Análise de Técnicas de Similaridade Textual em Repositório Institucional de Produção Acadêmica

Edson Candido Rodrigues Filho, Rafael Divino Ferreira Feitosa

Instituto Federal Goiano – Campus Ceres (IFGoiano) – Ceres, GO – Brasil

edson.rodrigues@estudante.ifgoiano.edu.br,
rafael.feitosa@ifgoiano.edu.br

Abstract. *Facing the difficulty of finding works similar to the desired theme, this paper analyzes and compares text similarity techniques in the Institutional Repository of *****, focusing on two approaches: data compression similarity and clustering. The algorithms Damicore and K-Means were selected for the analysis. Data collection was performed using a web crawler, followed by the conversion of PDF documents to text. The results indicate that Damicore demonstrates superior efficiency in a qualitative approach, contributing to the organization and accessibility of data in the *****.*

Resumo. *Diante da dificuldade de encontrar trabalhos semelhantes ao tema desejado, este artigo analisa e compara técnicas de similaridade textual no Repositório Institucional do *****, focando em duas abordagens: similaridade por compressão de dados e por clusterização. Foram selecionados os algoritmos Damicore e K-Means para a análise. A coleta de dados foi realizada com um web crawler, seguida pela conversão de documentos PDF para texto. Os resultados indicam que o Damicore apresenta a melhor eficiência em uma abordagem qualitativa, contribuindo para a organização e acessibilidade dos dados no *****.*

1. Introdução

No ambiente acadêmico, muitas vezes nos deparamos com o desafio de escolher a direção a seguir no curso para criar um trabalho de conclusão que seja tanto interessante quanto relevante no campo escolhido. A escolha do assunto e o estabelecimento da metodologia são fases fundamentais que podem impactar de maneira significativa a qualidade do trabalho final (Gil, 2022). A opção do tema não apenas deve ressoar com os interesses pessoais do estudante, mas também considerar a relevância e a atualidade do assunto dentro da área de estudo, garantindo que a pesquisa contribua para o avanço do conhecimento. Depois de vencer essa fase inicial, ainda é necessário encontrar trabalhos relevantes que possam atuar como guias, mantendo a mesma linha de pesquisa e semelhança. Essa busca por referências adequadas é crucial, pois permite a contextualização da pesquisa dentro do panorama acadêmico existente, oferecendo um suporte teórico que enriquece a análise e fortalece as argumentações apresentadas.

No contexto de busca por trabalhos semelhantes, sistemas de recomendação são técnicas de software que fornecem sugestões de itens para os usuários. Eles podem também auxiliar em diversas decisões, como escolha de produtos, seleção de músicas e leitura de notícias. A recomendação é um atrativo importante para que os clientes retornem a utilizar os serviços (Su & Khoshgoftaar, 2009).

Em ambientes de busca, devido ao grande volume de informações, os sistemas

de recomendação podem ajudar a refinar os resultados e minimizar o tempo de busca, oferecendo um retorno de pesquisa de forma individualizada e ágil, apresentando informações realmente relevantes ao usuário (Adomavicius e Tuzhilin, 2020). Além disso, o aumento do volume de informações disponíveis na internet e em repositórios acadêmicos torna ainda mais crucial o aprimoramento de métodos eficientes de similaridade textual. Com o aumento da quantidade de documentos digitais, a habilidade de identificar e organizar conteúdos relacionados de forma eficaz não é apenas desejável, mas fundamental para a investigação acadêmica. Esta procura por técnicas que aprimorem a organização e a recuperação de informações tem estimulado o interesse em algoritmos e estratégias inovadoras que possam simplificar o acesso a informações relevantes em um oceano de dados.

O objetivo principal deste trabalho é analisar e comparar técnicas de similaridade textual aplicadas no Repositório Institucional do *****. Especificamente, investigaremos duas abordagens distintas: similaridade textual por compressão e similaridade textual por clusterização. A escolha da melhor técnica é fundamentada na obtenção de resultados relevantes e assertivos, visando aprimorar a organização e a acessibilidade dos dados no ***** e contribuir para a eficiência na recuperação de informações pelos usuários.

