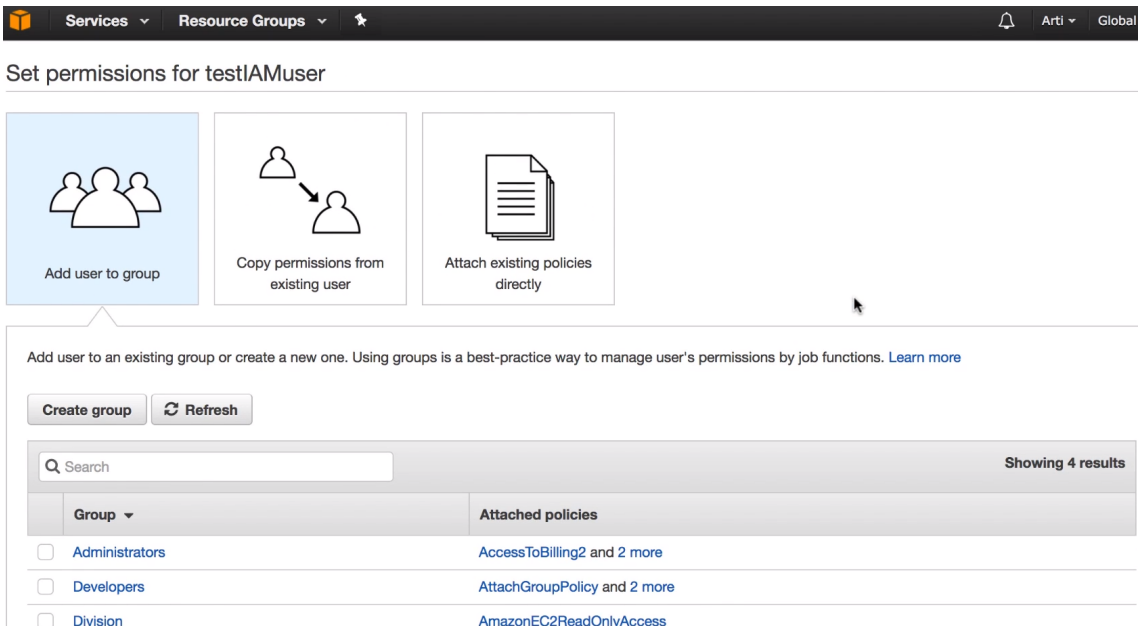


Figura 3.31. Exemplo de atribuição de permissão a um usuário [aws 2019c]

O serviço *aws key management service (KMS)* facilita a criação e o gerenciamento de chaves e o controle do uso de criptografia em uma ampla variedade de serviços da AWS e em seus aplicativos [aws 2019f]. A Figura 3.32 resume o comportamento desse serviço [stackoverflow 2018]. Existem dois tipos de chaves KMS: chaves mestras do cliente (CMKs) e chaves de dados (DKs). As chaves mestras do cliente nunca saem da infraestrutura da AWS e são geradas por chamada da API `CreateKey`. As chaves de dados são geradas por chamada da API `GenereateDataKey` que retorna uma versão "simples" e uma versão criptografada da chave. Essa criptografia é feita usando um CMK.

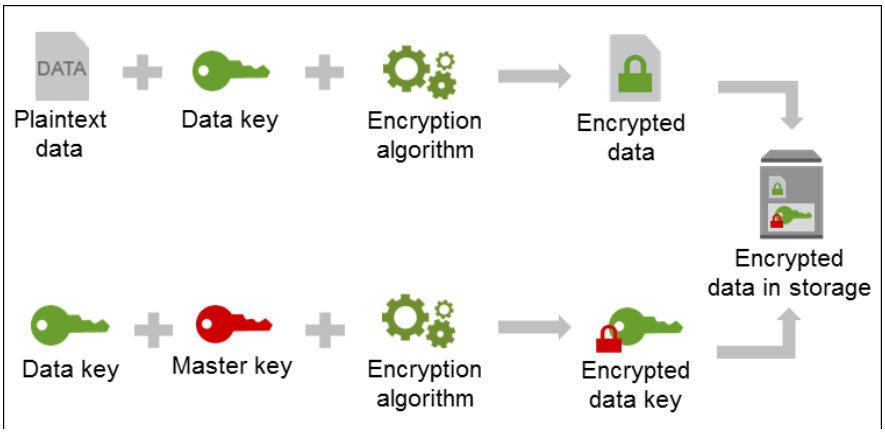


Figura 3.32. Visão geral do aws key management service [stackoverflow 2018]

3.4.3.3. AWS - Categoria de Monitoramento/Detecção

O *amazon guardduty* é um serviço de detecção de ameaças que monitora continuamente atividades mal-intencionadas ou e comportamentos não autorizados para proteger suas contas e cargas de trabalho da AWS [aws 2019d]. O serviço usa *machine learning*, detecção de anomalias e inteligência integrada contra ameaças para identificar e priorizar possíveis ameaças. A Figura 3.33 relata algumas ameaças encontradas e ordenadas por prioridade e a Figura 3.34 detalha uma das ameaças.

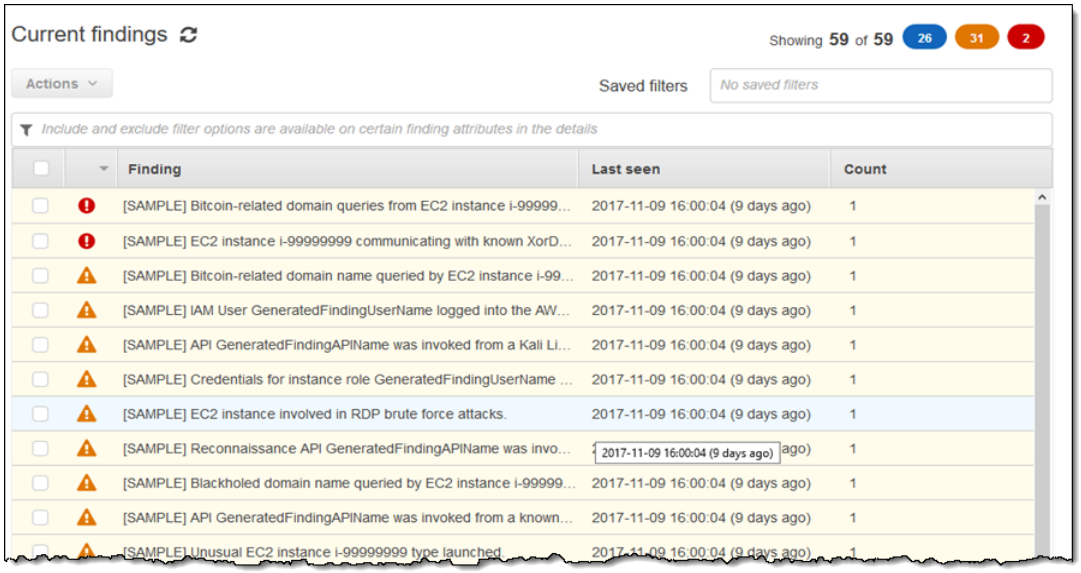


Figura 3.33. Ameaças detectadas [Barr 2017]

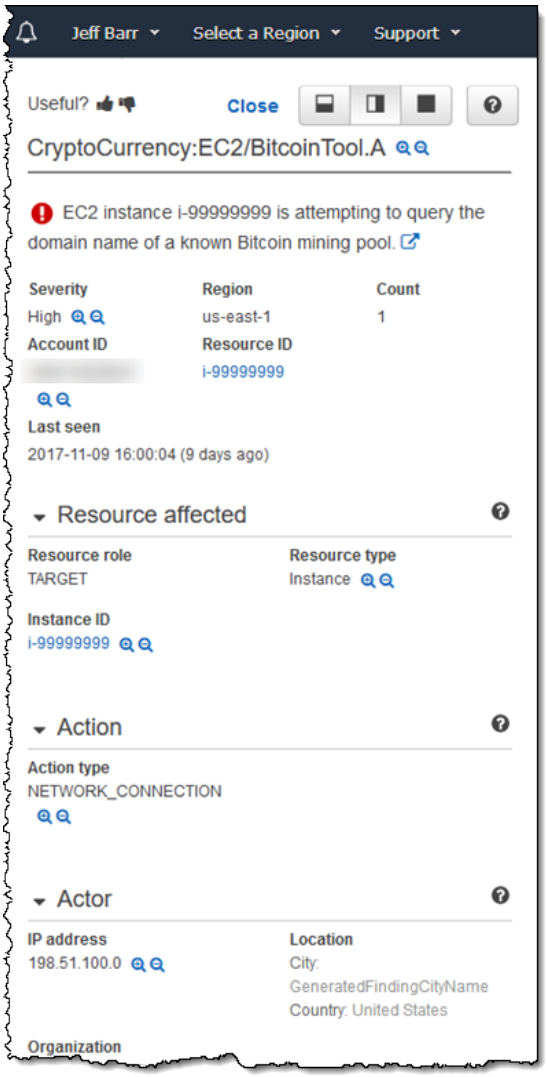


Figura 3.34. Detalhamento de uma das ameaças encontradas [Barr 2017]

O *aws config* é um serviço que permite acessar, auditar e avaliar as configurações dos recursos da AWS [aws 2019e]. O *aws config* monitora e grava continuamente os registros das configurações de recursos da AWS e lhe permite automatizar a avaliação das configurações registradas com base nas configurações desejadas. A Figura 3.35 demonstra uma visão geral da tela principal (*Dashboard*) do serviço *aws config*. Nessa tela, é possível (A) visualizar o número total de recursos que o *aws config* está registrando; (B) ver os tipos de recursos que o *aws config* está registrando, em ordem decrescente (por número de recursos). Caso o usuário selecione um tipo de recurso, o serviço abre a página de inventário de recursos; (C) escolher a exibição de todos os recursos também abre a página de inventário de recursos; (D) ver o número de regras não compatíveis; (e) ver o número de recursos não compatíveis; (f) ver as principais regras não compatíveis, em ordem decrescente (por número de regras); (G) escolher a exibição de todas as regras incompatíveis abre a página de regras.

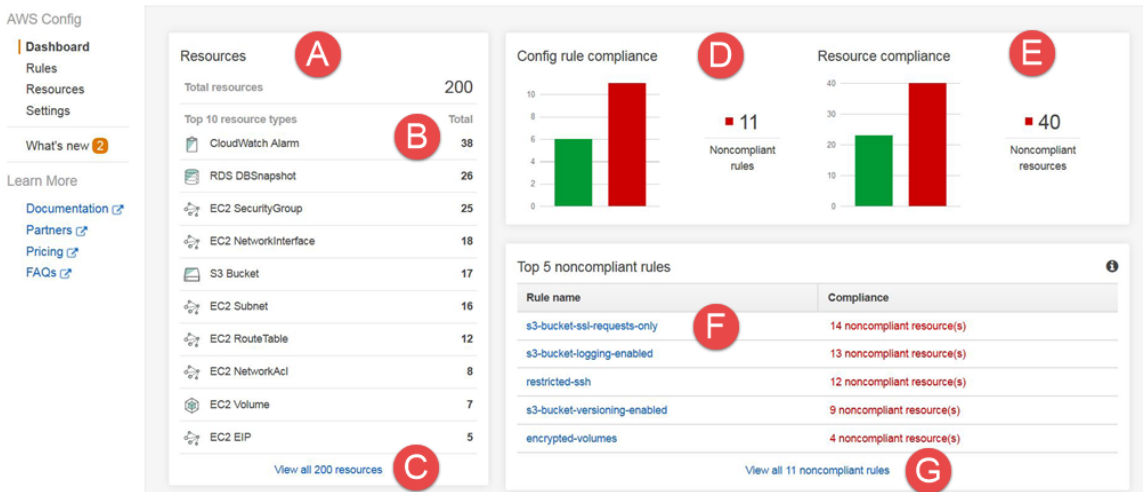


Figura 3.35. Dashboard do aws config [aws 2019e]

3.4.4. PostgreSQL

Existem algumas iniciativas de empresas e comunidades para auxiliar os usuários do SGBD relacional e gratuito PostgreSQL a se adaptarem às leis de proteção de privacidade. Entre as funcionalidades disponibilizadas, descreve-se neste capítulo, segundo as categorias da LGPD:

- Avaliação: até a escrita deste capítulo, não foram encontradas contribuições específicas dessa categoria.
- Prevenção: redação de dados e anonimização (*EnterpriseDB* e *PostgreSQL Anonymizer*); e criptografia de dados (*FUJITSU Enterprise Postgres* e *pgcrypto*).
- Monitoramento/Detecção: *pgAudit*.

3.4.4.1. PostgreSQL - Prevenção

Em termos de redação de dados, a empresa ***EnterpriseDB*** descreve uma forma, usando funções, visões, papéis e esquema padrão, para realizá-la de uma maneira que somente os usuários privilegiados consigam visualizar os dados no seu formato original [Linster 2018]. Os demais usuários, ao consultar os dados, visualizam de forma mascarada. Na Figura 3.36 tem-se um exemplo de função para mascarar a coluna `ssn` da tabela de

empregado (`employees`). Ela substitui todos os cinco primeiros números do `ssn` por 'x'. Adicionalmente, é incluído, ao final da função, o "`SECURITY DEFINER`" para especificar que a função deve ser executada com os privilégios do usuário que a possui.

```
CREATE OR REPLACE FUNCTION redact_ssn (ssn varchar(11))
RETURNS varchar(11)
/* substitui 020-12-9876 por xxx-xx-9876 */
AS
$$ SELECT overlay (ssn placing 'xxx-xx' FROM 1) ;$$
LANGUAGE SQL SECURITY DEFINER;
```

Figura 3.36. Função de redação ou mascaramento dos dados de ssn [Linster 2018]

Em seguida, é necessário criar uma visão da tabela de empregado (Figura 3.37) que chame a função definida anteriormente para a coluna `ssn` [Linster 2018]. Existem funções similares para as colunas de telefone (`phone`) e data de nascimento (`birthday`).

```
CREATE OR REPLACE VIEW redacteddata.employees
AS
SELECT
id,
name,
redact_ssn(ssn) ssn,
redact_phone(phone) phone,
redact_date(birthday) birthday
FROM employeedata.employees;
```

Figura 3.37. Exemplo de visão que realiza a chamada para as funções de mascaramento [Linster 2018]

Posteriormente, os usuários comuns obtêm acesso à visão criada e os usuários privilegiados obtêm acesso à tabela original. Além disso, o esquema padrão do papel do usuário comum passa a ser o esquema da visão (`redacteddata`) e o esquema padrão do papel do usuário privilegiado passa a ser o esquema dos dados originais (`employeedata`). A indicação dos esquemas é mostrada na Figura 3.38. Finalmente, os dados podem ser consultados pelos usuários comuns (Figura 3.39) e pelos usuários privilegiados (Figura 3.40).

```
ALTER ROLE redacteduser IN DATABASE mycompany SET search_path TO "$user", public, redacteddata;
ALTER ROLE privilegeduser IN DATABASE mycompany SET search path TO "$user", public, employeedata;
```

Figura 3.38. Alteração de esquema padrão para as papéis [Linster 2018]

```
SELECT * FROM employees;
id | name          | ssn         | phone      | birthday
---+---------------+-------------+------------+---------------------
 1 | Sally Sample  | xxx-xx-9345 | 5081234567 | 02-FEB-02 00:00:00
 2 | Jane Doe      | xxx-xx-9345 | 6171234567 | 14-FEB-02 00:00:00
 3 | Bill Foo      | xxx-xx-9345 | 9781234567 | 14-FEB-02 00:00:00
(3 rows)
```

Figura 3.39. Dados consultados por usuários comuns [Linster 2018]

```
SELECT * FROM employees;
id | name         | ssn         | phone      | birthday
---+--------------+-------------+------------+---------------------
 1 | Sally Sample | 020-78-9345 | 5081234567 | 02-FEB-61 00:00:00
 2 | Jane Doe     | 123-33-9345 | 6171234567 | 14-FEB-63 00:00:00
 3 | Bill Foo     | 123-89-9345 | 9781234567 | 14-FEB-63 00:00:00
(3 rows)
```

Figura 3.40. Dados consultados por usuários privilegiados [Linster 2018]

Outra forma de redação de dados confidenciais pode ser vista na extensão chamada *postgresql anonymizer* [Clochard 2018]. A Figura 3.41 mostra os dados originais. A Figura 3.42 mostra como a extensão pode ser criada e ativada no SGBD. Em seguida, na Figura 3.43, tem-se a criação de um papel para o usuário que verá os dados mascarados. Na Figura 3.44 apresenta-se a declaração das regras de mascaramento, onde o nome terá seus caracteres substituídos de forma randômica e o telefone só terá os 2 primeiros e os 2 últimos caracteres apresentados de forma real. Finalmente, na Figura 3.45, os dados são apresentados de forma mascarada.

```
SELECT * FROM people;
id   | name           | phone
-----+----------------+------------
T800 | Schwarzenegger | 0609110911
(1 row)
```

Figura 3.41. Dados originais [Clochard 2018]

```
CREATE EXTENSION IF NOT EXISTS anon CASCADE;
SELECT anon.mask_init();
```

Figura 3.42. Criação e ativação da extensão PostgreSQL Anonymizer [Clochard 2018]

```
CREATE ROLE skynet;
COMMENT ON ROLE skynet IS 'MASKED';
```

Figura 3.43. Criação do papel do usuário que verá os dados mascarados [Clochard 2018]

```
COMMENT ON COLUMN people.name IS 'MASKED WITH FUNCTION anon.random_last_name()';
COMMENT ON COLUMN people.phone IS 'MASKED WITH FUNCTION anon.partial(phone, 2, $$******$$, 2)';
```

Figura 3.44. Descrição do mascaramento das colunas de nome e telefone [Clochard 2018]

```
psql test -U skynet -c 'SELECT * FROM people;'
id   | name           | phone
-----+----------------+------------
T800 | Nunziata       | 06******11
(1 row)
```

Figura 3.45. Exibição de dados mascarados [Clochard 2018]

A criptografia de dados pode ser obtida através da aquisição da camada TDE proposta pela ***FUJITSU Enterprise Postgres*** [Downey 2019]. Essa camada não demanda alterações no SGBD e é disponibilizada de forma gratuita pela empresa. A Figura 3.46 mostra um exemplo da arquitetura com essa camada. Como pode ser visto na figura, os dados podem ser armazenados no banco de dados e no arquivo de backup de forma criptografada e somente os usuários autorizados que conseguem visualizar os dados originais. A criptografia pode ocorrer em nível de *tablespaces*, dados de backup, *log* de registro prévio de escrita (WAL), arquivos temporários e replicação de *streaming*.

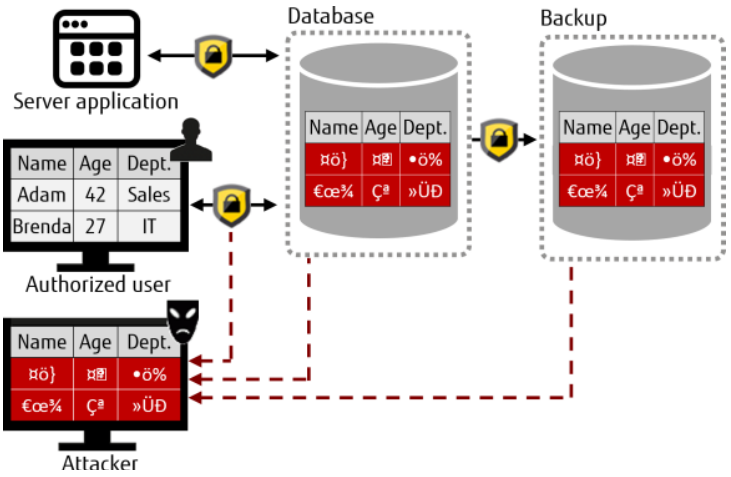


Figura 3.46. Arquitetura da camada TDE do FUJITSU Enterprise Postgres [Downey 2019]

Além disso, o PostgreSQL 10 possui algumas opções de criptografia, na própria camada do banco de dados, através do módulo *pgcrypto*, para: senha, colunas específicas, *storage* (nível de sistema e de bloco), dos dados durante o tráfego de rede, autenticação do host (cliente e servidor) por certificado e do lado do cliente [PostgreSQL 2019].

A Figura 3.47 mostra um exemplo de inserção de senha, usando as funções `crypt()` e `gen_salt()` para senhas *hashing*. A primeira função faz o *hashing* e a segunda prepara os parâmetros do algoritmo para ela.

```
INSERT INTO "login" (login, password, employee_id)
VALUES ('email', crypt('password', gen_salt('bf')));
```

Figura 3.47. Exemplo de criptografia de senha no banco de dados [stackoverflow 2013]

3.4.4.2. PostgreSQL - Monitoramento/Detecção

A auditoria de dados pode ser implementada através do módulo *pgAudit* [Riggs et al 2019]. Esse módulo é gratuito e provê um *log* detalhado de auditoria de objeto e/ou de sessão. Por exemplo, caso o(a) auditor(a) queira verificar qual a tabela que foi criada no período de janela de manutenção, ele(a) pode verificar o *log*, que informará os detalhes conforme demonstrado na Figura 3.48. É importante reparar que o usuário que criou a tabela tentou mascarar a execução do comando, ao incluí-lo em uma transação, com o uso do comando `EXECUTE`.

```
AUDIT: SESSION, 33, 1,FUNCTION, DO,,,"DO $$
BEGIN
    EXECUTE 'CREATE TABLE import' || 'ant_table (id INT)';
END $$."
AUDIT: SESSION, 33, 2,DDL, CREATE TABLE, TABLE, public.important table, CREATE TABLE important table(id INT)
```

Figura 3.48. Exemplo de registro no log de auditoria [Riggs et al 2019]

Na Figura 3.49 tem-se um exemplo de como o *log* de auditoria pode ser habilitado para todos os comandos DDL e de escrita (DML), descrevendo também as relações envolvidas nos comandos DML.

```
SET pgaudit.log = 'write, ddl';
SET pgaudit.log_relation = ON;
```

Figura 3.49. Comandos para habilitar o log de auditoria [Riggs et al 2019]

3.5. Conclusão

Este minicurso apresenta uma discussão sobre a LGPD e seus princípios, bem como os recursos disponíveis em alguns dos principais ambientes de bancos de dados existentes que podem auxiliar os controladores de dados sensíveis a manter o ambiente de banco de dados em conformidade com esta lei. Através das análises apresentadas no presente capítulo, percebe-se que os recursos para controle de acesso aos dados encontrados nos ambientes de Bancos de Dados analisados não são suficientes para contemplar os princípios da LGPD em relação a todas as questões de privacidade. Neste sentido é preciso também se preocupar com o sigilo, a integridade, o tempo de vida, o anonimato, o escopo de uso pela aplicação e a separação de funções dos dados.

Como já descrito por Elmasri e Navathe (2019), o avanço rápido do uso da tecnologia da informação (TI) na indústria, no governo e no meio acadêmico gera questões e problemas desafiadores com relação à proteção e ao uso de informações pessoais. Questões como quem e quais direitos à informação sobre indivíduos, para quais finalidades, tornam-se cada vez mais importantes à medida que seguimos para um mundo em que é tecnicamente possível conhecer quase tudo sobre qualquer um. Decidir como projetar considerações de privacidade na tecnologia para o futuro inclui dimensões filosóficas, legais e práticas.

Dessa forma, as organizações ainda se questionam qual o setor responsável em sua estrutura pela implantação e manutenção da LGPD, pela multidisciplinaridade inerente ao tema. Embora não tenha sido o foco do presente minicurso, essa questão ainda fica em aberto para futuras discussões.

Referências

Adithe, S., Singh, S. (2018) “Comprehensive Identity and Access Management in the Cloud”, SAPinsider, Volume 19, Issue 2, disponível em: https://sapinsider.wispubs.com/Assets/Articles/2018/May/Comprehensive-Identity-and-Access-Management-in-the-Cloud, acessado em outubro de 2019.

Amazon Web Services (AWS). (2019a) “Detalhes do Amazon Macie”, disponível em: https://aws.amazon.com/pt/macie/details/, acessado em outubro de 2019.

Amazon Web Services (AWS). (2019b) “Amazon Inspector - Guia do usuário”, disponível em: https://docs.aws.amazon.com/pt_br/inspector/latest/userguide/inspector_introduction.html, acessado em outubro de 2019.

Amazon Web Services (AWS). (2019c) “Conceitos básicos do AWS IAM”, disponível em: https://aws.amazon.com/pt/iam/getting-started/, acessado em outubro de 2019.

Amazon Web Services (AWS). (2019d) “Amazon GuardDuty”, disponível em: https://aws.amazon.com/pt/guardduty/, acessado em outubro de 2019.

Amazon Web Services (AWS). (2019e) “AWS Config – Guia do desenvolvedor”, disponível em: https://docs.aws.amazon.com/pt_br/config/latest/developerguide/viewing-the-aws-config-dashboard.html, acessado em outubro de 2019.

Amazon Web Services (AWS). (2019f) “Recursos do AWS Key Management Service”, disponível em: https://aws.amazon.com/pt/kms/features/, acessado em outubro de 2019.

Amazon Web Services (AWS). (2015) “2015 Webinar Series – AWS Inspector”, disponível em: https://www.youtube.com/watch?v=ddz0JmCTTsU, acessado em outubro de 2019.

Barr, J. (2017) “Amazon GuardDuty – Continuous Security Monitoring & Threat Detection”, disponível em: https://aws.amazon.com/pt/blogs/aws/amazon-guardduty-continuous-security-monitoring-threat-detection/, acessado em outubro de 2019.

Clochard, D. (2018) “Introducing PostgreSQL Anonymizer”, disponível em: https://blog.taadeem.net/english/2018/10/29/Introducing-PostgreSQL-Anonymizer, acessado em outubro de 2019.

Dean, B. (2017) “Privacy vs. Security”, disponível em: https://www.secureworks.com/blog/privacy-vs-security, acessado em outubro de 2019.

Downey, P. (2019) “Providing maximum data security with minimal impact to your business using transparent data encryption”, disponível em: https://www.postgresql.fastware.com/blog/transparent-data-encryption-tde, acessado em outubro de 2019.

Elmasri, R., Navathe, S. B. (2019) “Sistemas de Banco de Dados”, 7ª edição (versão traduzida), editora Pearson Universidades.

Feinberg, D., Heudecker, N., Adrian, M. (2018) “Magic Quadrant for Operational Database Management Systems”, In: Gartner Research, ID: G00346575, disponível em: https://www.gartner.com/en/documents/3891967, acessado em outubro de 2019.

Granet, E. (2018) “Static Data Masking for Azure SQL Database and SQL Server”, disponível em: https://azure.microsoft.com/pt-br/blog/static-data-masking-preview/, acessado em outubro de 2019.

Linster, M. (2018) “Creating a Data Redaction Capability to Meet GDPR Requirements Using EDB Postgres”, disponível em: https://www.enterprisedb.com/blog/creating-data-redaction-capability-meet-gdpr-requirements-using-edb-postgres, acessado em outubro de 2019.

Mahajan, G. (2019) “SQL Server Static Data Masking Example”, disponível em : https://www.mssqltips.com/sqlservertip/5939/sql-server-static-data-masking-example/, acessado em outubro de 2019.

Microsoft. (2019a) “Descoberta e classificação de dados SQL”, disponível em: https://docs.microsoft.com/pt-br/sql/relational-databases/security/sql-data-discovery-and-classification?view=sql-server-2017, acessado em outubro de 2019.

Microsoft. (2019b) “Segurança em nível de linha’, disponível em: https://docs.microsoft.com/pt-br/sql/relational-databases/security/row-level-security?view=sql-server-2017, acessado em outubro de 2019.

Microsoft. (2019c) “Mascaramento de dados dinâmicos”, disponível em: https://docs.microsoft.com/pt-br/sql/relational-databases/security/dynamic-data-masking?view=sql-server-ver15, acessado em outubro de 2019.

Microsoft (2019d) “Habilitar conexões criptografadas com o Mecanismo de Banco de Dados”, disponível em: https://docs.microsoft.com/pt-br/sql/database-

engine/configure-windows/enable-encrypted-connections-to-the-database-engine?view=sql-server-ver15, acessado em outubro de 2019.

Microsoft (2019e) "Criptografia de Dados Transparente (TDE)", disponível em: https://docs.microsoft.com/pt-br/sql/relational-databases/security/encryption/transparent-data-encryption?view=sql-server-ver15, acessado em outubro de 2019.

Microsoft. (2018a) "SQL Server and Azure SQL Database GDPR Guidance", disponível em: https://azurepartnerportal.blob.core.windows.net/media/Resources/SQL%20Server%20GDPR%20Guidance%20Paper.pdf, acessado em outubro de 2019.

Microsoft. (2018b) "Autenticação no SQL Server", disponível em: https://docs.microsoft.com/pt-br/dotnet/framework/data/adonet/sql/authentication-in-sql-server, acessado em outubro de 2019.

Microsoft. (2017a) "Server and Database Roles in SQL Server", disponível em: https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/sql/server-and-database-roles-in-sql-server, acessado em outubro de 2019.

Microsoft. (2017b) "Create Database Audit Specification (Transact-SQL)", disponível em: https://docs.microsoft.com/pt-br/sql/t-sql/statements/create-database-audit-specification-transact-sql?view=sql-server-ver15, acessado em outubro de 2019.

Microsoft. (2017c) "Sempre criptografados (mecanismo de banco de dados)", disponível em: https://docs.microsoft.com/pt-br/sql/relational-databases/security/encryption/always-encrypted-database-engine?view=sql-server-ver15, acessado em outubro de 2019.

Microsoft. (2017d) "Avaliação de Vulnerabilidades SQL", disponível em: https://docs.microsoft.com/pt-br/sql/relational-databases/security/sql-vulnerability-assessment?view=sql-server-ver15, acessado em outubro de 2019.

Microsoft. (2016a) "Auditoria do SQL Server (Mecanismo de Banco de Dados)", disponível em: https://docs.microsoft.com/pt-br/sql/relational-databases/security/auditing/sql-server-audit-database-engine?view=sql-server-ver15, acessado em outubro de 2019.

Microsoft. (2016b) "Tabelas temporais", disponível em: https://docs.microsoft.com/pt-br/sql/relational-databases/tables/temporal-tables?view=sql-server-ver15, acessado em outubro de 2019.

Microsoft (2014) "SQL Server – Object Level Permissions details", disponível em: https://gallery.technet.microsoft.com/scriptcenter/SQL-Server-Object-Level-fc2f1cb6, acessado em outubro de 2019.

Oracle. (2019a) "Secure Critical Data with Oracle Data Safe – Improve the Security of Cloud Databases with a Unified Control Center for Managing Sensitive Data", White Paper disponível em: https://www.oracle.com/a/tech/docs/dbsec/data-safe/wp-security-data-safe.pdf, acessado em outubro de 2019.

Oracle. (2019b) "Encryption and Redaction with Oracle Advanced Security", White Paper disponível em: https://www.oracle.com/a/tech/docs/dbsec/aso/advanced-security-wp-19c.pdf, acessado em outubro de 2019.

Oracle. (2019c) “Oracle Database Vault”, White Paper disponível em: https://www.oracle.com/a/tech/docs/dbsec/dbv/wp-dv-19c.pdf, acessado em outubro de 2019.

Oracle. (2018a) “Managing Oracle Database Encryption Keys in Oracle Cloud Infrastructure with Oracle Key Vault”, White Paper disponível em: https://docs.cloud.oracle.com/iaas/Content/Resources/Assets/whitepapers/manage-encryption-keys-oci-okv.pdf, acessado em outubro de 2019.

Oracle (2018b) “Oracle Label Security”, White Paper disponível em: https://www.oracle.com/technetwork/wp-dbsec-ols-201702-3634252.pdf, acessado em outubro de 2019.

Oracle. (2017) “Data Masking and Subsetting Guide”, disponível em: https://docs.oracle.com/en/database/oracle/oracle-database/12.2/dmksb/index.html, acessado em outubro de 2019.

PostgreSQL. (2019) “PostgreSQL 10 - Encryption Options”, disponível em: https://www.postgresql.org/docs/10/encryption-options.html, acessado em outubro de 2019.

Rajasekharan, D. (2017) “Accelerate Your Response to the EU General Data Protection Regulation (GDPR)”, Oracle White Paper, disponível em: https://www.oracle.com/technetwork/database/security/wp-security-dbsec-gdpr-3073228.pdf, acessado em outubro de 2019.

Riggs, S., Menon-Sem, A., Barwick, I. (2019) “pgAudit Open Source PostgreSQL Audit Logging”, disponível em: https://github.com/pgaudit/pgaudit/blob/master/README.md, acessado em outubro de 2019.

Stackoverflow. (2018) “Key Management Services”, disponível em: https://stackoverflow.com/questions/47904805/key-management-services, acessado em outubro de 2019.

Stackoverflow. (2013) “How do I encrypt passwords with PostgreSQL?”, disponível em: https://stackoverflow.com/questions/18656528/how-do-i-encrypt-passwords-with-postgresql, acessado em outubro de 2019.

Teixeira, B., Schwabe, D., Santoro, F., Baião, F., Luiza Campos, M., Verona, L., Laufer, C., Barbosa, S., Lifschitz, S., Costa, R. (2019). Privacy and Transparency within the 4IR: Two faces of the same coin. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 581-593). ACM.

Tutorialspoint. (2019) “SAP GRC Tutorial – SAP GRC - Overview”, disponível em: https://www.tutorialspoint.com/sap_grc/sap_grc_overview.htm, acessado em outubro de 2019.

Autores

Ana Carolina Brito de Almeida (Instrutora) é professora adjunta da Universidade do Estado do Rio de Janeiro (UERJ), alocada no Departamento de Informática e analista judiciário com especialidade em Informática no Tribunal Regional Federal da 2ª Região (TRF2), atuando como DBA. Realizou pós-doutorado com ênfase em *Big Data* na Universidade Federal do Rio de Janeiro (UFRJ) (Out/2014 a Abr/2015). Doutora em Informática com especialização em *Tuning* de BD pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) (2013). Mestre em Sistemas e Computação pelo Instituto Militar de Engenharia (IME/RJ) (2006). Pesquisadora na área de BD com ênfase em: (i) sistemas de (auto) sintonia de BD e (ii) ontologias. Já foi consultora em BD e ministrou cursos na Universidade Petrobras. Detalhes em http://lattes.cnpq.br/8306729029606464.

Letícia Dias Verona (Instrutora) possui mestrado em Informática pela UFRJ (2018) e graduação em Ciência da Computação pela UFRJ. Possui experiência com desenvolvimento de sistemas web, gestão de projetos e equipes. Pesquisadora em: integração de informações heterogêneas, análise de redes políticas e econômicas, metadados, ontologias, modelagem conceitual, BD e web semântica. Atualmente participa da coordenação do grupo de pesquisa de Dados Abertos, vinculado ao grupo GRECO (PPGI/UFRJ), cursando o doutorado em Gestão de Sistemas Complexos.
Detalhes em http://lattes.cnpq.br/2165808131875029.

Maria Luiza Machado Campos é professora no Departamento de Ciência da Computação da UFRJ, e uma das coordenadoras do grupo de pesquisas GRECO, atuando como pesquisadora e orientadora de mestrado e doutorado no PPGI da mesma universidade. Possui graduação em Engenharia Civil pela Universidade Federal do Rio Grande do Sul (1978), mestrado em Engenharia de Sistemas e Computação pela COPPE/UFRJ (1984) e doutorado em *Information Systems - University Of East Anglia* (1993), Inglaterra. Em 2015, realizou Pós-doutorado no *Laboratory of Applied Ontology*, CNRS, Italia. Foi coordenadora do Bacharelado em Ciência da Computação e do Programa de Pós-graduação em Informática, assim como Diretora Adjunta de Extensão do Instituto de Matemática da UFRJ. Participou de diversos projetos de desenvolvimento, pesquisa e extensão, assim como de numerosas orientações de trabalhos de conclusão de curso, dissertações de mestrado e de doutorado ao longo de mais de 30 anos de carreira. Seus principais temas de pesquisa estão associados à integração de informações heterogêneas, abordando principalmente os seguintes temas: metadados, ontologias, modelagem conceitual, banco de dados, *data warehousing* e web semântica. Detalhes em http://lattes.cnpq.br/0659658820912418.

Fernanda Araujo Baião é Professora no Departamento de Engenharia Industrial da PUC-Rio. Seus temas de pesquisa são nas áreas de Ciência de Dados, Modelagem Conceitual e Ontologias, Gestão de Processos de Negócio (BPM) e Integração de Dados Distribuídos através de Alinhamento de Ontologias, particularmente investigando a interação entre Ciências Cognitivas e Processamento de Linguagem Natural com BPM e Gestão de Dados, assim como o desenvolvimento de sistemas de Alinhamento de Ontologias para apoio à Integração de Dados sobre ambientes distribuídos de larga escala. De outubro de 2018 a Agosto de 2019 atuou como Cientista de Dados chefe em uma iniciativa na Caixa Econômica Federal para Prevenção de Fraudes no Programa Seguro desemprego. De 2004 a 2018 foi Professora da Universidade Federal do Estado do Rio de Janeiro. De 2001 a 2004 atuou como post-doc na COPPE/UFRJ, onde obteve seu título de Doutorado (2001) e de Mestrado (1997) em Engenharia de Sistemas e Computação. No ano de 2000 foi pesquisadora visitante na *University of Wisconsin-Madison* (USA). É autora de mais de 150 publicações com revisões por pares, muitas em colaboração com pesquisadores renomados na comunidade científica internacional resultantes da sua participação em projetos de pesquisa nacionais e internacionais, como o *Brazilian Institute of WebScience Research*, financiado pelo CNPq, e o Rise-BPM (rise-bpm.eu), financiado pela União Europeia. Coordena ou já coordenou projetos de pesquisa financiados pelo CNPq e FAPERJ. Participa e coordena diversos comitês de programa e e de editoração de conferências e periódicos nacionais e internacionais, como a Applied Ontology Journal, ER, BPM, dentre outras. Desenvolveu expertise valioso em projetos de transferência de conhecimento entre a Academia e a Indústria, e atuou como líder técnico em projetos de P&D sobre Data Science, BPM, Arquitetura Empresarial, Gestão de Dados e Segurança da Informação, em domínios de Exploração e Produção de Óleo e Gás, Seguros, Gestão de Serviços de TI e Predição de Fraudes. Detalhes em `http://lattes.cnpq.br/5068302552861597`.

Capítulo 4

Técnica de Ensino de Matemática para Alunos com Deficiência Visual com suporte Informatizado

Angélica Fonseca da Silva Dias, José Antonio dos Santos Borges e Júlio Tadeu Carvalho da Silveira

Abstract

This short course aims to introduce a methodology, easy to disseminate and reproduce, that produces better quality and comprehensiveness in the teaching of mathematics in inclusive classes with the visually impaired students. The course introduces techniques for writing and reading texts with mathematical and graphics content, with intensive use of specific computational tools that enable the formation of students, at the middle and higher levels, to meet acceptable criteria of quality and content, also favoring the access and permanence of blind and low vision students in math-based university careers.

Resumo

O objetivo deste minicurso é apresentar uma metodologia de fácil reprodução e disseminação, que permite aumentar a qualidade e abrangência do ensino de matemática em classes inclusivas com alunos com deficiência visual. O curso introduz técnicas de escrita e leitura de textos com conteúdo e gráficos matemáticos, com uso intensivo de ferramentas computacionais específicas que propiciam que a formação dos alunos, tanto no nível médio quanto superior, obedeça a critérios aceitáveis de qualidade e conteúdo, favorecendo também o acesso e permanência de estudantes cegos e com baixa visão nas carreiras universitárias com base matemática.

ADVERTÊNCIA: para leitura e editoração correta deste texto é necessário ter instalada a fonte SimBraille no computador.

4.1. A utilização de computadores por pessoas com deficiência visual

Desde 2006, a Sociedade Brasileira de Computação (SBC), vem patrocinando estratégias que promovam o aumento do acesso participativo e universal do cidadão brasileiro ao conhecimento. Reconhecendo a grande quantidade de pessoas no país que

convive com enormes diferenças sociais, financeiras, físicas e mentais, o que é levado em conta não é somente apoiar a produção de soluções inclusivas para adaptação de sistemas computacionais interativos para uso por pessoas com deficiência "que possam ser generalizados para múltiplos dispositivos", mas adequar os artefatos tecnológicos criados a esta enorme diversidade social. (Melo, A.M., 2014)

Nos últimos anos diversas áreas do ensino para pessoas com deficiência visual no Brasil utilizam cada vez mais sistemas computadorizados com síntese de voz para atender a alunos com deficiência visual. Os sintetizadores vêm pré-configurados para produzir uma fala razoavelmente precisa de textos convencionais na língua portuguesa, tanto em termos fonéticos quanto prosódicos, permitindo uma leitura fluente e agradável (Ferreira, 2014).

No Brasil, entre os sistemas de maior utilização está o Dosvox (NCE-2010), um sistema composto por cerca de 100 programas que atendem a grande parte das necessidades computacionais de uma pessoa com deficiência visual. Ele usa apenas a síntese de voz e o teclado para permitir a comunicação completa entre o computador e o usuário que não enxerga. Neste processo ele utiliza um estilo de comunicação muito intuitivo, baseado no uso de menus interativos falados, que conduzem o usuário por opções que são selecionadas através das setas do teclado ou por abreviaturas de uma única letra (Borges, 2009), O Dosvox não tenta simplesmente sonorizar os ícones e textos apresentados na tela, mas apresenta um diálogo sonoro facílimo de assimilar e de interagir, mesmo quando é usado por crianças pequenas ou pessoas com variados níveis de deficiência visual. (Dias et al 2016).

Figura 4.1 - DOSVOX

Por outro lado, esse diálogo específico tem que ser construído para cada função que deve ser oferecida. Isso o torna, por um lado, um sistema de operação muito simples, agradável e de rápida curva de aprendizado; por outro lado, à medida que a Informática evolui, novos programas precisam ser construídos, o que nem sempre ocorre, tanto por dificuldades técnicas quanto pelo esforço e recursos envolvidos.

Em direção oposta à operação simplificada do Dosvox, existem diversos programas conhecidos por "leitores de tela" no qual o teclado é usado para escolher interativamente e sonorizar um dos muitos elementos desenhados ou escritos na tela. A forma mais usual de operação é caminhar entre os itens desenhados usando as setas, caminhando sequencialmente entre os itens, ou a tecla TAB que pula para o próximo item clicável (link, por exemplo). É possível também usar sequências predefinidas de teclas (ou seja, teclas de atalho) que vai conduzindo o processo de leitura de lugares específicos do display (por exemplo, a barra de ferramentas) ou controlando uma leitura sequencial dos elementos (por exemplo, o texto de uma página da web).

Os leitores de tela mais usados no Brasil são o NVDA (produto gratuito, de origem australiana), o Virtual Vision (brasileiro) e o Jaws (americano). Mesmo tendo teclas de controle um pouco diferentes, comparativamente, seu uso não é muito diferente, até

porque a sequência de navegação é predefinida pelos utilitários que eles estão lendo na tela e pelo sistema operacional.

A vantagem dos leitores de tela é que a operação é feita exatamente sobre o que está desenhado na tela. Mas, pelo fato de que a localização é feita sem ver, o usuário precisa ser treinado previamente em cada programa a utilizar, tendo em geral que decorar uma quantidade enorme de informações para um uso razoável do sistema. Em outras palavras, a curva de aprendizado é muito mais lenta, a operação é mais complexa, mas o acesso é mais completo.

Infelizmente, tanto o Dosvox quanto os leitores de tela não vem dando um suporte razoável para a escrita matemática a ser utilizada por cegos. A razão mais importante é que o texto matemático tem várias características peculiares, que não obedecem às regras de leitura e escrita usuais da língua portuguesa.