2. Fundamentação Teórica

Jurafsky *et al.(2024),* afirmam que o Processamento de Linguagem Natural (PLN) é fortemente influenciado pelo estudo da linguagem humana ao longo da história. No entanto, uma das principais dificuldades do PLN é lidar com as ambiguidades inerentes à linguagem natural, o que torna a tarefa de interpretação complexa. De acordo com Zavaglia (2003), a ambiguidade na linguagem natural é um fenômeno complexo que pode ocorrer em diferentes níveis, como fonético, morfológico, sintático, semântico, pragmático e de discurso. Esse aspecto torna a interpretação de textos uma tarefa desafiadora para os sistemas computacionais, que não possuem a mesma capacidade dos humanos de discernir contextos ambíguos (Zavaglia, 2003).

O PLN está intimamente relacionado às técnicas de similaridade textual, pois ambas envolvem a compreensão e manipulação de textos. Os vetores de palavras desempenham um papel crucial na determinação da similaridade textual utilizando métricas como o coeficiente de similaridade do cosseno e a distância Euclidiana. Essas técnicas são fundamentais para diversas aplicações em PLN, tais como busca semântica e detecção de plágio. A precisão e eficácia dessas aplicações dependem diretamente do desenvolvimento contínuo das técnicas de similaridade textual, as quais são constantemente aprimoradas com os avanços no campo do PLN Jurafsky *et al*. (2024).

As técnicas de similaridade textual desempenham um papel crucial em diversas tarefas como recuperação de informações, agrupamento de documentos, e sumarização de texto, entre outras. Conforme discutido por Gomaa e Fahmy (2013), a medição da similaridade entre palavras, frases, parágrafos e documentos é essencial para essas aplicações. Este estudo aborda diferentes técnicas para a similaridade textual, dividindo-as em três categorias principais: baseadas em string, baseadas em corpus e baseadas em conhecimento. Enquanto as abordagens baseadas em string operam diretamente sobre sequências de caracteres para calcular a similaridade, as baseadas em corpus utilizam informações de grandes coleções de textos para determinar a similaridade semântica, e as baseadas em conhecimento utilizam redes semânticas para avaliar a similaridade entre conceitos. A combinação dessas técnicas proporciona uma análise mais

abrangente e precisa da similaridade textual, beneficiando significativamente a interpretação e organização de informações em contextos científicos e práticos.

Foram selecionados dois algoritmos para aplicar a técnica de similaridade textual: Damicore (Sanches *et al.*, 2011) e K-Means, ambos pertencentes à abordagem baseada em corpus. Essa escolha se justifica pela natureza da ferramenta de estudo, que contém uma ampla coleção de textos, favorecendo a aplicação de algoritmos que utilizam o conteúdo do corpus para analisar e identificar similaridades entre os documentos.

O Damicore é uma metodologia avançada que se destaca na análise de similaridade textual, especialmente em tarefas de agrupamento e classificação. Conforme descrito por Medeiros Cesar (2016), a técnica Damicore envolve um conjunto de ferramentas que integram várias etapas para analisar dados textuais de maneira eficiente e eficaz. A metodologia é baseada no cálculo da distância de compressão normalizada (NCD), utilizando compressores como o PPMd, que mede a similaridade entre instâncias convertendo-as em strings binárias e avaliando a compressão resultante. A NCD é a medida da similaridade entre dois textos x e y com base na fórmula:

$$\text{NCD(x,y)} = \frac{C(xy) - min(C(x), C(y))}{max(C(x), C(y))}$$

onde $C(x)$ é o tamanho do texto x após a compressão, e $C(xy)$ é o tamanho dos textos x e y comprimidos após a concatenação. Quanto menor o valor da NCD, mais similares são os textos.

A importância do Damicore na similaridade textual reside em sua capacidade de operar de forma quase independente de configuração, facilitando a aplicação em diversos tipos de dados e contextos. Essa flexibilidade permite que pesquisadores e profissionais de diferentes áreas utilizem a metodologia para identificar padrões e agrupar dados textuais sem a necessidade de um profundo conhecimento em programação. Além disso, o Damicore oferece uma abordagem robusta para a classificação, permitindo a análise precisa e confiável de grandes volumes de dados textuais, o que é crucial para aplicações em aprendizado de máquina e mineração de dados (Medeiros Cesar, 2016).