Em primeiro lugar, a escrita de matemática é tipicamente não linear, ou seja, escrevem-se textos matemáticos com os elementos posicionados fora de uma linha única (há índices em cima, subíndices em baixo, frações, somatórios, etc.; em segundo lugar, ela exibe ambiguidades que exigem uma fala não coloquial, e muitas vezes peculiar. Por exemplo, $2^{1/2}$ (dois elevado a um meio) deve ser falado diferentemente de 2 1/2 (dois inteiros e um meio). (ECS, 2015). (Dias *et al*, 2018).

4.2 Cegos em carreiras com base matemática (STEM[1])

Observamos que, graças ao uso de tecnologia computacional, foi viabilizada a presença de pessoas cegas em muitos cursos universitários, especialmente naqueles em que os textos de suporte não são matemáticos, que, já vimos, não vinham, até pouco tempo, sendo bem suportados. Infelizmente, isso não vale para as carreiras em que a matemática é um elemento central, em que o número de pessoas cegas cursando é próximo de zero.

Será que a razão é que o raciocínio exigido nas disciplinas aqui envolvidas é incompatível com as restrições impostas pela cegueira? Ou seria apenas a falta de suporte computacional adequado a causa principal de haver tão poucos estudantes cegos em cursos das carreiras com base matemática? (Cryer, 2013)

Há exemplos que desmentem a primeira conjectura. Talvez o mais impressionante seja o do matemático inglês Nicholas Saunderson (1682-1739), que viveu em uma época em que não havia tecnologia de escrita para cegos. Mesmo assim, ele é considerado um dos maiores matemáticos da história – pela Universidade de Cambridge (Reino Unido). É importante salientar que Saunderson era o que conhecemos hoje como *superdotado,* alguém capaz de formular e resolver equações matemáticas complexas mentalmente, sem usar a escrita para auxiliar o raciocínio (ver Wikipedia).

Uma pessoa mediana, entretanto, não consegue adquirir estas facilidades de abstração e cálculo mental senão com muitos anos de treinamento especializado. O que é necessário então é desenvolver um ferramental diferenciado, facilmente operável por pessoas cegas simples e até com baixa cultura, e que dê conta dos detalhes para permitir independência no estudo. É preciso que a pessoa, possivelmente com a ajuda do computador, seja capaz de escrever e ler textos matemáticos de mediana complexidade, sozinha e com segurança.

---

[1] *STEM – abreviatura de science, technology, engineering and mathematics*.

Somente há pouco tempo é que estas soluções computacionais vêm aparecendo e trazendo muito boas perspectivas para os alunos com deficiência visual. Através delas, prevê-se a possibilidade de que, em pouco tempo, seja possível viabilizar a presença bem sucedida de uma pessoa cega em classes mistas, com total proficiência de leitura e escrita de textos envolvendo matemática de razoável complexidade.

4.3 Apresentação dos principais temas estudados neste curso

Este texto descreve as iniciativas computacionais com algumas soluções baseadas no uso de informática para o ensino de matemática para cegos e algumas questões relacionadas à educação inclusiva e aprendizagem colaborativa mediada por computador, entre alunos com ou sem deficiência visual.

Começamos com um breve apanhado sobre a escrita de matemática Braille, mostrando algumas vantagens e dificuldades deste método. Em seguida mostramos algumas formas de produzir mecanicamente (usando impressora Braille) um texto simples em Braille matemático, usando o Braille como forma de apresentação, constatando que mesmo esta automatização não tem conseguido contornar as dificuldades essenciais que esta técnica apresenta e está sujeita.

Mostramos em seguida as dificuldades existentes atualmente para que as formas usuais de escrita matemática convencionais no computador (editores de equações e as convenções de escrita L_AT^EX) sejam utilizadas por pessoas cegas. Alternativamente, descrevemos a técnica de escrita baseada numa codificação especial acessível (AsciiMath), a leitura em síntese de voz com uma prosódia específica (SonoraMat).

Finalmente mostramos algumas questões relacionadas com a elaboração e leitura de gráficos pelos alunos, apresentando o Geoplano e o Multiplano como elementos importantes no aprendizado. Para operacionalização usando o computador, apresentamos uma ferramenta computacional criada especialmente para atender às demandas do desenho sem feedback visual (Grafivox).

Com o uso destes três mecanismos: o AsciiMath, SonoraMat e Grafivox, pretende-se que o cursista adquira habilidades para:

- Dominar as técnicas de escrita matemática no computador, com ferramentas simples, adotando uma escrita linear (semelhante às expressões que os desenvolvedores de software inserem nos programas de computador, utilizando um simples editor de textos).
- Ler sonoramente um texto matemático, escrito por qualquer pessoa (inclusive pelo próprio aluno). Este texto é reproduzido por fala sintetizada no computador, de forma inteligível e fluente.
- Produzir gráficos matemáticos simples, de forma interativa, que podem ser reproduzidos em impressoras de alto-relevo, que serão facilmente "lidos" por alunos cegos, ou em forma impressa convencional, pelos alunos que enxergam.

Em síntese, este capítulo aborda, de maneira estruturada, alguns conceitos teóricos relacionados ao ensino de matemática para cegos, com a apresentação de técnicas computacionais ainda pouco conhecidas, que facilitam esta atividade, e que constituem o núcleo deste minicurso. A seção 4.4 apresenta um breve referencial teórico sobre as técnicas tradicionais de representação e manipulação de matemática por estudantes cegos, em particular, o Braille Matemático. A seção 4.5 apresenta brevemente dois estilos de uso do computador, mostrando que, em ambos, são grandes as dificuldades ao usar as técnicas de escrita de matemática que são comumente adotadas para usuários

sem cegueira. A seção 4.6 aborda o AsciiMath, uma técnica simples de linearização da escrita matemática e o SonoraMath, complemento computacional que reproduz o texto escrito segundo regras prosódicas peculiares. A seção 4.7 apresenta o Geoplano e o Multiplano, ferramentas importantes para o domínio dos conceitos geométricos básicos. A seção 4.8 mostra uma alternativa para que os cegos sejam capazes de criar interativamente gráficos matemáticos simples no computador, o Grafivox. A Seção 4.9 contém uma breve discussão sobre os resultados alcançados com estas técnicas.

4.4 – Representação e manipulação de matemática por pessoas cegas através do Braille

Em 1824 Louis Braille, um estudante cego construiu um sistema eficaz de escrita e leitura autônomas para pessoas com deficiência visual. Com o advento da escrita Braille os cegos se tornaram proprietários de um competente sistema simbólico, e encontraram a ferramenta fundamental que lhes proporcionou uma revolução semiótica. Tal revolução aumentou significativamente a gama dos fenômenos, corpos e objetos que puderam então ser absorvidos para serem compartilhados com as pessoas que enxergam. (Souza, 2017).

O Braille é um sistema de transcrição que pode ser lido por toque. Nele, os caracteres são representados por conjuntos de seis pontos, numa matriz de 3 linhas e 2 colunas, que são conhecidos como células (também chamadas de *celas*, uma corruptela do inglês "*cell*"). Com 6 pontos é possível representar 63 arranjos, sem contar com o espaço, o que é suficiente para o alfabeto e muitos outros caracteres.

Para identificar os pontos dentro da célula usam-se números de 1 a 6 como na figura 4.2:

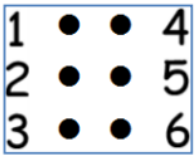
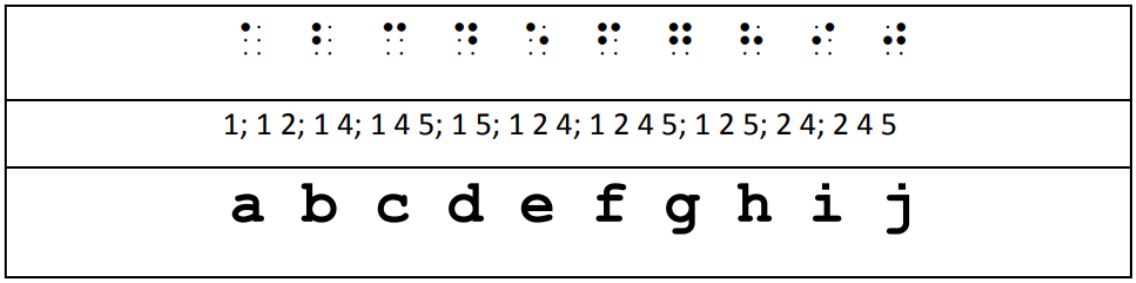


Figura 4.2 – pontos Braille

4.4.1 Representação de caracteres em textos

Louis Braille criou um código fácil de decorar, como mostrado a seguir:

a) Ele usou os 4 pontos superiores para as letras de a até j, eliminando algumas combinações que seriam difíceis de identificar, sendo cego. Por exemplo, ele usou apenas o pontinho 1 para representar a letra a, mas deixou de lado a possibilidade de usar só o pontinho 2 ou só o 4 ou só o 5, pois o leitor se confundiria. Veja as escolhas de Braille:



b) Todas as letras até agora não usaram a linha inferior. Então Braille usou a mesma sequência de pontos agregado ao ponto 3 para as próximas 10 letras. Em seguida agregou os pontos 3 e 6 para as letras restantes.

k l m n o p q r s t

u v x z

Nota: o w é exceção: ⠺ pois não existia em francês naquela época.

c) Indicar maiúsculas também é simples: basta usar os pontos 4 e 6 antes da palavra, por exemplo : ⠠⠁⠃⠁⠙⠑ para representar a palavra Abade.

Duas vezes este símbolo indica uma palavra em caixa alta: ⠠⠠⠁⠃⠁⠙⠑ para representar a palavra ABADE.

d) Acentos: Bem... Acentos não seguem uma regra tão simples: as letras mais os acentos são representados por códigos especiais. Pior que isso, dependendo da língua, são diferentes. Veja na tabela 1 abaixo a relação completa de letras, inclusive, os acentos.

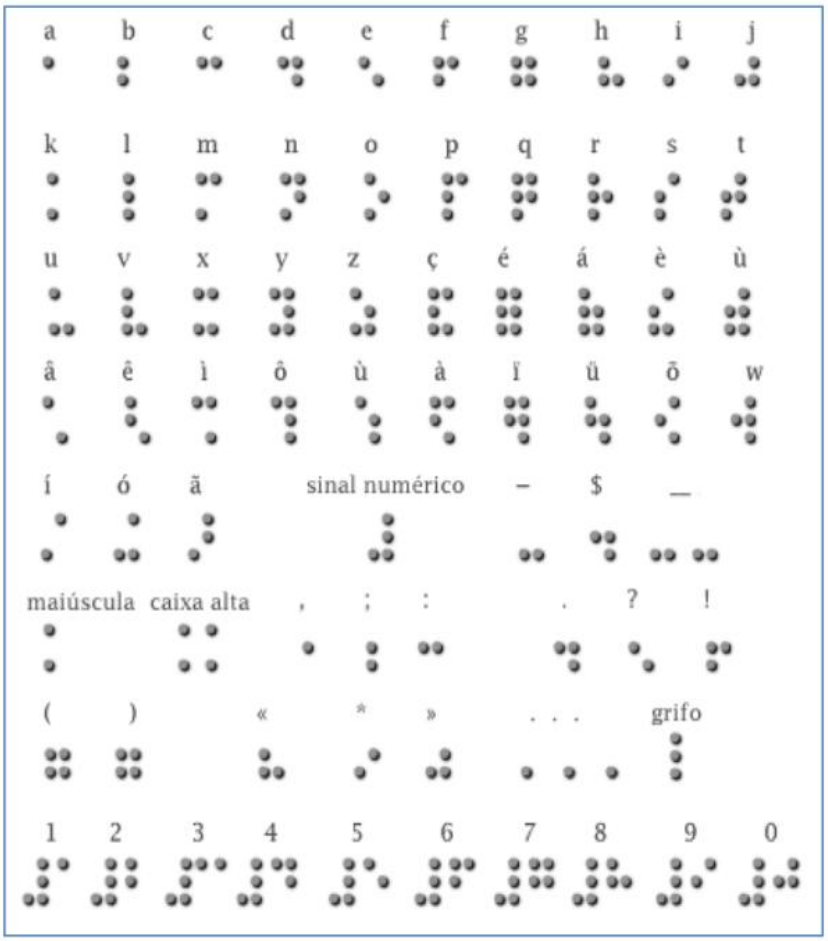


Tabela 1 – códigos Braille usados no Brasil

A escrita Braille pode ser feita manualmente utilizando um dispositivo chamado Reglete, no qual se pressiona um Punção, como visto na figura 4.3. Neste dispositivo, se cria os pontos da direita para a esquerda, de forma que, ao virar o papel, eles estejam com os pontos em relevo para a direita. Opcionalmente pode-se utilizar uma máquina de escrever especial, como mostrado na figura 4.4.

A prática da escrita manual, entretanto, vai além dos objetivos deste curso.

Figura 4.3 – Reglete e Punção

Figura 4.4 Máquina de escrever em Braille

4.4.3 Representação de matemática

A representação de símbolos matemáticos é simples.

a) Para representar os dígitos numéricos, usam-se as letras de a até j em Braille, sendo o número precedido por um sinal # (pontos 3 4 5 6).

   123 = #abc = ⠼⠁⠃⠉

b) O ponto e a vírgula decimais são hoje representados, respectivamente pelos pontos 3 e 2.

   123.000,00 = #abc,jjj,00 = ⠼⠁⠃⠉⠲⠚⠚⠚⠂⠚⠚

c) Símbolos aritméticos

   \+ ⠬ - ⠤ × ⠡ ÷ ⠌ = ⠨⠅ (⠣) ⠜

d) Representação de expressões

   Esta mesma técnica, com pequenas adaptações, permite a transcrição de expressões matemáticas que precisam ser adaptadas com um processo algorítmico que inclui sua linearização (que pode incluir parênteses), como está ilustrado na **Figura 4.5.**

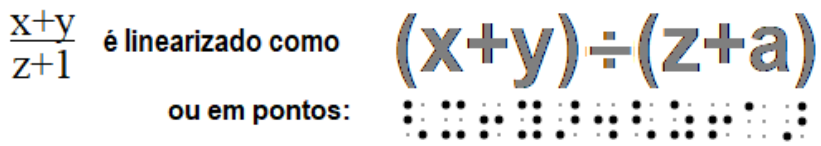


Figura 4.5: Expressão matemática simples e sua conversão para Braille.

e) Frações numéricas: abaixa-se os pontos do número do dividendo

2/3 ⠼⠃⠌⠼⠉

Detalhe técnico:

Ao linearizar expressões é comum ter que adicionar parênteses que não existiam na expressão original. Neste caso é conveniente usar nestes parênteses uma codificação especial, com os pontos

(⠐) ⠐

IMPORTANTE: O BRAILLE MATEMÁTICO NÃO É UNIVERSAL

Todos os símbolos matemáticos têm equivalência em Braille, mas devemos notar que os códigos de Braille matemático usados não são universais. No Brasil, por exemplo, adotou-se um código unificado com a Espanha e Portugal (Anjos, 2016), o que trouxe sérias consequências, como a dificuldade em usar a ampla literatura em Braille americano (Nemeth Code) (NBA, 1972) e a impossibilidade de usar os vários programas para transcrição computadorizada para matemática em Braille.

4.4.4 Um pequeno exercício usando a folhinha de treinamento

Disponibilizamos uma folha na Internet, em que os pontos Braille foram marcados como pequenas bolinhas. Desta forma é fácil treinar usando lápis e borracha.

Baixe e imprima esta folhinha a partir da página do link:

http://intervox.nce.ufrj.br/~antonio2/cursobraille/folhinha_braille.pdf

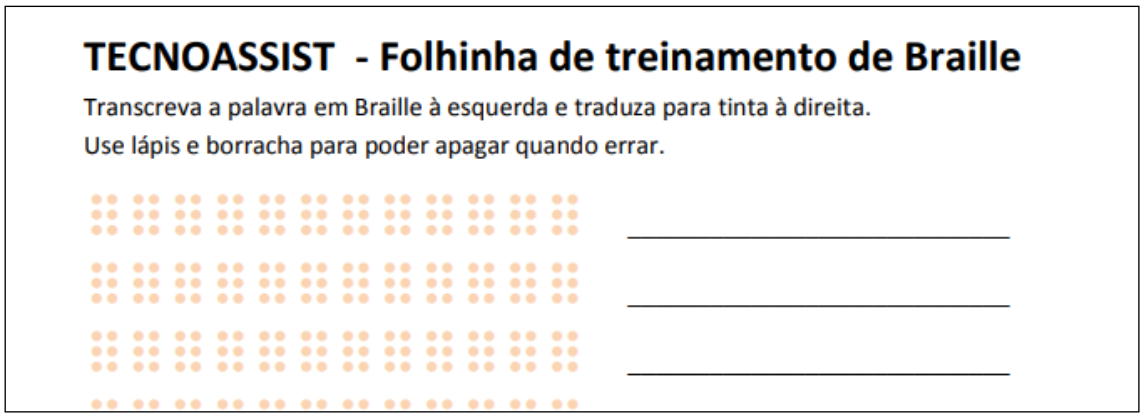


Figura 4.6: Folhinha de treinamento Braille

Produza na folhinha de treinamento usando um lápis para marcar os pontinhos Braille:

1+2=3

(2+4) × 3 = 12

5 ÷ 2 = 2,5

2/3

1/3 + 1/2 = 5/3

4.4.5 Impressão computadorizada de Braille

No Brasil, houve ampla disponibilização de impressoras computadorizadas de Braille promovidas pelo MEC como parte do Plano Nacional do Livro Didático em Braille (Borges, 2001). Foram criadas também ferramentas tecnológicas automáticas para transcrição com rapidez e boa usabilidade, como o Braille Fácil. (ver figura 4.7).

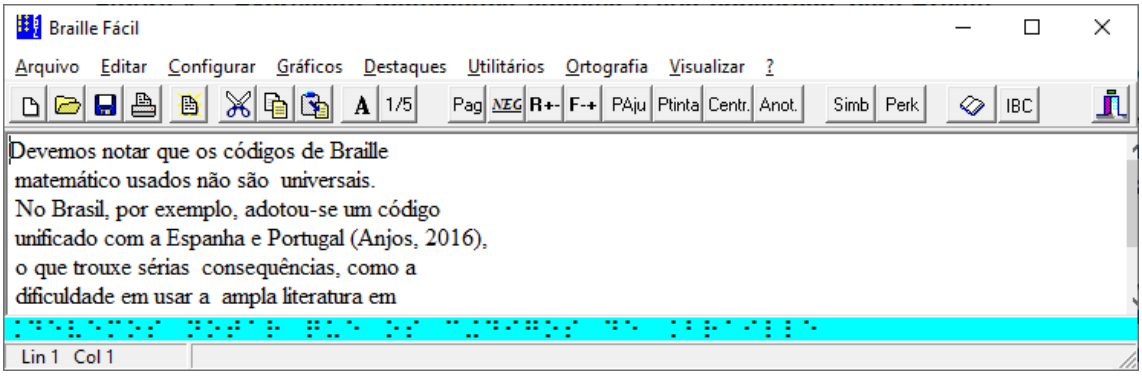


Figura 4.7 – Programa Braille Fácil

O uso deste programa é quase trivial. Simplesmente se digita o texto na tela e seleciona-se a opção de Arquivo/Imprimir em Braille. Naturalmente, é necessário que ao computador esteja acoplada uma impressora Braille. O uso do programa Braille Fácil transcende os objetivos deste curso.

Um detalhe: alguns símbolos matemáticos na forma textual diferem dos símbolos usuais de texto simples. Para evitar discrepâncias na tradução para Braille, o programa permite que todos os símbolos matemáticos sejam desambiguados simplesmente precedendo-os pelo sinal de crase (`).

4.4.6 O declínio da escrita Braille

A disponibilidade das ferramentas tecnológicas ajudou a incrementar a produção de textos em Braille, mas não impediu o declínio da utilização desta técnica no país. Este é um fenômeno mundial, provocado pelo uso cada vez mais intenso do computador e outros dispositivos similares associado a ferramentas de acessibilidade. Por exemplo, nos Estados Unidos, estimava-se que na década de 1950, 40 por cento das pessoas cegas tinha contato com o Braille. Em 2009, menos de 10 por cento das crianças cegas o aprendia [NFB, 2009].

No Brasil, o ensino da matemática para cegos, até poucos anos atrás, vinha utilizando como suporte a escrita Braille, que embora seja intrinsecamente unidimensional, é capaz de prover mecanismos de manipulação de textos não lineares. Mas há uma grande dificuldade: além do declínio do conhecimento de Braille pelos alunos, também é raro que os professores das classes convencionais estejam habilitados para esta forma de escrita e leitura.

Desta forma, nós recomendamos a escrita matemática em Braille na escola nos primeiros anos, em particular no ensino da matemática básica, pois essa forma de escrita tem a tendência de gerar um contato mais íntimo com a matemática e um aumento na abstração, na medida em que existe um consenso de que o processo de escrita Braille favorece o contato íntimo com os símbolos o que se traduz numa relação mais íntima com a matemática, fundamental no ensino básico. Entretanto, quando se trata de um ensino mais avançado, que exige interação com o professor e com os colegas num nível mais complexo, e usando textos de certo tamanho, o uso do computador é bem mais eficiente.

4.5 – A escrita tradicional de matemática no computador e seus elementos de inacessibilidade para cegos

Existe uma preponderância de editores de textos para literatura, mas um número muito reduzido de ferramentas para escrita de textos matemáticos. Isso vale também quando se trata de pessoas com deficiência visual.

A maior parte das pessoas que enxergam faz uso de duas formas de escrever matemática:

a) usando uma codificação textual simples com marcações (conhecida como L_AT^EX)

b) montando as fórmulas matemáticas com suporte de um editor gráfico, usando o mouse, com o qual se vai posicionando interativamente os símbolos num painel na tela do computador. Um exemplo é o *Equation Editor* do Microsoft Word.

A escrita em L_AT^EX é conceitualmente simples: utiliza-se um editor comum de textos, usando os caracteres + - * / = para os símbolos aritméticos usuais. O restante dos símbolos matemáticos e as indicações de posicionamento gráfico são representados pela \ (barra invertida) seguida por uma palavra-chave. A figura a seguir ilustra o texto digitado e sua representação quando o texto for impresso por um utilitário de conversão.

```
\[
      \left(
            5a + \frac{3ab^{2}}{2a^{2}-{b\over 2}}
      \right) *(a^{3}+b^{3})
\]
```

que representa

$$\left(5a + \frac{3ab^2}{2a^2 - \frac{b}{2}}\right) * (a^3 + b^3)$$

Figura 4.8: Representação L_AT^EX de uma expressão matemática

Frisamos que a escrita textual simples do L_AT^EX pode ser executada sem dificuldade usando qualquer editor de textos simples, em particular o Sistema Dosvox. Entretanto, o texto escrito apresenta uma forma de escrita muito rebuscada, tornando sua leitura árdua e sujeita a interpretações equivocadas quando realizada através de um sintetizador

de voz. Já a leitura da produção de editores gráficos de matemática não é suportada pelo Dosvox.

A alternativa são os softwares leitores de tela, que são capazes de reproduzir o conteúdo que está escrito na tela, dando também suporte interatividade neste conteúdo, possivelmente usando o mouse ou o teclado. Porém, estes softwares também dão suporte muito limitado à leitura da tela quando ela apresenta textos matemáticos e quando se utilizam editores gráficos para matemática.

Como alternativa, Silveira et al. (2011) apresentaram um artefato tecnológico capaz de elaborar material instrucional com símbolos matemáticos, sendo estes convertidos automaticamente para o formato texto, que pode ser reproduzido pelos leitores de tela. Além disso, essa ferramenta pode gerar o mesmo conteúdo no formato MathML (formato textual, semelhante a HTML, usado pelos navegadores de internet para apresentar visualmente textos matemáticos). No entanto, o estudo realizado sobre este artefato mostra algumas limitações como a navegação em fórmulas extensas e a incorporação desses componentes em um ambiente mais versátil para editoração do que os navegadores Web.

4.6 – Alternativas de escrita e leitura de matemática por cegos: o AsciiMath e o SonoraMat

Simplificando, existem dois problemas mais importantes para o aluno cego quando o assunto é matemática:

1. Escrever matemática no computador, com ferramentas simples, usando uma escrita linear (algo semelhante ao que os programadores fazem quando criam programas convencionais em um editor de textos de linhas de comando).

2. Ler um texto matemático escrito por ele próprio ou por outras pessoas, traduzindo o texto criado para uma fala que o representasse de forma clara.

O Dosvox era, até pouco tempo atrás, frágil quando o assunto era suporte à matemática. Tinha algumas ferramentas específicas, como calculadora sonora e planilha especializada, mas elas não davam suporte razoável à escrita de matemática e à operacionalização de cálculos matemáticos mais sofisticados.

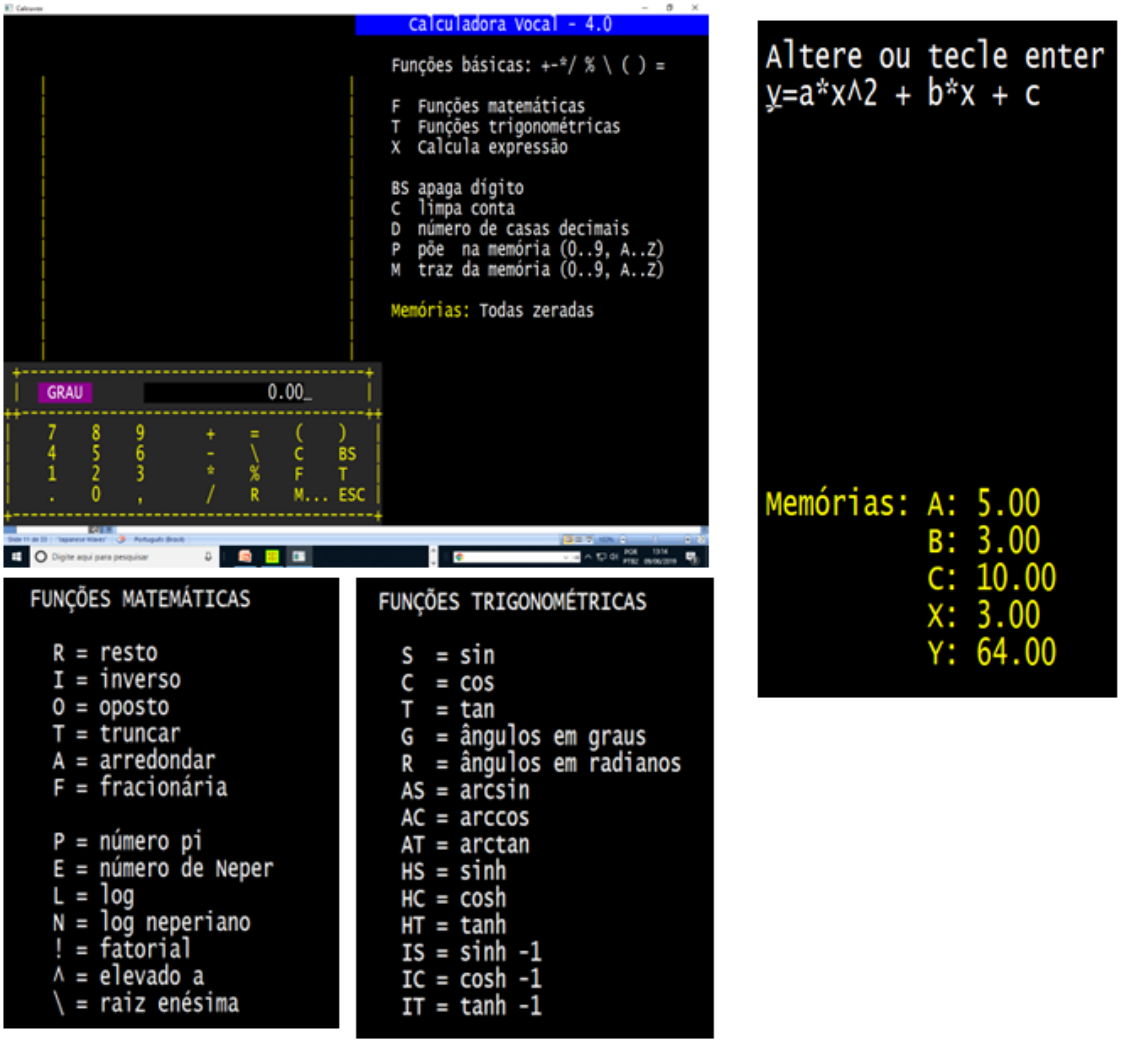


Figura 4.9 – Calculadora sonora do Dosvox

Nota: Uma importante abordagem foi criada e usada com sucesso por diversos alunos de nível médio e superior, tendo por base o estilo de interface do Dosvox: o MatVox (Silveira, 2010),que agrega facilidades de cálculo matemático a um editor de textos. Porém, o problema de leitura de matemática convencional, também não foi resolvido por esta abordagem.

Um breve exercício:

Ative a calculadora sonora do Dosvox, usando as letras UC. Coloque na memória X da calculadora o valor 5 e depois calcule o valor da expressão: $x^2 + 2x + 3$

4.6.1 Como um cego criaria um texto contendo expressões matemáticas no computador?

As alternativas mais óbvias para uma pessoa cega criar matemática, portanto seriam os formatos LaTeX, usados por 90% dos matemáticos para gerar textos científicos, ou o MathML (que não está sendo apresentado aqui), usado pelos navegadores da web para mostrar textos matemáticos. Infelizmente, ambos são complicados de escrever e sua leitura usando síntese de voz, é quase ininteligível.

Uma terceira alternativa foi escolhida: o formato AsciiMath (Gray, 2007), uma forma fácil de escrever matemática para a qual a *renderização* (geração do gráfico em papel) é compatível com todos os navegadores da atualidade. Este formato é similar à escrita de fórmulas matemáticas de linguagens de computação como Fortran ou Python, ensinadas nos cursos STEM.

Os principais símbolos de AsciiMath são os seguintes:

+ - * / = ( ) para os símbolos matemáticos básicos

sqrt para raiz quadrada

sum para somatório

^ para representar a situação em que um símbolo sobe (por exemplo, num expoente)

_ para representar quando um símbolo desce.

Importante: o gerador de renderização gráfica elimina os parênteses que são colocados apenas para agrupamento, por exemplo 2^(x+y) seria desenhado por 2^{x+y}

Isso é verdade também quando representamos frações.

Por exemplo, uma equação do segundo grau: x^2 + bx + c = 0

A figura 4.10 mostra um exemplo mais completo.

```
sum_(i=1)^n i^3=((n(n+1))/2)^2
```

$$\sum_{i=1}^{n} i^3 = \left(\frac{n(n+1)}{2}\right)^2$$

Figura 4.10: Fórmula digitada em AsciiMath e renderizada

Essa solução foi integrada ao Sistema Dosvox, que, a partir de 2018, passou a editar e imprimir fórmulas matemáticas de grande complexidade, misturadas a textos comuns, bastando para isso que os textos em AsciiMath fossem precedidos e sucedidos por um caractere especial (acento grave), de forma que seja fácil distinguir o que é texto e o que é matemática. Falaremos em seguida sobre a sua impressão.

Símbolos especiais de AsciiMath

AsciiMath suporta praticamente todos os símbolos matemáticos usados no ensino superior. A tabela 2 a seguir apresenta alguns dos mais utilizados, inclusive a sua equivalente em LATEX

Type	TeX alt	See
+		$+$
-		$-$
*	cdot	$\cdot$
**	ast	$\ast$
***	star	$\star$
//		$/$
\\	backslash setminus	$\backslash$
xx	times	$\times$
-:	div	$\div$
|><	ltimes	$\ltimes$
>|<	rtimes	$\rtimes$
|><|	bowtie	$\bowtie$
@	circ	$\circ$
o+	oplus	$\oplus$
ox	otimes	$\otimes$
o.	odot	$\odot$
sum		$\sum$
prod		$\prod$
^^	wedge	$\wedge$
^^^	bidwedge	$\bigwedge$
vv	vee	$\vee$
vvv	bigvee	$\bigvee$
nn	cap	$\cap$

Type	TeX alt	See
2/3	frac{2}{3}	$\frac{2}{3}$
2^3		2^3
sqrt x		$\sqrt{x}$
root(3)(x)		$\sqrt[3]{x}$
int		$\int$
oint		$\oint$
del	partial	∂
grad	nabla	∇
+-	pm	$\pm$
O/	emptyset	$\emptyset$
oo	infty	∞
aleph		$\aleph$
:.	therefore	$\therefore$
:'	because	$\because$
|...|	|ldots|	$\|\ldots\|$
|cdots|		$\|\cdots\|$
vdots		$\vdots$
ddots		$\ddots$
|\ |		$\|\ \|$
|quad|		$\|\ \|$
/_	angle	$\angle$

Type	TeX alt	See
=		$=$
!=	ne	$\neq$
<	lt	$<$
>	gt	$>$
<=	le	$\leq$
>=	ge	$\geq$
-<	prec	$\prec$
-<=	preceq	$\preceq$
>-	succ	$\succ$
>-=	succeq	$\succeq$
in		$\in$
!in	notin	$\notin$
sub	subset	$\subset$
sup	supset	$\supset$
sube	subseteq	$\subseteq$
supe	supseteq	$\supseteq$
-=	equiv	$\equiv$
~=	cong	$\cong$
~~	approx	$\approx$
prop	propto	$\propto$

Tabela 4.2 – Alguns símbolos de AsciiMath

Nota: A relação completa de símbolos pode ser encontrada em www.asciimath.org

Renderizando AsciiMath

Existem alguns programas de uso público na internet capazes de renderizar textos em AsciiMath, mas nós optamos por criar um editor bem simples, que pudesse ser utilizado por alunos cegos e também por seus professores e colegas que enxergam para gerar textos matemáticos com grande beleza gráfica. Este programa pode também ser utilizado acoplado a leitores de tela.

Com o Intermat, uma pessoa qualquer consegue gerar textos matemáticos com grande beleza quando impresso, e que apresenta acessibilidade para deficientes visuais. O programa também aceita códigos HTML, que são simples e muito conhecidos, para diminuir a curva de aprendizado dos estudantes, como mostrado na figura abaixo.

Nota: Repare a inclusão de marcações para destaque gráfico em HTML (h2 e h3) e o uso de crases para indicar os trechos contendo códigos de matemática.

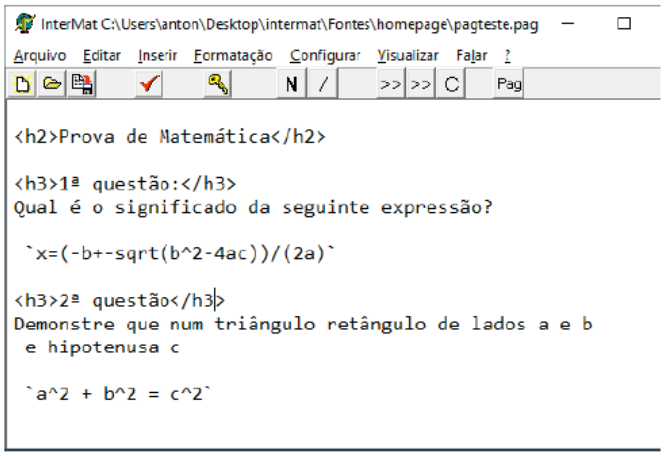


Figura 4.11: Um trecho de prova criado no InterMat (ou no Dosvox)

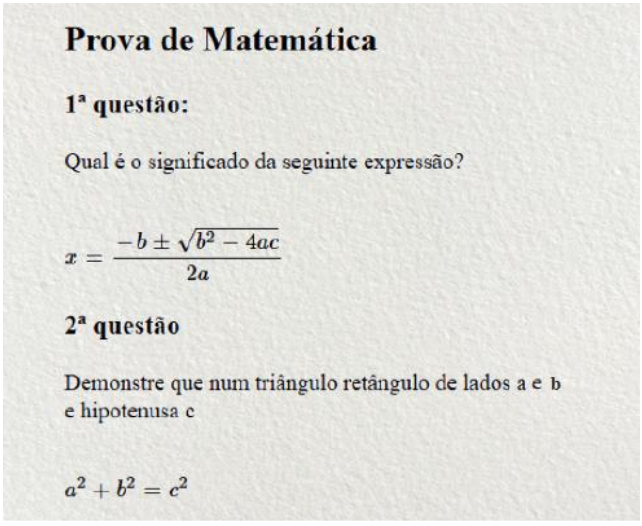


Figura 4.12: Impressão na tela ou papel através do InterMat

Um breve exercício:

Instale o programa InterMat no seu computador, digite e visualize a as seguintes expressões matemáticas:

$ax^2 + bx + c = 0$

$a_1x^2 + a_2x + c = 0$

$$x' = \frac{-b + \sqrt{\Delta}}{2.a}$$

$$x'' = \frac{-b - \sqrt{\Delta}}{2.a}$$

Digite esta mesma última fórmula expandindo Delta para b^2-4ac

Lendo AsciiMath num sintetizador de voz

Apesar da simplicidade e conveniência deste formato, Ainda que um cego, devidamente orientado, possa escrever matemática com grande precisão e beleza com o AsciiMath, ele não conseguiria ler uma fórmula razoavelmente complexa de forma fluente e compreensível. Para isso foi desenvolvido no NCE/UFRJ o SonoraMat, uma ferramenta de leitura e elaboração de textos matemáticos. O SonoraMat é um programa acessório que quando executado, interage com Dosvox, Intermat e outros programas para interpretar a fala das expressões matemáticas.

Operação:

1. execute o programa SonoraMat
2. execute o Edivox ou SonoraMat.
3. Aperte as letras ALT-H para conectar os programas ao SonoraMat. Esta operação é feita apenas uma vez antes da primeira leitura do texto.
4. digite as expressões matemáticas precedendo-as e sucedendo-as com o sinal de crase.
5. Use os procedimentos normais para leitura, por exemplo, usando as setas.
6. Ao final, feche o programa SonoraMat

Nota:

Todos os textos matemáticos serão interpretados automaticamente no Dosvox ao usar as setas. No Intermat, que não é um programa tipicamente sonoro, é preciso ativar a configuração de fala automática das linhas, e neste caso as expressões serão faladas junto com as teclas Cima e Baixo deste editor.

Um breve exercício:

Leia em síntese de voz os resultados do exercício anterior, usando o Intramat ou o Editor Edivox do sistema Dosvox.

Para isso, instale antes o Sonoramat no computador a partir do site:

http://intervox.nce.ufrj.br/sonoramat

4.7 O Geoplano e o Multiplano

Antes de ensinar uma pessoa cega a desenhar é importante passar para ela alguns conceitos básicos de geometria, sendo o Geoplano uma ótima ferramenta. Este objeto é formado por uma placa de madeira onde são cravados pregos, fomando uma malha composta por linhas e colunas dispostas de acordo com a figura a seguir:

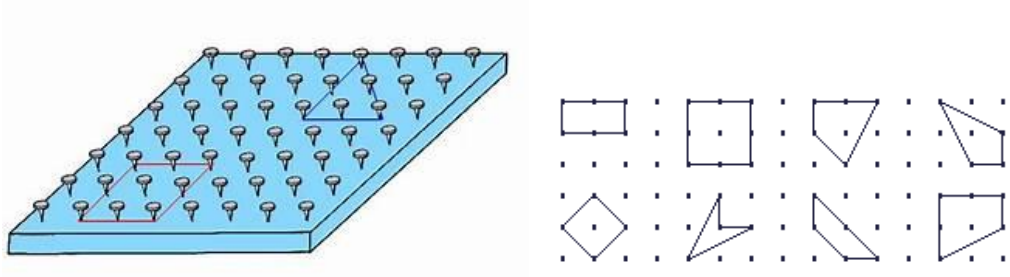


Figura 4.13 – Geoplano e alguns desenvolvimentos

O Geoplano introduz a pessoa cega à Geometria Euclidiana, em particular ao Plano Cartesiano. Ele pode ser utilizado para construir figuras simples, unindo os preguinhos por elásticos, como na figura a seguir. O Geoplano também é muito útil para desenvolver estratégias para cálculo de perímetro, área, figuras simétricas, arestas, vértices, construção de polígonos, exploração espacial entre outros.

Exercício:

Use um geoplano para:

- Construir dois quadrados de áreas diferentes.
- Calcular a área e o perímetro de cada figura.
- Dividir os quadrados em triângulos de mesma área.
- Descobrir a área de cada triângulo encontrado.
- Construir um retângulo de área igual a 12

Provavelmente você não terá acesso a um geoplano real (embora seja muito fácil adquirir um através da Internet). Mas você pode usar um geoplano virtual:

https://apps.mathlearningcenter.org/geoboard/

Devemos finalmente, indicar que há iniciativas importantes que se apresentam como alternativas ao Geoplano. Destacamos o Multiplano, dispositivo inventado por Rubens Ferronato em sua tese de mestrado, e posteriormente industrializado. [Ferronato, 2002].

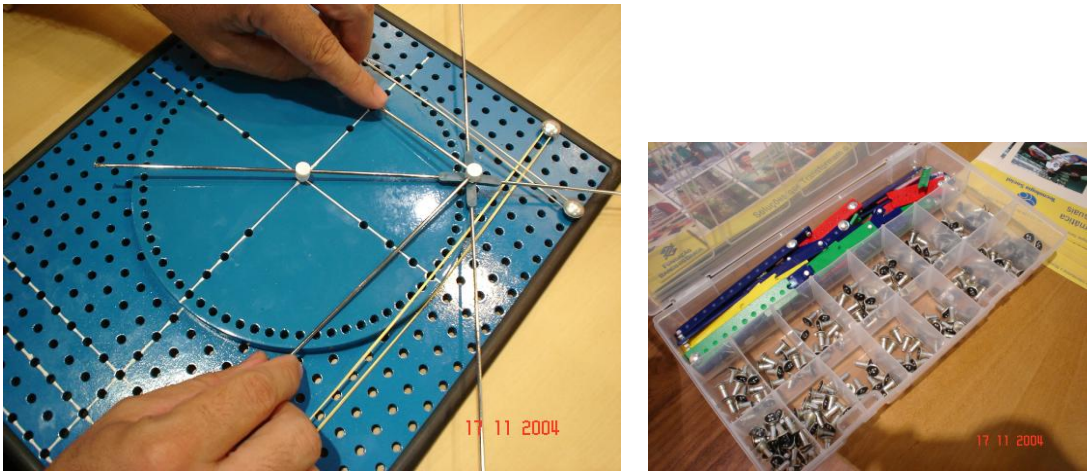


Figura 4.14 – O multiplano e algumas de suas peças

O Multiplano é base de uma metodologia cujos resultados são apontados como muito efetivos em vários níveis de ensino (da pré-escola à universidade). Entretanto o estudo desta ferramenta transcende os objetivos deste texto, especialmente pelo nosso foco no suporte computacional ao ensino de matemática para cegos.

4.8 Criação interativa de gráficos por pessoas cegas

Em matemática, o desenho é crucial para complementar e compreender a informação escrita. Quanto mais técnica ou complexa é a expressão matemática, mais a sua representação gráfica se firma como uma linguagem na qual podemos expressar ideias e conceitos de maneira concisa, clara e interessante. Até pouco tempo atrás, porém, as tecnologias disponíveis de desenho para cegos permitiam apenas a leitura ou a escrita indireta ou com interveniência de outros dispositivos. A verdade é que os cegos não têm sido estimulados a desenhar os gráficos, apenas consumi-los.

Estudos demonstram que a partir de experiências com alunos cegos (Borges, 1998), pode-se produzir, com treinamento mínimo, gráficos legíveis – para pessoas cegas ou não. Isto, entretanto, só se viabilizou pela disponibilidade das impressoras Braille e das máquinas fusoras, hoje presentes em centros de apoio pedagógico nas escolas e universidades públicas.