O algoritmo K-Means é uma técnica de aprendizado não supervisionada amplamente utilizada para a clusterização de dados. Conforme discutido por Skinner (2019), o K-Means é essencial na formação de clusters, que agrupam objetos semelhantes com base em características específicas, facilitando a análise de grandes volumes de dados textuais. O processo iterativo do K-Means começa com a seleção inicial de k centróides, que representam os clusters. Em seguida, cada ponto de dados é atribuído ao centróide mais próximo, e os centróides são recalculados como a média dos pontos atribuídos a cada cluster. Este processo é repetido até que os centróides não mudem significativamente.

A importância do K-Means na similaridade textual reside na sua capacidade de agrupar documentos ou frases com base em suas características textuais, permitindo uma organização eficiente e a descoberta de padrões dentro dos dados. No contexto de sistemas de recomendação de textos acadêmicos, o K-Means pode ser utilizado para agrupar artigos semelhantes, facilitando a recomendação de leituras relevantes para os usuários (Skinner, 2019).

A implementação do K-Means iterativo, conforme detalhado no trabalho de

(Skinner, 2019), ajusta o número de clusters dinamicamente, começando com um único cluster e incrementando o valor de k até atingir uma configuração satisfatória. Esta abordagem elimina a dificuldade de determinar previamente o número ideal de clusters, resultando em grupos de tamanhos controlados e com pouca discrepância interna. Esse método é particularmente útil quando o número ótimo de clusters não é evidente, permitindo uma clusterização mais adaptativa e precisa dos dados textuais.

3. Materiais e Métodos

Para iniciar o processo de análise de similaridade textual, foi necessário coletar uma base de dados substancial a partir do *****. Esse processo de coleta foi realizado através de um web crawler que percorreu as páginas do repositório e extraiu os links para os documentos de interesse. A coleta de dados em bases científicas pode ser significativamente aprimorada com o uso de ferramentas automatizadas como web scrapers. Graciano e Ramalho (2023) apresentam o ScraperCI, um protótipo de web scraper, como uma solução eficiente para a automação da coleta de dados científicos, destacando o potencial de tais ferramentas na extração e gestão de grandes volumes de informação disponíveis na Web. A pesquisa enfatiza que o uso de scrapers pode aumentar a produtividade e favorecer a recuperação de dados em um contexto acadêmico.

Após baixar os arquivos em formato PDF, foi necessário convertê-los para o formato texto (.txt) e realizar uma limpeza nos dados extraídos. A conversão de arquivos PDF para texto é uma etapa de grande relevância em projetos de análise de dados que contêm grandes volumes de informações extraídas de documentos. De acordo com Lima (2018), a conversão de diferentes tipos de arquivos, como PDF, para o formato texto é fundamental para garantir que o conteúdo possa ser facilmente analisado.

O método utilizado foi um algoritmo em shell script que percorre o diretório atual, lendo cada arquivo PDF, convertendo-o para o formato texto (.txt) e, em seguida, realizando a limpeza dos dados ao remover caracteres especiais indesejados. Após a conversão e limpeza, o script organiza os arquivos, movendo os arquivos PDF para um diretório específico chamado filesPDF e os arquivos de texto limpos para outro diretório denominado fileTXT, garantindo, assim, a separação e organização dos diferentes tipos de arquivos.

Ao realizar a conversão e limpeza dos dados, foi utilizado o algoritmo Damicore para calcular a similaridade textual entre os documentos. Esta etapa envolveu a aplicação do algoritmo sobre a base de dados de arquivos em formato TXT, gerando um arquivo no formato ".phylip" que contém as informações de similaridade.

O arquivo .phylip gerado pelo Damicore foi utilizado como base para identificar as cinco maiores similaridades entre os documentos, já que o algoritmo Damicore forneceu tanto a árvore filogenética quanto a matriz de proximidade. A partir dessa matriz, foi realizado um processo que organiza os documentos de acordo com sua similaridade com outros. Esse processo considera as distâncias entre os documentos e os organiza de forma que os cinco mais próximos de cada um sejam destacados.