Com a entrada de estudantes cegos em carreiras STEM, na UFRJ, tornou-se urgente viabilizar que os alunos conseguissem produzir gráficos simples para apresentar em seus trabalhos universitários. Para isso, foi agregado ao sistema Dosvox um utilitário que, através de uma pequena linguagem gráfica pudesse produzir gráficos simples, compostos por elementos básicos (linhas, curvas, funções, eixos, etc). O programa Grafivox é capaz de reproduzir, na tela ou numa impressora em tinta ou Braille, o desenho produzido, automaticamente escalonado e adaptado às características daqueles equipamentos (que no caso das impressora Braille são bastante distintos, pela baixíssima resolução que apresentam). É possível também gravar a forma gráfica gerada em PNG ou JPG, para importação num editor de textos ou outro utilitário qualquer.

A linguagem gráfica que foi desenvolvida, também denominada de Grafivox, foi tornada também compatível com o sistema Intermat, possibilitando desta forma que os textos matemáticos criados ou exibidos por aquele utilitário pudessem ter ilustrações gráficas.

A interação do programa é trivial, e como a maior parte dos programas do Dosvox, baseada num menu, controlado pelas setas, para escolha da função desejada: edição (interativa ou por edição direta), visualização, impressão e configuração, como mostrado na figura 4.15.

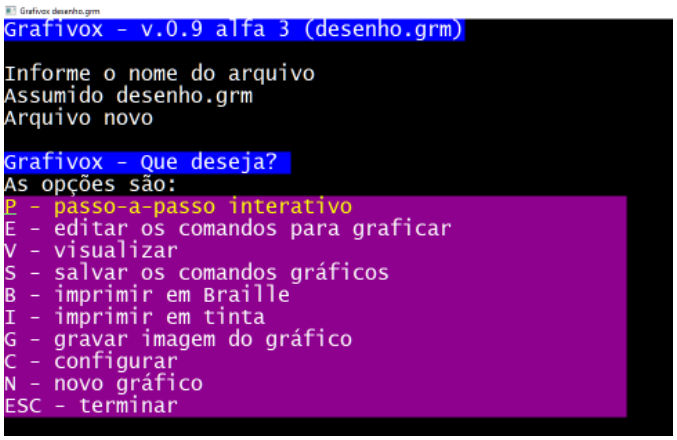


Figura 4.15: Opções gerais do programa Grafivox

4.8.1 A linguagem Grafivox

A figura 4.16 mostra à esquerda um pequeno programa escrito no Grafivox, e à direita, o resultado impresso em tinta e em Braille.

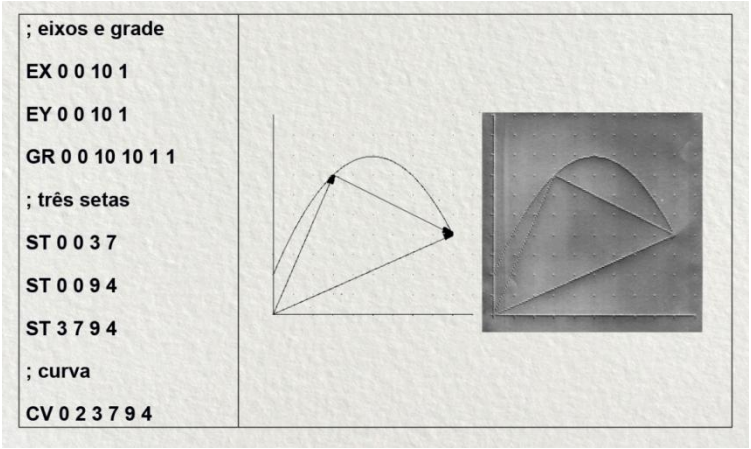


Figura 4.16: Pequeno gráfico gerado no Grafivox, impresso automaticamente em tinta e em impressora Braille.

Como se pode observar pela figura 4.16, os comandos dessa linguagem são expressos por abreviaturas de duas letras (p.ex.: RT para reta, RG para retângulo), seguidas por coordenadas (cartesianas, polares, relativas, fórmulas ou pontos descritos a partir de arquivos) que descrevem a forma indicada pelo comando. A lista de possibilidades está mostrada na figura 4.17.

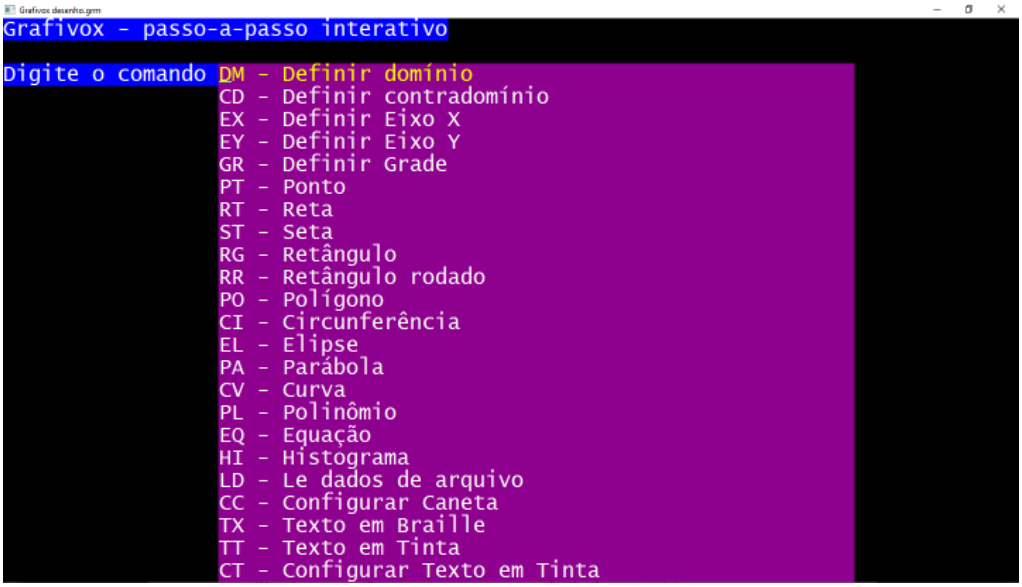


Figura 4.17: Comandos da linguagem Grafivox

O programa permite duas formas que podem ser usadas de forma misturada para criar o programa. A primeira, voltada para iniciantes, é chamada de edição interativa, em que o usuário escolhe a função com as setas (ou teclando a abreviatura), e a partir daí, se abre um formulário para preenchimento das opções, como visto na figura 4.18. Os comandos vão sendo acumulados na ordem em que são inseridos.

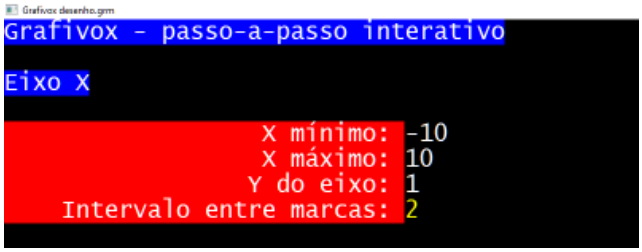


Figura 4.18: Especificação interativa dos parâmetros do comando EX (eixo X)

A segunda forma de operação é simplesmente um editor de linhas, em que o usuário tecla os comandos já com os parâmetros. Parece difícil ter que decorar tantas possibilidades de parâmetros, mas num desenho típico a variedade de comandos usados é pequena, e a ordem dos parâmetros intuitiva. Em outras palavras, nossa observação mostra que depois de poucas horas, os usuários quase não recorrem mais à opção interativa (exceto para sanar alguma dúvida).

Exercício:

Crie na linguagem Grafivox o desenho

- de dois quadrados concêntricos, de lado 2 e 4.
- dentro do quadrado mais interno trace dois segmentos de reta unindo os vértices não adjacentes.
- crie uma circunferência que toque nos quatro vértices do quadrado externo.

Visualize e imprima o gráfico criado na linguagem Grafivox no programa Intermat ou diretamente através do programa Grafivox do Dosvox.

4.9 Uma breve discussão sobre a metodologia utilizada

Neste texto foi apresentada a problemática de ensino e transcrição de matemática para cegos, tanto no que se refere a textos quanto gráficos. Foram apresentadas algumas alternativas com uso intensivo do computador, que podem ser aplicadas em diversas modalidades, tanto presencial quanto à distância, com ou sem estratégias de colaboração.

A base metodológica se centra no uso de um formato específico de escrita (AsciiMath), no uso de uma ferramenta de leitura de textos matemáticos (SonoraMat) e num sistema integrado interativo, totalmente acessível, que utiliza uma linguagem gráfica simples (Grafivox) Os programas apresentados são uma base mínima para o desenvolvimento acadêmico, e certamente precisam ser atualizados incessantemente ou substituídos por outras soluções, visando atender aos requisitos crescentes de inclusão acadêmica nos cursos que tem como base a matemática nas universidades.

É esperado que a aplicação das ideias mostradas neste texto em cursos preparatórios para professores de nível médio e universitários possibilite a capacitação de uma massa crítica, habilitada a utilizar e ensinar matemática com suporte tecnológico, com resultados muito mais promissores, como os que temos observado na UFRJ. De forma ainda mais conveniente, todos estarão usando uma tecnologia brasileira, simples, de distribuição gratuita e compartilhada.

Temos certeza de que, com o uso amplo destas ideias, o número de alunos que poderão ser beneficiados alcançará a marca de muitos milhares de estudantes, justificando plenamente sua aplicação no cenário brasileiro, tão carente deste tipo de ferramentas.

Agradecimentos:

Laboratório de Aplicações e Pesquisas em Tecnologia Assistiva. Projeto com a Chancela SBC.

Referências:

Anjos, D. Z. - Código matemático unificado: da definição às diferenças semióticas na conversão da tinta ao Braille - Encontro Nacional de Educação Matemática – Educação Matemática na Contemporaneidade: desafios e Possibilidades – 2016, disponível em http://www.sbem.com.br/enem2016/anais/pdf/5413_2991_ID.pdf

Borges, J.A. Do Braille ao Dosvox – diferenças nas vidas dos cegos brasileiros – Rio de Janeiro: UFRJ/COPPE, 2009.

Borges, J. A. e Chagas Jr, G. J. F., Impressão Braille no Brasil: o papel do Braivox, Braille Fácil e Pintor Braille. Anais do I Simpósio Brasileiro sobre Sistema Braille, 2001.

Borges, J. A.; Borges, P. P., Matemática para alunos cegos. CIÊNCIA HOJE, v. 348, p. 1, 2018.

Borges, J. A.; Jensen, L. R. Cegos e Computador: Uma Interação que Explora o Potencial do Desenho. In: IX Simpósio Brasileiro de Informática Educativa, 1998, Recife. Anais do SBIE'98, 1998.

Cryer, H. Teaching STEM subjects to blind and partially sighted students: Literature review and resources. RNIB Centre for Accessible Information, Birmingham - 2013: Literature review #6.

Dias, A. F. S.; Franca, J.B.S.; Borges, M. R. S. Silva. Tecnologia Assistiva: Um Survey com portadores de deficiência visual em ambiente virtual de aprendizagem a partir do Modelo TAM. In: XVIII Conferência Internacional sobre Informática na Educação, TISE 2013, Porto Alegre

Dias, A.F.S; França, J.B.S.; Borges, J.A.S.; Silveira, J.T.C.; Carvalho. M.F.C.; Borges, M.R.S.B. Matemática, Computação e Braille: Desafios da Pedagogia, da Semiótica e da Síntese da Fala. CBIE – Congresso Brasileiro de Informática na Educação. Fortaleza , CE, 2018. https://br-ie.org/pub/index.php/sbie/article/view/8175

Ferreira, T.A.C.S - Sistema Online de Síntese de Fala em Português – Tese de Mestrado - Faculdade de Ciências e Tecnologia do Departamento de Engenharia Eletrotécnica e de Computadores - Universidade de Coimbra - 2014

Ferronato, Rubens. A construção de instrumento de inclusão no ensino da matemática. Dissertação de mestrado, UFSC, Florianópolis - SC. 2002, disponível em https://repositorio.ufsc.br/bitstream/handle/123456789/82939/PEPS2320-D.pdf?sequence=1&isAllowed=y

Gray, James (2007), "ASCIIMathML: now everyone can type MathML", MSOR Connections, 7 (3): 26–30.

Melo, A.M.. Acessibilidade e Inclusão Digital em Contexto Educacional. 3º Congresso Brasileiro de Informática na Educação (CBIE 2014). 3ª Jornada de Atualização em Informática na Educação (JAIE 2014). Dourados, MS.

NCE-UFRJ. (2010) “Projeto Dosvox”. Núcleo de Computação Eletrônica da Universidade Federal do Rio de Janeiro. http://intervox.nce.ufrj.br/dosvox

N.F.B. (National Federation for the Blind) - The Braille Literacy Crisis in America Facing the Truth, Reversing the Trend, Empowering the Blind; Jernigan Institute; March 26, 2009.

Silveira, H.M.; Martini, L.C. MATVOX: um aplicativo para deficientes visuais que proporciona a implementação de algoritmos e cálculos matemáticos em um editor de texto. Brazilian Symposium on Computers in Education - Simpósio Brasileiro de Informática na Educação – SBIE, Campinas – SP. 2011.

Souza, J. B. - O que vê a cegueira - A escrita Braille e sua natureza semiótica. - Ed. UFPB –2017.

Biografia Resumida dos Autores:

Angélica F. S. Dias possui Doutorado em Informática pela Universidade Federal do Rio de Janeiro (PPGI - 2018) com ênfase em Gestão de Sistemas Complexos. Mestre em Sistemas de Informação pela UFRJ. MBA em E-Business pela COPPEAD e DBM - Inteligência e Database Marketing/NCE/UFRJ. Aperfeiçoamento em Gerência Avançada de Projetos/UFRJ. Graduação em Processamento de Dados/UNESA. Foi Diretora da Área de Extensão do Instituto Tércio Pacitti/UFRJ, Coordenadora Acadêmica dos Cursos de pós-graduação no NCE/UFRJ (2003-2005). Professora Convidada pelos: Instituto de Economia, Instituto Tércio Pacitti (NCE) e Programa de Pós-Graduação em Informática (PPGI). Foi Professora convidada do Coppead/UFRJ, Ministério da Educação, IBMEC e Cecierj. Tem experiência nas áreas de Administração Pública, Gerência de Projetos, Ciência da Computação e Educação. Atualmente é pesquisadora e Coordenadora de Extensão pelo Instituto Tércio Pacitti/NCE/UFRJ. Temas de interesse: Gestão de Conhecimento, Trabalho e Aprendizagem Cooperativa apoiada por computador (CSCW e CSCL), Tecnologia Assistiva, Economia Circular com ênfase em Sustentabilidade, Gestão Estratégica da Informação e Educação. Artigos e capítulos de livros publicados em conferências nacionais e internacionais.

Lattes: **http://lattes.cnpq.br/8795875378897586**

José Antonio dos Santos Borges possui graduação em Matemática mod. Informática pela Universidade Federal do Rio de Janeiro (1980), graduação em Piano - Conservatório Brasileiro de Música (1977), mestrado (1988) e doutorado em Engenharia de Sistemas e Computação pela COPPE/UFRJ (2009). Atualmente é Coordenador da Pós-graduação em História das Ciências e das Técnicas e Epistemologia da UFRJ – É analista de Tecnologia da Informação no Instituto Tércio Pacitti da UFRJ (NCE/UFRJ), onde trabalha desde 1975. É especialista em Tecnologia Assistiva, tendo desenvolvido grande

quantidade de sistemas para acesso de pessoas com deficiência aos computadores. Atuou também em síntese de voz, sistemas para cartografia tátil adaptada, computação gráfica e CAD para microeletrônica. Premiado duas vezes com a medalha de Excelência Acadêmica do Instituto de Matemática da UFRJ. Tem diversos artigos e capítulos de livros publicados em conferências nacionais e internacionais.

Lattes: **http://lattes.cnpq.br/1957526921210046**

Júlio Tadeu Carvalho da Silveira possui graduação em Informática (Bacharelado) pela Universidade Federal do Rio de Janeiro (1990) e mestrado em Engenharia de Sistemas e Computação pela COPPE/UFRJ (1996). Atualmente é Técnico de Tecnologia da Informação, no Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais/NCE, da Universidade Federal do Rio de Janeiro e Professor Assistente do Centro Universitário Carioca. Tem experiência na área de Ciência da Computação. Atuando principalmente nos seguintes temas: Robótica, Planejamento de Trajetória, Trajetória Evitando Obstáculos. Tem diversos artigos publicados em conferências nacionais e internacionais.

Lattes: **http://lattes.cnpq.br/7081474226160086**

Capítulo 5

Fusão de dados para Ambientes Inteligentes

Claudio M. de Farias, Gabriel Caldas, Gabriel Costa, Luis Filipe Kopp, Beatriz A. Campos

Abstract

This chapter proposal is to present the main multisensor data fusion concepts and its application on smart environments. It will be shown the main concepts about data multisensor fusion techniques, models, classification and main definitions about smart environments, its relation to multisensor data fusion, multisensor data fusion applications to smart environments, as well as open research opportunities for future research.

Resumo

A proposta do minicurso é apresentar os principais conceitos sobre fusão de dados e a aplicação desse conceitos em Ambientes Inteligentes. Serão abordados os conceitos fundamentais sobre técnicas de fusão de dados, modelos de fusão de dados, suas classificações, e as principais definições de Ambientes Inteligentes, sua relação com fusão de dados, aplicações para fusão de dados em Ambientes Inteligentes, bem como questões em aberto que apontem possibilidades de pesquisas futuras.

5.1. Introdução

Os recentes avanços nas tecnologias de comunicação e computação fomentaram um enorme crescimento no número de dispositivos inteligentes disponíveis para uso [Farias et al.2016]. A integração desses objetos inteligentes na Internet originou o conceito de Internet das Coisas (IoT - *Internet of Things*) [Whitmore et al.2015]. A IoT pode ser compreendida como um mundo de objetos interligados, capazes de serem identificados, endereçados, controlados e acessados via Internet. Esses objetos podem se comunicar uns com os outros, com outros recursos disponíveis na web, com sistemas de informação e usuários humanos. As aplicações de IoT envolvem interações entre vários dispositivos heterogêneos, a maioria deles interagindo diretamente com seu ambiente físico, seja coletando variáveis ou atuando sobre o meio. Entre essas aplicações pode-se citar o monitoramento

e sistemas de decisão presentes ambientes que vão desde: (i) pessoas e partes do corpo, (ii) peças de equipamentos, (iii) eletrodomésticos e carros, (iv) estruturas civis de grande porte como prédios, pontes e plataformas petrolíferas. Essas conjuntos de aplicações e sensores criam o que se conhece como Ambiente Inteligentes [Liu et al.2019].

Novos desafios emergem neste cenário, bem como várias oportunidades a serem exploradas. Uma dessas oportunidades refere-se extração de informação útil para os usuários finais a partir dos grandes volumes de dados produzidos por esses ambientes inteligentes. Nesse contexto, são necessárias técnicas para promover a descoberta do conhecimento a fim de explorar plenamente o uso dos dispositivos da IoT. As técnicas de fusão de dados [Nakamura et al.2007] tratam da associação, correlação e combinação de dados e informações de fontes únicas e múltiplas para obter uma representação consistente e precisa de um objeto ou ambiente do mundo real. Uma vez que os dados produzidos pelos ambientes inteligentes são geralmente dinâmicos e heterogêneos, torna-se importante investigar técnicas de fusão de dados nesse contexto. O emprego dessas técnicas de fusão de dados é útil para revelar tendências nos dados amostrados, diminuir o volume de dados trafegados (e assim minimizar o consumo dos recursos), descobrir novos padrões de variáveis monitoradas, realizar previsões, e com isso aumentar a eficácia dos processos de tomada de decisão, reduzindo os tempos de resposta de decisões e permitindo uma percepção mais inteligente e rápida do ambiente monitorado.

As Tecnologias da Informação e da Comunicação (TIC) desempenham um papel vital na solução dos problemas ambientais causados pela degradação da natureza. Apesar de serem partes do problema por também consumirem energia e serem fontes de poluição, as TICs possuem potencial de contribuir para a redução do consumo de energia através da otimização das operações em diversas áreas (como a geração e distribuição de energia elétrica, controle de tráfego, construção, controle industrial) e, consequentemente, diminuir o desperdício [Huang et al.2018]. Dessa forma, as TICs assumem um importante papel na busca por soluções para o crescimento sustentável e verde das nações.

Um dos campos de pesquisa relacionado ao uso das Tecnologias da Informação e Comunicação (TICs) como provedoras de soluções para os desafios ambientais são os espaços inteligentes (*smart spaces*) [Marchenkov2018]. Um espaço inteligente (ou ambiente de computação pervasiva) pode ser caracterizado como um ambiente com diversos dispositivos, conectados em rede, os quais possuem capacidade de processamento e sensoriamento e que auxiliam os usuários finais na execução de suas tarefas de forma mais eficiente.

As redes elétricas inteligentes (*smart grid*) [Zame et al.2018] podem ser citadas como um dos exemplos de ambientes inteligentes. A rede elétrica inteligente é um tipo de rede de fornecimento e distribuição de energia elétrica (composta por subsistemas para geração, distribuição, transmissão e consumo de energia elétrica) que utiliza as TICs para predizer o comportamento do sistema elétrico e, em caso de problemas (como quedas de energia), atuar (como, por exemplo, ligar sistemas de refrigeração de baterias) [Zame et al.2018].

Há diversos trabalhos na literatura discutindo a questão dos ambientes inteligentes que lidam com o aparecimento de objetos com capacidade de monitoramento, processamento e comunicação nos últimos anos [Culman et al.2019], [Zame et al.2018],

[Santos et al.2019], [de Farias et al.2019], [dos Santos et al.2018], [de Farias et al.2018], [Rogova and Snidaro2019] e [Jorge et al.2018]. Diante disto, aparece o cenário da Internet das Coisas (*Internet of Things* – IoT) [dos Santos et al.2018] onde objetos podem se conectar à Internet e prover comunicação entre usuários, dispositivos (D2D), máquinas (M2M) e formarem novas aplicações. Recentemente, a computação de borda (*Edge Computing*) é vista como o próximo passo dos sistemas de Internet das Coisas [Li et al.2018] [Liu et al.2019]. A computação na borda é uma arquitetura de TIC aberta e distribuída que apresenta poder de processamento descentralizado, capacitando tecnologias de computação móvel e Internet das Coisas (IoT). Na computação na borda, os dados são processados pelo próprio dispositivo, computador ou servidor local, em vez de serem transmitidos para um *data center*. Essa arquitetura têm se mostrado promissoras para o processo de tomada de decisão em ambientes inteligentes [Huang et al.2018].

Nos últimos anos o campo de IoT tem observado diversas mudanças que impactaram o projeto e operação dessas redes. Dentre as diversas mudanças apresentadas há o surgimento das redes de sensores compartilhadas, que ao invés de assumir um projeto de rede tradicional específico para uma única aplicação alvo, a infraestrutura de sensoriamento e comunicação das redes IoT é compartilhada por múltiplas aplicações que podem pertencer a usuários diferentes, otimizando assim a utilização de recursos. Pelo fato de compartilhar a mesma infraestrutura de sensoriamento e comunicação por diversas aplicações, esse tipo de rede passa a ser uma das mais promissoras soluções para as aplicações voltadas para ambientes inteligentes. Há uma série de potenciais vantagens de um projeto de rede de múltiplas aplicações, como a redução significativa nos custos de implantação da rede por permitir que múltiplas aplicações dividam os mesmos nós e infraestrutura de comunicação e sensoriamento, melhorando a utilização global de recursos. Porém, apesar desse potencial, a adoção das redes compartilhadas apresenta novos desafios, os quais devem ser suplantados para se usufruir plenamente de suas vantagens. Tais desafios estão relacionados a funções básicas necessárias para a operação e a gerência das redes, como fusão de dados, escalonamento de tarefas, integração de tarefas, segurança de rede dentre outras, que devem ser adaptadas para esse novo ambiente compartilhado.

Com o aumento no número de aplicações dividindo uma mesma infraestrutura de IoT, há consequentemente um aumento na quantidade de dados coletados pelos sensores (e por consequência um aumento no número de transmissões). A partir dos dados obitodos pelos sensores informações úteis podem ser extraídas. Nós sensores possuam limitações críticas de recursos (baterias removíveis, pouco poder de processamento e memória), apesar disso eles podem realizar ações tais como o processamento de dados como uma forma de solucionar o problema do aumento do número de transmissões em redes de sensores.

Uma forma promissora de reduzir o número de transmissões na rede e, conseqüentemente, de reduzir o consumo de energia dos nós sensores consiste na utilização de técnicas de fusão de dados [Chakraborty et al.2019]. Fusão de dados pode ser definida como o processamento de múltiplas fontes de dados a fim de obter um único dado de saída considerado melhor em termos de precisão ou custo [Chakraborty et al.2019]. Tradicionalmente, as técnicas de fusão de dados processam múltiplas fontes de dados de uma única aplicação para gerar um único dado de saída. Entretanto, podem existir situações nas quais algumas aplicações, que compartilham uma mesma rede IoT, demandam da-

dos que podem ser adquiridos por uma mesma unidade de sensoriamento. Por exemplo, considere duas aplicações dentro do cenário de redes elétricas inteligentes: uma aplicação de monitoramento de linhas de transmissão suspensas (MLTS) e uma aplicação de monitoramento de baterias (MB). A aplicação de MLTS deve suspender a transmissão de energia por uma dada linha de transmissão caso o valor de temperatura ultrapasse um determinado limite (65 ° C). A aplicação de monitoramento de bateria deve suspender o armazenamento de energia de uma bateria caso o valor da temperatura esteja acima de 144 °C. Nota-se que as duas aplicações compartilhariam o sensor de temperatura (fazem uso dos mesmos dados gerados por ele). Nessas situações, não é desejável que os dados relativos às mesmas fontes de dados (unidades de sensoriamento) sejam coletados repetidamente para ambas as aplicações. Bastaria que, para uma mesma fonte de dados e para um determinado instante de tempo, o dado fosse coletado uma única vez e armazenado para posterior consulta e uso pelas respectivas aplicações. Entretanto, as técnicas de fusão de dados tradicionais, quando aplicadas aos dados de uma mesma fonte destinados a múltiplas aplicações, podem resultar na perda da semântica dos dados para algumas destas aplicações. Tal perda se dá porque as aplicações possuem diferentes intervalos de dados. Ao realizar a fusão de dados de aplicações com intervalos tão distintos para o mesmo tipo de dados acabam-se gerando resultados em intervalos que não pertencem a nenhuma das aplicações. Em outras palavras, a semântica dos dados é diferente em aplicações distintas e, por isso, quando a fusão de dados é realizada com técnicas tradicionais há perda da semântica das aplicações (já que os dados resultantes não estão nos intervalos de nenhuma aplicação).

A fim de superar este desafio, a fusão de conhecimento [Dong et al.2014a] surge como uma importante ferramenta para descobrir e limpar erros presentes nas fontes de dados. Os erros podem ser cometidos no processo de extração de conhecimento das fontes. Comparando com a fusão de dados, que visa resolver conflitos de fontes, a fusão de conhecimento considera uma dimensão adicional de erros - os erros cometidos pelos extratores de conhecimento. Entretanto, a fusão do conhecimento foi proposta para um cenário rico em memória e processamento: a web [Preece et al.2000]. Geralmente, os dispositivos na área de ambientes inteligentes têm restrições severas de recursos [Nakamura et al.2007], como energia e processamento. Ainda, de acordo com o modelo de dados de Dasarathy [Nakamura et al.2007], a fusão do conhecimento foi aplicada principalmente em cenários com maiores níveis semânticos de dados (como decisões) [Nakamura et al.2007], enquanto em ambientes inteligentes os dispositivos frequentemente produzem dados em baixos níveis semânticos.

Portanto, o desafio consiste em criar/adaptar técnicas de fusão de dados e ou conhecimento que considerem os dados coletados de uma mesma fonte e destinados a diferentes aplicações realizando tantas reduções quanto possível sem que haja perda da semântica dos dados.

Adicionalmente, a existência de um maior número de aplicações em uma rede IoT pode possibilitar que o processo de tomada de decisão de uma aplicação influencie no processo de tomada de decisão de outra aplicação provocando alterações no comportamento das aplicações. Por exemplo, em geral uma aplicação de MTLS deve deixar que a transmissão de energia elétrica aconteça enquanto o valor informado da temperatura da linha de transmissão indicar que a temperatura está em um valor considerado ideal

(entre 40°C e 65 °C). No entanto, caso o valor da temperatura esteja acima de 65 °C, a aplicação de MLTS indicará que a transmissão de energia deve ser temporariamente interrompida. A aplicação de MLTS decide que deverá avisar a aplicação MB sobre a interrupção da transmissão de forma que essa possa comandar o desligamento da bateria, evitando assim que a bateria fique desnecessariamente ligada, o que pode diminuir a vida útil da mesma. Portanto, a fim de gerir as aplicações em uma rede IoT de forma eficiente surge a noção de integração. A integração é definida como sendo a capacidade de realizar a troca de informação e a colaboração entre aplicações para atingir objetivos comuns [Martins et al.2018]. Nesse contexto, mecanismos responsáveis pela integração de aplicação devem ser concebidos de forma a tornar possível a realização da comunicação entre aplicações distintas através da troca de informação (como a decisão da aplicação de MLTS enviar alerta de interrupção de transmissão para a aplicação MB). Além disso, esse mecanismo de integração de aplicações deve levar em consideração também que decisões de aplicações consideradas mais prioritárias devem ser tratadas em primeiro lugar. Portanto outro desafio consiste em desenvolver metodologias/mecanismos descentralizados voltados para aplicações de ambientes inteligentes que fazem uso de RSAC os quais possibilitem integrar diferentes aplicações dentro da própria rede.

Na busca de investigar soluções para os desafios expostos, este minicurso envisiona apresentar os conceitos fundamentais atrás da fusão de dados para ambientes inteligentes bem como expôr seus principais desafios. Também buscar-se-á apresentar novas técnicas de fusão de dados que busquem resolver os problemas apresentados acima.

O minicurso está organizado da seguinte forma: (i) na seção 2 serão apresentados os conceitos fundamentais de fusão de dados; (ii) na seção 3 serão apresentados os conceitos básicos e exemplos de ambientes inteligentes; (iii) na seção 4 serão apresentados técnicas de fusão de dados que atendam às demandas dos ambientes inteligentes; (iv) na seção 5 será apresentado o *framework* micropython bem como o desenvlvimento de um caso de exemplo usando o referido *framework*. Por fim na seção 6 serão apresentadas breves conclusões.

5.2. Fusão de Dados

Nesta seção serão apresentados os principais conceitos relacionados à fusão de dados. Também serão apresentados os desafios existentes quando se utilizam técnicas de fusão de dados. Serão apresentadas as classificações das técnicas, as técnicas mais tradicionais e os modelos de fusão de dados.

De uma forma mais geral, a fusão de dados pode ser vista como "um processo de múltiplos níveis que lida com a detecção, associação, correlação e estimação de dados provenientes de múltiplos sensores" (Departamento de Defesa dos EUA 1991). No domínio das IoT, técnicas de agregação de dados simples (médias aritméticas, a busca por máximos e mínimos, dentre outras) têm sido usadas para a redução do tráfego de dados com o intuito de reduzir o consumo de energia dos nós sensores. A agregação de dados pode ser definida como a combinação de dados de diferentes nós fontes usando funções triviais (*i.e*, máximo, mínimo, média) que realizam a supressão de mensagens redundantes, e consequentemente, reduzem a quantidade de dados. A eficiência dos algoritmos de agregação de dados depende da correlação entre os dados gerados pelas diferentes fontes

de informação [de Farias et al.2019]. A correlação pode ser espacial, quando os valores gerados por sensores próximos são relacionados; temporal, quando as leituras de sensores mudam lentamente ao longo do tempo; ou semântica, quando as informações de diferentes pacotes de dados podem ser classificadas sob o mesmo grupo semântico, como por exemplo os dados que são gerados por sensores colocados em uma mesma sala. Esse aspecto favorece a eliminação de redundância (uma das metas das técnicas de agregação de dados), mas garante também a acurácia dos dados. Isso é importante, pois a sumarização dos dados pode representar uma perda na acurácia [de Aquino et al.2018], que é um requisito típico para muitas aplicações RSSFs. A acurácia pode ser definida como o grau de proximidade entre a medição observada e o seu real valor esperado. Com uma correlação eficiente dos dados originais é possível alcançar uma maior redução da quantidade de dados para uma mesma acurácia dos dados agregados.

Outros dois conceitos importantes para a eficiência do mecanismo de agregação de dados são: grau e latência [Nakamura et al.2007]. O grau de agregação é definido como o número de pacotes agregados em um único pacote de transmissão; enquanto a latência pode ser medida como o tempo entre os pacotes recebidos no nó sorvedouro e os dados gerados nos nós fontes [de Farias et al.2016]. É importante que a relação entre esses dois conceitos seja equilibrada para que haja eficiência na redução da quantidade de dados por um lado, mas que também não haja atrasos exagerados na entrega final dos dados, por outro lado. A fusão de dados pode ser categorizada em diversos aspectos, a saber: relacionamento entre fontes de dados, nível de abstração e o propósito da fusão de dados. De acordo com o relacionamento entre fontes de dados, a fusão de dados pode ser classificada como complementar, redundante e cooperativa [Farias et al.2016]:

- Complementar: Quando a informação provida pelas fontes representa pedaços de um cenário maior, a fusão pode ser aplicada para obter informações mais completas a cerca do cenário. A fusão complementar busca a completude, formando uma nova informação através da composição de diversas outras (como sensores que verificam a presença de pessoas em quatro cantos diferentes de um aposento e ao fundir essa informação temos a visão completa do aposento).

- Redundante: Se duas ou mais fontes independentes proveem o mesmo pedaço de informação, estes pedaços podem ser fundidos para aumentar a confiabilidade da informação. A fusão de redundância pode ser usada para aumentar a confiabilidade, precisão e credibilidade da informação. Em RSSF, a fusão de redundâncias pode prover informação de alta qualidade e prevenir nós sensores de transmitirem dados iguais (vários sensores de temperatura avaliando a temperatura de uma caldeira industrial).

- Cooperativo: Duas fontes são cooperativas quando a informação provida por elas é fundida em uma nova informação (normalmente mais complexa do que a original), que do ponto de vista da aplicação representa melhor a realidade (um sensor de temperatura e um sensor de fumaça combinando informações para detectar um incêndio).

Quanto ao nível de abstração, [Nakamura et al.2007] a fusão de dados pode ser classificada em quatro níveis:

- Sinal (*signal*): lida com sinais uni ou multi-dimensionais vindos dos sensores (em geral dados brutos vindos dos sensores). Pode ser usado em aplicações de tempo real ou como um passo intermediário entre fusões.

- Pixel: usado em imagens que podem ser utilizadas em tarefas de processamento de multimídia.

- Característica (*feature*): lida com características (ou atributos) extraídas de sinais (como, por exemplo, a temperatura de uma sala) e imagens, como forma e velocidade.

- Símbolo (*symbol*): neste tipo de fusão, a informação é um símbolo que representa uma decisão (como por exemplo, há um símbolo indicando a ação acionar alarme em caso de incêndio), e portanto, esse tipo de fusão também é conhecida como fusão de decisões.

De acordo com o nível de abstração dos dados manipulados, a fusão da informação também pode ser classificada em 4 categorias:

- Fusão de baixo nível: Também conhecida como fusão em nível de sinal. Dados sem processamento são usados como entrada da fusão, combinados em novos dados mais precisos que os originais (aqui se enquadram os dados extraídos das unidades de sensoriamento, como voltagem, amperagem ou campo eletromagnético).

- Fusão de nível médio: atributos ou características de uma entidade (como a forma, textura e posição) são fundidos para a obtenção de um mapa de características que pode ser útil em outras tarefas. Este tipo de fusão também é conhecida como fusão de atributos (dados de temperatura e campo eletromagnético para encontrar danos em linhas de transmissão).

- Fusão de alto nível: decisões ou representações simbólicas são usadas como entrada e são combinadas para obter uma decisão mais confiável ou com uma visão mais ampla do cenário (por exemplo, a decisão de que uma linha de transmissão está danificada e a decisão de que a bateria está danificada gerando a decisão de usar outra linha de transmissão de energia elétrica).

- Fusão de múltiplos níveis: acontece quando a fusão de dados utiliza dados de diferentes níveis de abstração (por exemplo, quando dados de decisão são fundidos com dados do tipo característica, como um dado vindo de um sensor de presença, indicando que não há pessoas na sala, combinado com uma decisão de aquecimento ao realizar a fusão pode concluir que nenhuma ação deve ser dada).

Dasarathy [Martins et al.2018] apresenta outra bem-conhecida classificação para fusão de dados que leva em consideração a abstração dos dados de entrada e saída.O modelo Dasarathy identifica 5 categorias:

- *Data In–Data Out* (DAI-DAO): Nessa classe, a fusão de dados lida com dados em nível de sinale o resultado também em nível de sinal, possivelmente mais preciso ou confiável.
- *Data In–Feature Out* (DAI-FEO): Nessa classe, a fusão de dados usa dados brutos como entrada para extrair atributos ou características que descrevem uma atividade como saída.
- *Feature In–Feature Out* (FEI-FEO): nessa classe, a fusão de dados trabalha sobre um conjunto de características para melhorar ou refinar uma característica ou atributo, ou para extrair novos.
- *Feature In–Decision Out* (FEI-DEO): nessa classe, a fusão de dados usa uma série de características de uma entidade gerando uma representação simbólica ou uma decisão.
- *Decision In–Decision Out* (DEI-DEO): nessa classe, decisões podem ser fundidas de forma a obter novas decisões ou dar ênfase a decisões anteriores.

Ainda outra forma de classificação é baseada no propósito dos métodos de fusão, ou seja, que tipo de informação busca-se extrair dos dados coletados [Nakamura et al.2007]. De acordo com esse critério a fusão de dados pode ser realizada com diferentes objetivos como inferência, estimação, classificação, agregação e compressão.

Métodos de inferência, estimação e classificação são muitas vezes aplicados em fusões de decisão. Nesse caso, uma decisão é tomada baseada no conhecimento da situação percebida. A inferência se refere a transição de uma proposição provavelmente verdadeira, a qual a veracidade é creditada como resultado de uma inferência anterior. Métodos clássicos de inferência são a Inferência Bayesiana [de Farias et al.2019] e a teoria da acumulação de crenças de Dempster-Shafer [de Farias et al.2017].

Métodos de compressão e agregação são usados apenas para redução do volume de dados. A agregação é usada para resolver os problemas de implosão e *Overlapping*. No primeiro, os dados sensoriados são duplicados na rede devido a alguma estratégia de roteamento. O *overlapping* acontece quando dois nós diferentes disseminam os mesmos dados. Os métodos de compressão não são métodos de fusão de dados propriamente ditos, uma vez que eles apenas consideram as estratégias de codificação dos dados. O código de Huffman se enquadra nos métodos de compressão de dados.

5.3. Ambientes Inteligentes e a Internet das Coisas

Conforme apresentado na seção 5.1, um espaço inteligente (ou ambiente de computação pervasiva - *smart space*) pode ser caracterizado como um ambiente com diversos dispositivos, conectados em rede, os quais possuem capacidade de processamento e sensoriamento e que auxiliam os usuários finais na execução de suas tarefas de forma mais eficiente.

Diversos ambientes inteligente surgiram recentemente de forma a gerar mais conforto e/ou segurança para os usuários. A questão ao se trabalhar com ambientes inteligentes é que ambientes distintos possuem requisitos distintos em termos de suas aplicações.

Claramente há uma disparidade na importância entre uma aplicação de controle de temperatura e um sistema de alarme de incêndio. Ou ainda entre um sistema de controle cardíaco de pacientes em um hospital e o controle de iluminação. Ainda uma mesma aplicação pode possuir requisitos diferentes em ambientes distintos. O controle de refrigeração é menos importante em uma casa do que é um hospital.

O termo edifício inteligente (*Smart Building*) segundo [Liu et al.2019] é definido como edifícios equipados com dispositivos inteligentes instalados de forma a minimizar o consumo de energia e sem comprometer o conforto e a segurança do usuário. Portanto, um passo importante rumo a um estilo de vida mais sustentável é melhorar a eficiência energética dos edifícios. Embora os avanços recentes no campo da ciência dos materiais tenham conseguido reduzir o consumo diretamente na estrutura dos edifícios, ainda existem problemas a serem superados [Marchenkov2018]. Um deles relaciona-se a grande quantidade de energia que continua sendo consumida por diversos equipamentos, tais como os aparelhos de controle de temperatura e de iluminação, tanto em edifícios residenciais como comerciais.

Aplicações para edifícios inteligentes mais comumente encontradas na literatura [de Farias et al.2019] [Whitmore et al.2015] incluem: aplicações de controle de aquecimento, ventilação e sistemas de ar condicionado (HVAC); aplicação de iluminação; aplicação de sombreamento; aplicação de qualidade do ar e controle de janelas; aplicações de desligamento de dispositivos; aplicações domésticas (por exemplo, televisores, máquinas de lavar); aplicações de segurança (controle de acesso) e segurança de dispositivos. Tais aplicações muitas vezes fazem uso do mesmo tipo de dado ambiental, como luz, vibração, temperatura, presença, químicos (como fumaça ou gases) e voltagem.

No contexto de *smart building*, recentemente, foram apresentadas na literatura novas propostas, [de Farias et al.2017] [de Farias et al.2019] [Caldas et al.2015] propondo soluções de controle e monitoramento que fazem uso de RSSF. No contexto de *smart grid* não existem soluções de monitoramento devido ao seu alto custo [Gungor et al. 2009]. Nesses ambientes *smart grid* a instalação e manutenção de cabos de comunicação são procedimentos custosos e por isso que os sistemas de monitoramento cabeados não são amplamente utilizados atualmente. Portanto, há uma necessidade urgente por sistemas de monitoramento e diagnóstico sem fio de baixo custo que melhorem a confiabilidade e a eficiência do sistema através da melhora do gerenciamento dos sistemas de geração e transmissão de energia.

O sistema de monitoramento da integridade da estrutura (SHM, do inglês *Structural Health Monitoring*) permite prever danos (fraturas) e, conseqüentemente, antecipar consertos evitando acidentes. Em aplicações construídas para esse fim, os dispositivos de sensoriamento são usados para capturar medidas relacionadas à própria estrutura bem como os eventos externos que afetam a estrutura monitorada e enviar tais medidas para uma entidade centralizada de coleta de dados. A tomada de decisão do SHM quanto a existência de dano é feito de forma centralizada e é efetuada nessa entidade. Normalmente, o monitoramento de estruturas (SHM) é efetuado utilizando-se sensores analógicos cabeados. Entretanto, ao transferir parte do processo de decisão quanto a existência de um dano ou quanto ao mau funcionamento de um dispositivo para dentro dos sensores se obtêm um SHM inteligente. Esse SHM inteligente faz uso de mecanismos descentraliza-

dos capazes de detectar e predizer danos (fraturas) ou mal funcionamento de dispositivos de forma precisa e confiável dentro da própria rede . Ao empregar SHM inteligente por exemplo em estruturas ou dispositivos ligados a uma usina eólica inserida no contexto de *smart grid* permite que tais sistemas gerem energia de forma mais confiável e economicamente eficaz.

O termo Rede elétrica inteligente (*Smart Grid*) segundo [de Farias et al.2018] é definida como sendo uma rede transmissão bidirecional (do gerador para o consumidor e vice-versa) de eletricidade que faz uso das TICs nos sistemas de geração, transmissão, distribuição e uso final de energia com o intuito de melhorar a eficiência, confiabilidade e segurança da geração e fornecimento de energia. Uma vantagem adicional do *smart grid* é que a geração de energia pode ser feita através do uso de energias renováveis e não-poluentes como a energia solar e a energia eólica. A transmissão ser bidirecional implica que será possível conhecer em tempo real a condição de fios, cabos, transformadores e o consumo até de dispositivos específicos instalados em fábricas, escritórios ou domicílios. Com isso será possível controlar esses dispositivos e usar tarifas que variem de acordo com a carga do sistema, hora do dia ou estação para incentivar a conservação de energia, reduzindo seu uso não econômico em horas de demanda máxima.