Após essa análise, foi gerado um script em SQL que insere essas informações em uma base de dados. O SQL gerado contém instruções para armazenar, para cada documento, a lista dos cinco mais próximos, identificando o documento alvo, o documento similar, o nível de proximidade (do mais próximo ao menos próximo) e o tipo de similaridade utilizada. Esses dados permitem a posterior recuperação e consulta

da relação de proximidade entre os documentos, facilitando análises mais aprofundadas com base nas relações textuais entre eles.

A análise de similaridade textual utilizando K-Means foi realizada a partir da base de dados limpa em formato TXT. O objetivo foi identificar os cinco documentos mais próximos para cada documento da base de dados, utilizando técnicas de aprendizado de máquina e processamento de linguagem natural. A partir dos arquivos TXT limpos, os documentos foram lidos e transformados em embeddings utilizando o modelo BERTopic, especializado para a língua portuguesa. Esses embeddings foram então convertidos em uma matriz 2D, que serviu de entrada para o algoritmo de clustering. Foi utilizado a pontuação do coeficiente de silhueta para encontrar o número ótimo de clusters. Este coeficiente mede a qualidade do clustering, com valores mais altos indicando uma melhor separação entre os clusters. Estas técnicas são detalhadas a seguir.

Embeddings são representações vetoriais de palavras, frases ou documentos que capturam o significado semântico do texto em um espaço multidimensional. Essa técnica é fundamental para o PLN, pois transforma informações textuais em formatos que podem ser usados por algoritmos de aprendizado de máquina, preservando relações semânticas entre palavras. No estudo de Santos (2022), os embeddings são utilizados pelo BERTopic para representar vetorialmente os termos, permitindo a identificação de tópicos relevantes dentro de um conjunto de dados textuais.

O BERTopic é uma ferramenta avançada para a extração de tópicos que combina embeddings com técnicas de clusterização. Essa abordagem permite agrupar documentos similares em clusters temáticos, facilitando a análise de grandes volumes de texto. Segundo Santos (2022), o BERTopic utiliza métodos como TF-IDF (Term Frequency-Inverse Document Frequency) para observar a relevância dos termos em cada cluster, além de proporcionar visualizações que ajudam na compreensão dos dados extraídos.

O coeficiente de silhueta é uma métrica amplamente utilizada para avaliar a qualidade de agrupamentos (clusters) em análise de clustering. Ele mede a eficiência da organização dos elementos dentro dos clusters, considerando a proximidade entre os dados do mesmo grupo e a separação em relação a outros clusters. Os valores do coeficiente variam entre -1 e 1, onde valores próximos de 1 indicam uma boa separação entre os clusters e maior homogeneidade interna. De acordo com Oliveira et al. (2020), essa métrica permite avaliar a qualidade do agrupamento após a formação dos clusters. Após aplicar o algoritmo K-Means aos embeddings e determinar o número ideal de clusters, utilizamos o coeficiente de silhueta para avaliar a qualidade dos agrupamentos formados. Os documentos são então atribuídos a diferentes clusters com base na similaridade de seus conteúdos.

Em seguida da identificação dos documentos mais próximos, o algoritmo gera um script SQL que insere essas informações em uma base de dados, facilitando a consulta e o armazenamento. Para cada documento, os cinco mais próximos são armazenados com a respectiva ordem de proximidade, tornando os dados organizados e acessíveis para futuras análises.

4. Resultados

Os resultados preliminares indicam que o Damicore se mostra superior ao K-Means na detecção de semelhanças textuais. Esta conclusão é fundamentada na avaliação

qualitativa dos agrupamentos, onde se constatou que o Damicore geralmente produz agrupamentos mais consistentes e pertinentes em contraste com os resultados obtidos pelo K-Means. A habilidade do Damicore de identificar nuances de similaridade parece ser uma vantagem significativa em relação à abordagem mais convencional do K-Means.