De acordo com os estudos realizados pela EPRI (*Eletric Power Rsearch Institute*) [Whitmore et al.2015], a rede elétrica inteligente irá gerar uma redução de emissão de gases que varia de 60 a 211 milhões de toneladas de CO_2 nos Estado Unidos até 2030. A comunicação dentro da rede elétrica pode ser feita através fibras ópticas e via tecnologias que usam fios elétricos de média e baixa voltagem como canais para comunicação IP.

Outra aplicação que vêm ganhando destaque é a aplicação de Fazendas Inteligentes [Culman et al.2017]. Estas são fazendas controladas por sensores e ajudam os agricultores em melhorar o desempenho das colheitas e/ou rebanho. Tipicamente, aplicações para fazendas Inteligentes incluem sensores para a medição de umidade do solo, controle de temperatura, irrigação, avaliação da evapotranspiração dos ambientes.

Nas fazendas Inteligentes [Culman et al.2019] [de Farias et al.2017] atuais diversos algoritmos de predição tentam predizer como os fatores edafoclimáticos (do clima e do solo) podem interferir na produção e alterar decisões de irrigação e insumos com base neles. Alguns algoritmos fazem uso de lógica *fuzzy* como no trabalho de [Culman et al.2019] que modela as variáveis ambientais como variáveis *fuzzy* e tenta alterar em tempo real o comportamento dos sistemas de irrigação.

Em [de Farias et al.2017] os autores usam uma técnica de fusão de dados entre os dados de sistemas metereológicos e dos sensores na fazenda de forma a controlar a irrigação. Neste trabalho usou-se um modelo de fusão de dados baseado em teoria de crença no qual conforme os dados metereológicos acertavam ou erravam, ajustavam os pesos de suas informações na decisão de irrigação.

Há ainda o conceito de estufas inteligentes (*smart greenhouses*) que tentam retirar as complexidades impostas pelas intempéries do ambiente através do total isolamento das culturas plantadas em estufas controladas através de sensores [de Farias et al.2017]. Nesse tempo de ambiente o controle de temperatura bem como o controle de iluminação são extremamente importantes.

Hospitais Inteligentes [Hassan et al.2019] são hospitais monitorados por sensores e câmeras que tentam monitorar as condições dos pacientes e das instalações dos hospitais. Várias aplicações são similares à das casas inteligentes. Entretanto há diversas aplicações distintas como a verificação de infecções e contaminação de ambientes. Sensores químicos são muito importantes. Ainda dentro deste contexto o uso de sensores multimídia como câmeras é deveras importante para a predição de quedas de pacientes e monitoramento de movimentação.

Uma segunda categoria de aplicações de hospitais inteligentes diz respeito ao monitoramento dos sinais vitais dos pacientes através das redes de sensores corporais (do inglês *body sensor networks*). O monitoramento de sinais vitais é extremamente complexo e muito suscetível a falsos alarmes. Os falsos alarmes são alarmes para a equipe do hospital sem que haja uma emergência ocorrendo. Esses falsos alarmes levam ao problema da fadiga de alerta no qual as equipes deixam de responder aos alarmes de uma rede corporal pensando ser um falso alarme. Sendo assim técnicas de fusão de dados para esse ambiente específico devem buscar reduzir os falsos alarmes tanto quanto possível.

Há ainda aplicações militares fazendo uso de fusão de dados para a verificação de saúde de soldados em tempo real. Além disso o uso de fusão de dados para a navegação em veículos autônomos [Blasch et al.2018].

Como pode ser observado cada ambiente possui particularidades que devem e desafios que devem ser tratados com extremo cuidado. Na próxima seção serão apresentadas diversas técnicas que buscam resolver esses desafios ainda que parcialmente.

5.4. Domínios de Aplicação de Fusão de Dados em Ambientes Inteligentes

Nesta seção serão apresentados diversas técnicas de fusão de dados usadas em cenários apresentados na seção 5.3.

Conforme dito na seção 5.2 as técnicas de fusão de dados podem ser organizadas de acordo com as saídas dos métodos. Uma técnica a nível de característica famosa é o filtro de média móvel. O Filtro de média móvel é um método amplamente usado no processamento de sinais digitais por ser simples e capaz de reduzir o ruído do sinal. O filtro computa a média aritmética de um número de sinais de entrada para produzir cada ponto do sinal de saída. Dado um sinal z = (z(1), z(2), ...), o verdadeiro sinal x = (x(1), x(2), ...) é estimado por:

$$x(k) = \frac{1}{N}\sum_{i=0}^{M-1} z(k-i), \forall k \geq M, \quad (1)$$

Onde M é o tamanho da janela de fusão, z = z(1),z(2), z(3) ... são os dados de entrada e x = x(1), x(2), x(3) ... são os dados estimados pelo método.

A janela de fusão é o parâmetro mais importante para essa equação uma vez que M é usado para a detecção de qualquer mudança no nível do sinal; quanto maior o valor de M, mais claro será o sinal.

Apesar de fácil compreensão e baixa complexidade, o filtro de média móvel lida apenas com medições em nível de sinal e característica, mas não consegue lidar com

níveis semânticos mais altos, como decisões.

No filtro de média móvel, todos os dados possuem o mesmo peso. Em um ambiente com múltiplas aplicações, algumas medidas podem ser mais importantes para uma determinada aplicação do que para outra aplicação. Por exemplo, uma aplicação de detecção de incêndio é mais importante do que uma aplicação de AVAC e valores de temperatura mais altos são mais importantes para a detecção de incêndio do que para a aplicação de AVAC. Um valor de temperatura de 50 C possui pouco valor para a aplicação de AVAC, uma vez que o intervalo de operação desta aplicação não considera valores de temperatura tão altos, entretanto é um ótimo indicativo de incêndio. Se todas as medidas forem consideradas da mesma forma, pode-se encontrar um resultado que não reflete o ambiente sensoriado corretamente, um resultado tendencioso.

No Filtro de média móvel aperfeiçoado, há a necessidade de avaliar determinada medida considerando uma aplicação alvo. De forma a tornar essa abordagem possível, pode-se modificar a abordagem tradicional para a apresentada na equação 10:

$$x(k) = \frac{1}{N}\sum_{i=0}^{M-1} \mu z(k-i), \forall k \geq M, \mu > 0, \mathrm{N} = \sum_{i=0}^{M-1} \mu \tag{4}$$

Onde M é o tamanho da janela de fusão, z = z(1),z(2), z(3) ... são os dados de entrada, x = x(1), x(2), x(3) ... são os dados estimados pelo método, μ é o peso dado baseado na importância da aplicação e N é a soma de todos os pesos . Um especialista de domínio deve escolher os pesos apropriadamente.

Um outro trabalho já usa o conceito de distribuir pesos diferentes no filtro de média móvel. O EWMA (Exponentially weighted moving average), uma variante do filtro de média móvel apresentada em [Farias et al.2016], distribui pesos para os dados de cada sensor como forma de melhor avaliar o ambiente. Nossa proposta é diferente do EWMA uma vez que ao invés de distribuir pesos para os dados de diferentes sensores, assume-se que todos os sensores são capazes de monitorar o mesmo tipo de informação (temperatura por exemplo), mas a relevância dessa informação variará de acordo com os requisitos de cada aplicação que for utilizá-la.

Entretanto dentro dos ambientes inteligentes as técnicas de fusão de dados a nível de decisões possuem ainda mais destaque uma vez que evitam que os dados coletados precisem sair da rede. Entretanto essas técnicas devem ser capazes de ser implementadas em sensores.

Uma das técnicas de fusão de dados mais utilizadas a nível de decisão é a Inferência Bayesisana (muitas vezes chamada de *Naive Bayes*). A inferência bayesiana utiliza uma combinação de evidências de acordo com certas regras de probabilidades. O grau de incerteza é representado por probabilidades condicionais como mostrado na equação 1:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \tag{1}$$

Onde Pr(Y | X) representa a probabilidade de que a hipótese Y seja verdadeira dada a informação X. Essa probabilidade é obtida através da multiplicação de Pr(Y), a

probabilidade anterior da hipótese Y, por Pr(X| Y), a probabilidade de que X seja verdadeiro dado que Y seja verdadeiro; Pr(X) pode ser considerada uma constante normalizadora. O maior problema da Inferência Bayesiana é que as probabilidades Pr(X) e Pr(X|Y) devem ser conhecidas a priori sem nenhuma garantia dos valores reais. Não há testes com dados reais para garantir que as probabilidades representem cenários reais. Apesar disso, a Inferência Bayesiana é usada no cenário de IoT devido à sua simplicidade e baixo consumo de recursos.

Para estender os métodos probabilísticos, tais como a Inferência Bayesiana, uma abordagem recursiva é proposta neste trabalho. Na Inferência Bayesiana aprimorada (equação 8), a expressão BI(Y | X) representa a crença de que a aplicação Y terá um determinado comportamento dado o resultado da aplicação X. Por exemplo, sempre que uma aplicação de detecção de incêndio detecta um incêndio em potencial, a aplicação de AVAC será desligada, uma vez que quando um determinado limiar de temperatura é encontrado, e esse limiar representa a ocorrência de um incêndio, não há necessidade de manter a aplicação de AVAC (que tem como objetivo prover conforto aos ocupantes do edifício) em operação. Nesse momento, todos os dados de temperatura providos pelos sensores serão relevantes somente para a aplicação de detecção de incêndio. [Farias et al.2014]:

$$BI(A \mid B) = \frac{BI(B \mid A)\, BI(A)}{BI(B)} \tag{2}$$

os termos BI(A) e BI(B) são inferências bayesianas que consideram o conjunto de estados de cada aplicação separadamente.

Entretanto essa extensão ainda sofre com a limitação da Inferência Baeysiana tradicional: as probabilidades devem ser conhecidas a priori, ou sejam antes do início da operação do sistema de decisão [Farias et al.2014].

Esse método é baseado na teoria de acumulação de crença de Dempster-Shafer, que é uma teoria introduzida por Dempster-Shafer (DEMPSTER E SHAFER 1974), e generaliza a teoria bayesiana. Dempster-Shafer lida com crenças ou funções de massa da mesma forma que a teoria Bayesiana lida com probabilidades. A teoria de probabilidades provê um formalismo que pode ser usado para a representação de conhecimentos incompletos, combinação de evidências e atualização de crenças.

Um conceito fundamental no sistema de inferência de Dempster-Shafer é o quadro de discernimento, que é definido como segue. Seja $\Theta = \{\theta_1, \theta_2, ..., \theta_N\}$ o conjunto de todos os estados possíveis que descrevem o sistema, dado que θ é exaustivo e mutuamente exclusivo, no sentido em que o sistema está em apenas um dado estado θ_i, onde $1 \leq$ i $\leq$ N. O intervalo 1 até N é chamado de quadro de discernimento porque seus elementos são usados para discernir os estados do sistema.

Os elementos do conjunto 2^{Θ} são chamados de hipóteses. Na teoria de Dempster-Shafer, baseado na evidência E, uma probabilidade é relacionada para cada hipótese $H \in 2^{\Theta}$, de acordo com a função de massa $m : 2^{\Theta} \rightarrow [0,1]$ que satisfaz as condições abaixo:

$$m(\emptyset) = 0 \tag{3}$$

$$m(H) \geq 0, \forall H \in 2^{\Theta} \tag{4}$$

$$\sum_{H \in 2^{\Theta}} m(H) = 1 \tag{5}$$

O grau de crença de uma hipótese é definido pela função bel(H) como visto em seguida $bel : 2^{\Theta} \rightarrow [0,1]$ sobre Θ como:

$$bel(H) = \sum_{A \subseteq H} m(A), \tag{6}$$

onde $bel(\emptyset) = 0$, e $bel(\Theta) = 1$.

A partir da crença podemos calcular dois outros graus de pertencimento: dúvida (dou) e plausibilidade (pl) como mostrado nas equações 4 e 5:

O grau de dúvida em H pode ser expressado de forma intuitiva em termos da função de crença $bel : 2^{\Theta} \rightarrow [0,1]$ como:

$$dou(H) = bel(\neg H) = \sum_{A \subseteq \neg H} m(A). \tag{7}$$

Para expressar a plausibilidade de cada hipótese, a função $pl : 2^{\Theta} \rightarrow [0,1]$ sobre Θ é definido como:

$$pl(H) = 1 - dou(H) = \sum_{A \cap H = \emptyset} m(A) \tag{8}$$

A plausibilidade intuitivamente diz que quanto menor a dúvida na hipótese H, mais plausível é.

Dessa forma, para combinar o efeito de dois conjuntos de probabilidades (m1 e m2) sobre uma mesma hipótese, a teoria de Dempster-Shafer define a seguinte regra de combinação satisfazendo $m_1 \oplus m_2(\emptyset) = 0$:

$$m_1 \oplus m_2(H) = \frac{\sum_{X \cap Y = H} m_1(X) m_2(Y)}{1 - \sum_{X \cap Y = \emptyset} m_1(X) m_2(Y)} \tag{9}$$

Na inferência de Dempster-Shafer, probabilidades não são associadas as hipóteses a priori. Ao invés disso, probabilidades são associadas somente quando a informação de suporte está disponível, ou seja, quando há dados coletados para serem avaliados. Por outro lado, a inferência de Dempster-Shafer é totalmente dependente dos dados coletados, o que significa que ser houver poucos dados coletados o resultado pode ser impreciso ou tendencioso. Para superar esse problema deve-se analisar quantos dados de entrada são necessários para cada aplicação.

Escolher entra a inferência Bayesiana e Dempster-Shafer não é uma decisão trivial, uma vez que há uma troca entre a acurácia da inferência Bayesiana e a flexibilidade trazida por Dempster-Shafer [Nakamura et al.2007]. Em um cenário não estático (como o das RSAC), a acurácia da Inferência Bayesiana é reduzida devido as mudanças no ambiente. Nesse caso, a flexibilidade trazida pela inferência de Dempster-Shafer passa a ser interessante, uma vez que a função de crença se adaptará aos novos dados enquanto as probabilidades da inferência Bayesiana se mantêm estáticas [Suganuma et al.2018].

O método Dempster-Shafer foi aperfeiçoado, pois é uma variante dos métodos probabilísticos que não necessita conhecer as probabilidades dos eventos a priori, conforme apresentado no capítulo 5.2. No Dempster-Shafer aperfeiçoado, cada aplicação terá seu próprio conjunto de hipóteses. Nessa abordagem, a função belief representará a crença que uma aplicação Y apresentará um determinado comportamento dado o resultado de outra aplicação X. Nesse contexto, H é o conjunto de estados que representam o comportamento de uma aplicação. Assim deriva- se a equação 9:

$$m_1 \oplus m_2(H) = \frac{\sum\limits_{X \cap Y = H} DS_1(X) DS_2(Y)}{1 - \sum\limits_{X \cap Y = \emptyset} DS_1(X) DS_2(Y)} \tag{10}$$

Onde DS1 e DS2 são inferências Dempster-Shafer sobre as condições das aplicações isoladas e o numerador da equação representa a crença de que as duas aplicações estarão em um dado estado simultaneamente.

Considerando o exemplo da ocorrência de um incêndio, X é uma aplicação de detecção de incêndio e Y é uma aplicação de AVAC. O numerador da equação 9 representa o grau de dúvida que aparece quando DS1 infere sobre um estado da aplicação, por exemplo que em uma dada condição de temperatura deve haver uma alerta de incêndio, DS2 indicará que na mesma condição de temperatura, a aplicação Y irá desligar o arcondicionado. Não há necessidade da aplicação de AVAC permanecer ativa durante um incêndio, uma vez que o uso de um sistema de ventilação aumentaria os riscos de problemas na fiação elétrica que poderiam aumentar o incêndio. Isto é importante pois será possível evitar o desperdício de recursos através da eliminação de estados repetidos de duas ou mais aplicações. O denominador significa a plausabilidade dessa hipótese.

Uma outra família de técnicas que surgiu om o intuito de lidar com as incertezas sobre os dados coletados é a lógica subjetiva. Lógica Subjetiva (LS) é um tipo de lógica que utiliza probabilidade para calcular o grau de certeza ou crença em determinado indivíduo ou na ocorrência de um evento. A ideia é que para a ocorrência de um dado evento, em que não é possível estimar ao certo sua probabilidade de ocorrência, seja adicionado um valor de incerteza [Jøsang2001].

A LS trabalha utilizando quatro parâmetros, e são eles [Jøsang et al.2006]: *belief* (b, crença)e *disbelief* (d, descrença), que são influenciados por eventos passados provenientes do indivíduo, sendo estes eventos positivos e negativos, onde eventos positivos aumentam a crença e vice-versa; *uncertainty* (u) (incerteza), grau de incerteza (ou ignorância) que se possui acerca de um evento, este parâmetro diminui conforme o número de observações do indivíduo aumenta; α ou a, pode ser considerado como a confiança inicial

que se pode ter sobre um novo nó da rede ou como a probabilidade real de acontecimento de um evento, caso este valor seja conhecido.

$$\omega_x^A = (b, d, u, a) \tag{11}$$

Juntos, os quatro parâmetros formam a *opinião* de um indivíduo em outro. A expressão presente em 11 denota a opinião de um indivíduo A acerca de x [Jøsang et al.2006]. As propriedades dos parâmetros estão presentes nas equações 12 e 13. Note que, para o maior grau possível de incerteza, $u = 1$, os valores de b e d são zero e pode-se dizer que é o estado de completa ignorância acerca da ocorrência de um determinado evento. Em contra partida, para o menor grau possível de incerteza sobre um evento ($u = 0$), diz-se que a opinião sobre este é dogmática e pode obter este valor após infinitas observações [Jøsang et al.2006].

$$b, d, u, a \in [0, 1] \tag{12}$$

$$b + d + u = 1 \tag{13}$$

Para calcular os parâmetros b, d e u são utilizadas as fórmulas presentes em 14. Onde p corresponde ao número de eventos positivos gerados por um indivíduo e n o número de eventos negativos gerados pelo mesmo indivíduo. Em 14, k é uma constante, com valor geralmente 1 ou 2 que determina o quão rápido a crença no elemento avaliado é construída [**?**].

$$b = \frac{p}{p+n+k}, d = \frac{n}{p+n+k}, u = \frac{k}{p+n+k}. \tag{14}$$

É possível ilustrar a a utilização da LS através de dois nós A e B, que se relacionem de alguma forma. No exemplo, A possui uma opinião acerca de B. Com valores fictícios, esta opinião pode ser expressa pela equação $w_B^A = (0.24, 0.43, 0.34, 0.5)$. Neste caso, $b = 0.24$, $d = 0.43$, $u = 0.34$ e $a = 0.5$. O espaço de opinião de A com relação a B pode ser mapeado no interior de um triângulo equilátero, onde b, d e u identificam a posição da opinião no espaço [Jøsang et al.2006], como mostra a Figura 5.1.

Caso existam, para um mesmo evento, duas opiniões diferentes, o operador de '*consenso*' tem por objetivo calcular um acordo entre as duas opiniões [Jøsang et al.2006]. Para tal, supõe-se que A e B possuem cada um, uma opinião acerca de x, onde estas opiniões são: $\omega_x^A = (b_x^A, d_x^A, u_x^A, a_x^A)$ e $\omega_x^B = (b_x^B, d_x^B, u_x^B, a_x^B)$. Segundo [Jøsang2016, Jøsang et al.2006] o calculo do consenso entre duas opiniões se divide em dois casos, e são descritos como as equações 15 e 5.4. A equação 5.4 também pode ser utilizada nos casos em que uma das opiniões é dogmática [Martinsson2005]. Uma demonstração gráfica do consenso aplicado é demonstrada na Figura 5.2, onde ele é aplicado em para a junção de duas opiniões distintas.

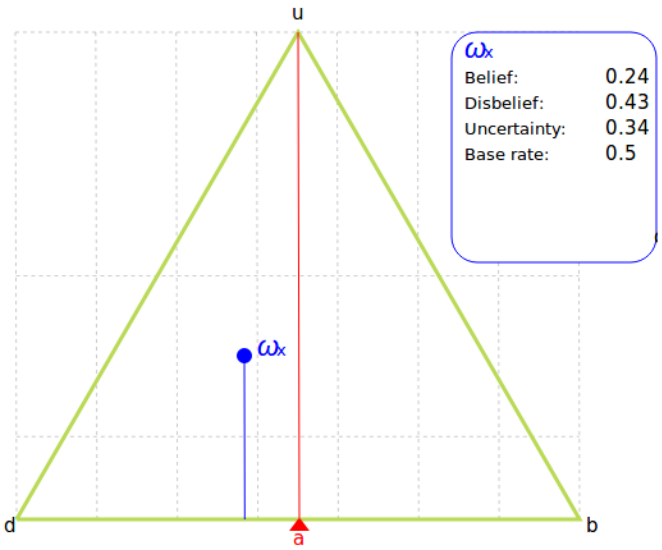


Figura 5.1. Espaço de opinião de w_B^A [Jøsang et al.2006].

Caso I: $u_x^A + u_x^B - u_x^A u_x^B \neq 0$:

$$b^{AB} = \frac{b_x^A u_x^B + b_x^B u_x^A}{u_x^A + u_x^B - u_x^A u_x^B}, d^{AB} = \frac{d_x^A u_x^B + d_x^B u_x^A}{u_x^A + u_x^B - u_x^A u_x^B}, u^{AB} = \frac{u_x^B u_x^A}{u_x^A + u_x^B - u_x^A u_x^B}, a^{AB} = \frac{a_x^A + a_x^B}{2}. \quad (15)$$

Caso II: $u_x^A + u_x^B - u_x^A u_x^B = 0$:

$$b^{AB} = \gamma_x^A b_x^A + \gamma_x^B b_x^B, d^{AB} = \gamma_x^A d_x^A + \gamma_x^B d_x^B, u^{AB} = 0, a^{AB} = \gamma_x^A a_x^A + \gamma_x^B a_x^B.$$

Onde:

$$\gamma_x^A = \lim_{u_x^B \to 0, u_x^A \to 0} \frac{u_x^B}{u_x^A + u_x^B}.$$
$$\gamma_x^B = \lim_{u_x^B \to 0, u_x^A \to 0} \frac{u_x^A}{u_x^A + u_x^B}. \quad (16)$$

Os trabalhos recentes de [de Andrade Campos et al.2019] e [de Andrade Campos et al.] utilizaram a lógica sujetiva para detecção de anomalias. Nesse trabalhos os autores colocaram pesos nos canais de comunicação que eram modificados conforme as decisões se provavam corretas ou não. Dessa forma canais de comunicação pouco confiáveis podiam ser ignorados.

Finalmente, a fusão de conhecimento [Dong et al.2014a] surge como uma importante ferramenta para descobrir e limpar erros presentes nas fontes de dados. Os erros podem ser cometidos no processo de extração de conhecimento das fontes. Comparando com a fusão de dados, que visa resolver conflitos de fontes, a fusão de conhecimento considera uma dimensão adicional de erros - os erros cometidos pelos extratores de conhecimento. Entretanto, a fusão do conhecimento foi proposta para um cenário rico em recursos (memória e processamento): a web [Dong et al.2014a] [Dong et al.2014b] [Oliveira et al.2019]. Geralmente, os dispositivos na área de ambientes inteligentes têm

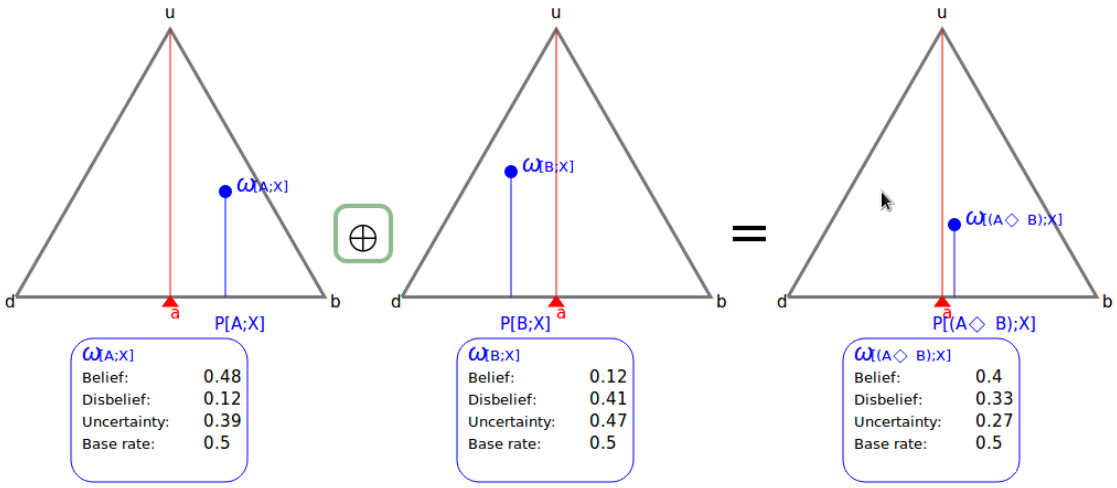


Figura 5.2. Representação gráfica do consenso entre duas opiniões.

restrições severas de recursos [Nakamura et al.2007], como energia e processamento. Ainda, de acordo com o modelo de dados de Dasarathy [Nakamura et al.2007], a fusão do conhecimento foi aplicada principalmente em cenários com maiores níveis semânticos de dados (como decisões) [Nakamura et al.2007], enquanto em ambientes inteligentes os dispositivos frequentemente produzem dados em baixos níveis semânticos (como dados brutos).

Recentemente, o termo *Ai on Edge* (*Artifical Intelligence*) passou a ser um dos mais requisitados em ambientes inteligentes e significa o uso de técnicas de aprendizado de máquina e/ou de Inteligência Artificial nas bordas das redes IoT. Vale destacar que apesar das redes neurais serem opções naturais neste tipo de aplicação há outras opções viáveis como redes neurais sem peso [Cardoso et al.2018] que são mais rápidas e consomem menos recursos do que as redes neurais tradicionais.

Outro ponto relevante é que as técnicas apresentadas podem fazer parte de técnicas mais complexas. Por exemplo nos trabalhos de [de Farias et al.2016] [de Farias et al.2019] [de Farias et al.2018] [de Farias and Pirmez2017] os autores buscam diferenciar eventos dentro de massas de dados para só então aplicar técnicas de fusão de dados. Ou seja pode haver momentos em que as técnicas apresentadas sejam somente mais uma etapa do sistema de decisão.

5.5. Construindo um sistema de decisão simples usando Fusão de Dados e Micropython

Nesta seção será apresentado como construir uma aplicação que faça uso de fusão de dados. Usar-se-á como domínio de aplicação um edifício inteligente. Como ambiente de desenvolvimento será usado o *framework* Micropython [Norris2016].

A ideia é implementar um filtro de média móvel bem simples que possa ser utilizado para controlar uma aplicação de controle de ventilação. A implementação do algoritmo para esta solução foi dada em Micropython – uma implementação do software da linguagem de programação baseada em Python, escrita em C, otimizada para microcon-

troladores.

O MicroPython é uma implementação enxuta e eficiente da linguagem de programação Python 3 que inclui um pequeno subconjunto da biblioteca padrão do Python e é otimizada para ser executada em microcontroladores e em ambientes restritos.

```
import pyb

# turn on an LED
pyb.LED(1).on()

# print some text to the serial console
print('Hello MicroPython!')
```

Figura 5.3. Hello world!

O MicroPython é um compilador e um tempo de execução completos do Python que são executados no *bare-metal*. Você recebe um *prompt* interativo (o REPL) para executar comandos imediatamente, com a capacidade de executar e importar scripts do sistema de arquivos interno. O REPL possui histórico, preenchimento de guias, modo de recuo automático e colar para uma ótima experiência do usuário.

O MicroPython se esforça para ser o mais compatível possível com o Python normal (conhecido como CPython), de modo que, se você conhece Python, já conhece o MicroPython. Por outro lado, quanto mais você aprende sobre o MicroPython, melhor você se torna no Python. Bibliotecas tradicionais do Python como a biblioteca *Math* já possuem versões equivalentes para Micropython. Recentemente a biblioteca de cálculos matriciais Numpy teve uma implementação feita para Micropython.

```
1  # a subset of the Python Math library
2
3  import math
4  import cmath
5
6  print(math.sqrt(5))
7  print(math.log10(100))
8  print(math.sin(12345) ** 2 + math.cos(12345) ** 2)
9  print(math.cosh(1) ** 2 - math.sinh(1) ** 2)
10 print(cmath.polar(1 + 1j))
11
```

Figura 5.4. Operações Matemáticas

Além de implementar uma seleção das principais bibliotecas Python, o MicroPython inclui módulos como "máquina"para acessar hardware de baixo nível. Dessa forma com o Micropython é possível acessar dados de sensores não inclusos com a placa bem como extrair informações do hardware das placas.

Pode-se simular o funcionamento de uma placa usando um simulador online através da página https://micropython.org/unicorn/. Nela pode-se ver um emaulador da placa pyboard (placa desenvolvida pelo projeto micropython) em pleno funcionamento. Diversos algoritmos e exemplos estão presentes para estudo.

Há uma IDE (*Integrated Development Environment*) chamada upycraft que permite o desenvolvimento em micropython bem como automatiza o processo de instalação de drivers e afins. Tendo disponibilidade para Windows, Linux e MacOSX a IDE é uma opção que facilita o desenvolvimento de aplicações em micropython.

```
from machine import Pin, I2C

# creat an I2C bus
i2c = I2C(scl=Pin('X1'), sda=Pin('X2'))

# scan for list of attached devices
dev_list = i2c.scan()

# write to and read from a device
i2c.writeto(0x42, b'\x04')
data = i2c.readfrom(0x42, 4)

# memory transactions
i2c.writeto_mem(0x42, 0x12, b'\x')
data = i2c.readfrom_mem(0x42, 0x12, 2)
```

Figura 5.5. Trabalhando com a máquina

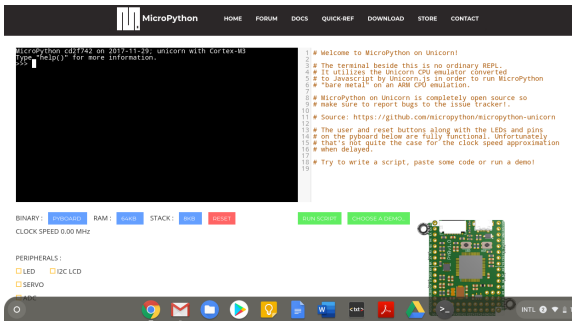


Figura 5.6. Emulador Pyboard

Com a finalidade de conseguir executar os casos de teste foi preciso um microcontrolador para compilar e executar o código construído a partir do algoritmo. O microcontrolador utilizado foi o ESP8266 (da plataforma NodeMCU), muito utilizado para desenvolvimentos IoT e difundido pela facilidade na sua utilização e instalação. A ESP8266 (Figura 5.9) possui um processador ESP8266-12E, que pode operar em 80MHz/160MHz, 4Mb de memória flash, 64Kb para instruções e 96Kb para dados.

Para o desenvolvimento do algoritmo do filtro de média móvel escolheu-se o uso de sensores DHT que medem a temperatura do ambiente. A ideia seria replicar o comportamento de um sistema de controle de ventilação. O ideal seria permanecer entre 18 e 23 graus. Portanto o sistema deveria ler algumas medidas de temperatura para encher a janela de fusão, realizar o filtro de média móvel, e se a temperatura estiver fora da faixa de conforto acionar o sistema de ventilação.

O algoritmo pode ser visto na Figura 5.8. Note que não necessariamente este é o modo mais eficiente de implementação mas para critérios didáticos o algoritmo é suficiente para demonstrar o quão simples é fazer um sistema com Micropython.

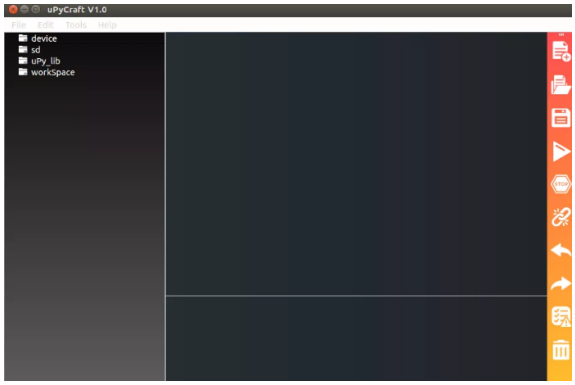


Figura 5.7. Interface do upycraft

Figura 5.8. NodeMCU: ESP8266 microcontrolador

5.6. Conclusões

Nesta seção serão apresentadas as considerações finais referentes aos assuntos abordados, realizando uma retrospectiva das principais questões discutidas e ressaltando os benefícios trazidos pelas fusão de dados para Ambientes Inteligentes.

Conforme pode ser observado diversas aplicações de ambientes inteligentes possuem requisitos em relação ao tempo de resposta das aplicações. Nesses casos o tempo necessário para coleta e processamento fora da rede pode ser maior do que o requerido pela aplicação. Não somente esse fato mas conforme as redes IoT se popularizam aumenta o número de aplicações que compartilham a mesma infraestrutura. Com isso além do aumento do número de mensagens na rede e do maior consumo de energia há a possibilidade de conflito entre diferentes aplicações em um mesmo ambiente.

Apesar da fusão de dados aparecer como uma das possíveis soluções para os desafios apresentados, as técnicas tradicionais não são adequadas para lidar com os múltiplos requisitos dessas múltiplas aplicações presentes nos ambientes inteligentes. Sur-

```
from machine import Pin
from time import sleep
import dht

sensor = dht.DHT22(Pin(14))
fus_wnw = []
maf = 0

while True:
  try:
    sleep(2)
    sensor.measure()
    temp = sensor.temperature()
    temp_f = temp * (9/5) + 32.0
    fus_wnw.add(temp_f)
    # escolheu-se uma janela de 5 elementos
    if (fus_wnw.lenght() == 5):
        # calculando o filtro
        for i in fus_wnw:
            temp +=i/5
    if (maf >= 23 and maf <= 21):
        activate_HVAC()
  except OSError as e:
    print('Failed to read sensor.')
```

Figura 5.9. Algoritmo do Filtro em Micropython

giram diversas novas técnicas capazes de lidar com esse desafio [de Farias et al.2019] [Caldas et al.2015] [de Farias et al.2018] [de Andrade Campos et al.] [de Farias et al.2016] [Farias et al.2016] [Farias et al.2014].

Embora tenham havido esforços ainda existem muitos desafios em aberto na área como: (i) fusão em *Streams* de dados, (ii) técnicas de fusão que lidam com os dados de múltiplas aplicações simultaneamente, (iii) mineração de dados, (iv) Sensoriamento Participativo, (v) predição de dados, (vi) *Ai on Edge* e (vii) fusão de dados com fontes heterogêneas, tanto no nível de dispositivo quanto no nível de aplicações.

Referências

[Blasch et al.2018] Blasch, E., Boril, J., Smrz, V., and Leuchter, J. (2018). Pilot interface considerations using high level information fusion. In *2018 IEEE Aerospace Conference*, pages 1–8. IEEE.

[Caldas et al.2015] Caldas, G., de Farias, C. M., Pirmez, L., and Delicato, F. C. (2015). S-leach: A leach extension for shared sensor networks. In *Proceedings of the International Conference on Wireless Networks (ICWN)*, page 121. The Steering Committee of The World Congress in Computer Science, Computer . . . .

[Cardoso et al.2018] Cardoso, D. O., Gama, J., and França, F. (2018). Weightless neural modeling for mining data streams. *SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE*, 83(1):26–43.

[Chakraborty et al.2019] Chakraborty, I., Chakraborty, A., and Das, P. (2019). Sensor selection and data fusion approach for iot applications. In *Recent Developments in Machine Learning and Data Analytics*, pages 17–33. Springer.

[Culman et al.2019] Culman, M., de Farias, C. M., Bayona, C., and Cruz, J. D. C. (2019). Using agrometeorological data to assist irrigation management in oil palm crops: A

decision support method and results from crop model simulation. *Agricultural water management*, 213:1047–1062.

[Culman et al.2017] Culman, M., Portocarrero, J. M., Guerrero, C. D., Bayona, C., Torres, J. L., and de Farias, C. M. (2017). Palmnet: An open-source wireless sensor network for oil palm plantations. In *2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC)*, pages 783–788. IEEE.

[de Andrade Campos et al.] de Andrade Campos, B., de Farias, C. M., and da Costa Carmo, L. F. R. Using trusted networks to detect anomaly nodes in internet of things. In *Proceedings of the 22nd International Conference on Information Fusion (FUSION 2019)*, volume 1.

[de Andrade Campos et al.2019] de Andrade Campos, B., de Farias, C. M., and da Costa Carmo, L. F. R. (2019). Utilização de redes de confiança para detecção de anomalias em rssf. In *IV Workshop sobre Regulação, Avaliação da Conformidade, Testes e Padrões de Segurança 2018*, volume 1. Galoá.

[de Aquino et al.2018] de Aquino, G. R. C., de Farias, C. M., and Pirmez, L. (2018). Data quality assessment and enhancement on social and sensor data. In *BiDu-Posters@ VLDB*.

[de Farias et al.2016] de Farias, C. M., Li, W., Delicato, F. C., Pirmez, L., Pires, P. F., and Zomaya, A. Y. (2016). Seraph: Service allocation algorithm for the execution of multiple applications in heterogeneous shared sensor and actuator networks. In *Management of Cyber Physical Objects in the Future Internet of Things*, pages 93–113. Springer.

[de Farias and Pirmez2017] de Farias, C. M. and Pirmez, L. (2017). A multisensor data fusion technique for multiapplication wireless sensor networks based on overlapping intervals. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 804–811. IEEE.

[de Farias et al.2018] de Farias, C. M., Pirmez, L., and Delicato, F. C. (2018). Density based multisensor data fusion for multiapplication wireless sensor networks. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 814–821. IEEE.

[de Farias et al.2017] de Farias, C. M., Pirmez, L., Delicato, F. C., Pires, P. F., Li, W., Zomaya, A. Y., de LF Jorge, E. N., and Juarez-Ramirez, R. (2017). Grown: A control and decision system for smart greenhouses using wireless sensor networks. In *Proceedings of the Australasian Computer Science Week Multiconference*, page 48. ACM.

[de Farias et al.2019] de Farias, C. M., Pirmez, L., Fortino, G., and Guerrieri, A. (2019). A multi-sensor data fusion technique using data correlations among multiple applications. *Future Generation Computer Systems*, 92:109–118.

[Dong et al.2014a] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014a). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM.

[Dong et al.2014b] Dong, X. L., Gabrilovich, E., Heitz, G., Horn, W., Murphy, K., Sun, S., and Zhang, W. (2014b). From data fusion to knowledge fusion. *Proceedings of the VLDB Endowment*, 7(10):881–892.

[dos Santos et al.2018] dos Santos, I. L., Costa, G. M. d. O., de Farias, C. M., Pirmez, L., Delicato, F. C., and Diego, M. d. A. (2018). A holistic approach to challenges in industry 4.0. In *Anais da V Escola Regional de Sistemas de Informação do Rio de Janeiro*, pages 24–29. SBC.

[Farias et al.2014] Farias, C., Pirmez, L., Delicato, F., Carmo, L., Li, W., Zomaya, A. Y., and de Souza, J. N. (2014). Multisensor data fusion in shared sensor and actuator networks. In *17th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.

[Farias et al.2016] Farias, C. M. D., Li, W., Delicato, F. C., Pirmez, L., Zomaya, A. Y., Pires, P. F., and Souza, J. N. D. (2016). A systematic review of shared sensor networks. *ACM Computing Surveys (CSUR)*, 48(4):51.

[Hassan et al.2019] Hassan, M. K., El Desouky, A. I., Elghamrawy, S. M., and Sarhan, A. M. (2019). Big data challenges and opportunities in healthcare informatics and smart hospitals. In *Security in Smart Cities: Models, Applications, and Challenges*, pages 3–26. Springer.

[Huang et al.2018] Huang, J., Liu, Z., Duan, Q., Atiquzzaman, M., Jo, M., and Haas, Z. J. (2018). Green computing and communications for smart portable devices. *Wireless Communications and Mobile Computing*, 2018.

[Jorge et al.2018] Jorge, E. N. d. L. F., de Farias, C. M., dos Santos, I. L., and de Souza Pereira, M. S. (2018). An analysis of the network of rural producers in the state of rio de janeiro. In *BiDu-Posters@ VLDB*.

[Jøsang2001] Jøsang, A. (2001). A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(03):279–311.

[Jøsang2016] Jøsang, A. (2016). *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated.

[Jøsang et al.2006] Jøsang, A., Hayward, R., and Pope, S. (2006). Trust network analysis with subjective logic. In *Proceedings of the 29th Australasian Computer Science Conference-Volume 48*, pages 85–94. Australian Computer Society, Inc.

[Li et al.2018] Li, H., Ota, K., and Dong, M. (2018). Learning iot in edge: Deep learning for the internet of things with edge computing. *IEEE Network*, 32(1):96–101.

[Liu et al.2019] Liu, Y., Yang, C., Jiang, L., Xie, S., and Zhang, Y. (2019). Intelligent edge computing for iot-based energy management in smart cities. *IEEE Network*, 33(2):111–117.

[Marchenkov2018] Marchenkov, S. (2018). Infrastructure multi-layer model for smart spaces middleware development. In *Proceedings of the 22st Conference of Open Innovations Association FRUCT*, page 51. FRUCT Oy.

[Martins et al.2018] Martins, G., de Farias, C. M., and Pirmez, L. (2018). Athena: A knowledge fusion algorithm for the internet of things. In *Proceedings of the 14th ACM International Symposium on QoS and Security for Wireless and Mobile Networks*, pages 92–99. ACM.

[Martinsson2005] Martinsson, H. (2005). An evaluation of subjective logic for trust modelling in information fusion.

[Nakamura et al.2007] Nakamura, E. F., Loureiro, A. A., and Frery, A. C. (2007). Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Computing Surveys (CSUR)*, 39(3):9.

[Norris2016] Norris, D. (2016). *Python for Microcontrollers: Getting Started with MicroPython*. Mcgraw-hill Education-Europe.

[Oliveira et al.2019] Oliveira, J., Farias, C. M., Pacitti, E., and Fortino, G. (2019). *Big Social Data and Urban Computing*. Springer.

[Preece et al.2000] Preece, A., Hui, K., Gray, A., Marti, P., Bench-Capon, T., Jones, D., and Cui, Z. (2000). The kraft architecture for knowledge fusion and transformation. In *Research and Development in Intelligent Systems XVI*, pages 23–38. Springer.

[Rogova and Snidaro2019] Rogova, G. L. and Snidaro, L. (2019). Quality, context, and information fusion. In *Information Quality in Information Fusion and Decision Making*, pages 219–242. Springer.

[Santos et al.2019] Santos, I. L., Pirmez, L., Delicato, F. C., Oliveira, G. M., Farias, C. M., Khan, S. U., and Zomaya, A. Y. (2019). Zeus: A resource allocation algorithm for the cloud of sensors. *Future Generation Computer Systems*, 92:564–581.

[Suganuma et al.2018] Suganuma, T., Oide, T., Kitagami, S., Sugawara, K., and Shiratori, N. (2018). Multiagent-based flexible edge computing architecture for iot. *IEEE Network*, 32(1):16–23.

[Whitmore et al.2015] Whitmore, A., Agarwal, A., and Da Xu, L. (2015). The internet of things—a survey of topics and trends. *Information Systems Frontiers*, 17(2):261–274.

[Zame et al.2018] Zame, K. K., Brehm, C. A., Nitica, A. T., Richard, C. L., and Schweitzer III, G. D. (2018). Smart grid and energy storage: Policy recommendations. *Renewable and Sustainable Energy Reviews*, 82:1646–1654.