Para alcançar esse resultado inicial, efetuamos uma análise qualitativa dos agrupamentos formados por ambas as técnicas. Por meio de uma seleção por amostragem, analisamos os textos que foram apontados como similares. Embora ainda não tenha sido utilizada uma métrica quantitativa para essa comparação, a amostragem revelou que os textos agrupados pelo Damicore frequentemente apresentavam títulos e conteúdos mais próximos entre si. Essa observação sugere que o Damicore pode ser mais eficaz na captura de similaridades semânticas, uma área que merece investigação mais aprofundada.

Por exemplo, ao analisar os dados, notamos que em certas ocasiões o Damicore identificou cinco trabalhos com títulos parecidos, enquanto o K-Means conseguiu agrupar apenas dois ou até um trabalho parecido em certas situações. Esta variação nos resultados não só indica uma grande discrepância na habilidade de ambos os algoritmos em detectar similaridades, mas também ressalta a capacidade do Damicore de proporcionar agrupamentos mais pertinentes e coerentes. A avaliação qualitativa dos textos reunidos pelo Damicore revelou que eles não só possuíam palavras-chave em comum, mas também tratavam de temas e conceitos parecidos, sugerindo uma conexão semântica mais profunda entre os documentos.

Por outro lado, os achados do K-Means frequentemente levaram a agrupamentos que continham textos que, mesmo com títulos parecidos, não evidenciavam uma conexão evidente em relação ao conteúdo ou ao contexto. Essa restrição pode ser creditada à característica do algoritmo K-Means, que costuma ser mais sensível a ruídos e outliers, gerando agrupamentos que, apesar de estatisticamente válidos, podem não espelhar a real similaridade dos textos.

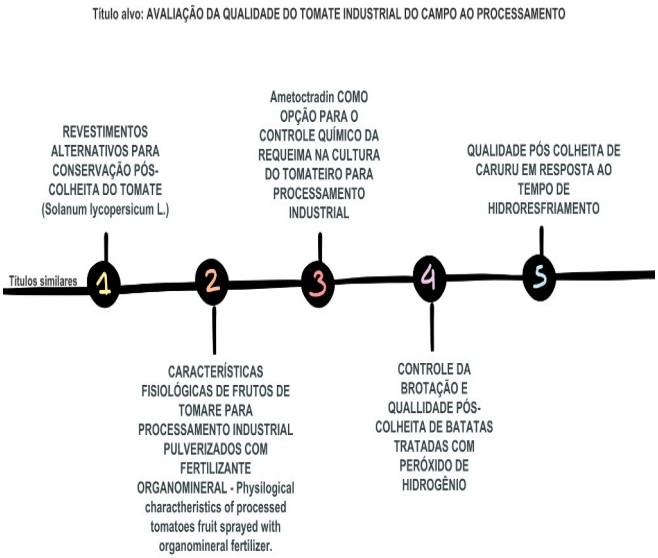

Figura 1. Exemplo de um bom caso com uso do Damicore.

Fonte: Elaboração própria (2024).

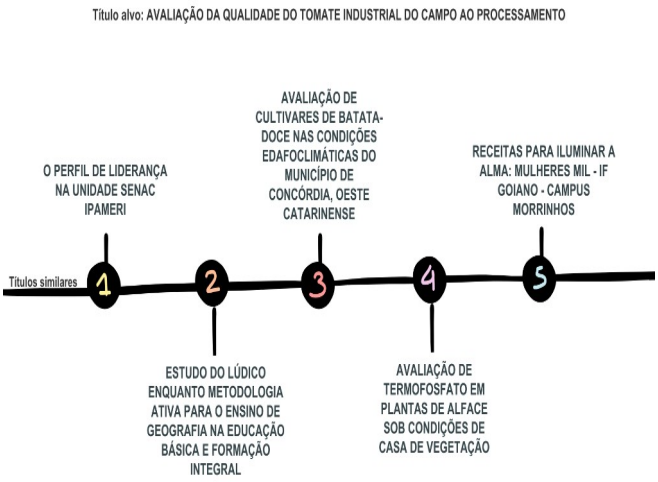

Figura 2. Exemplo de um caso ruim com uso do K-menas.

Fonte: Elaboração própria (2024).

No exemplo ilustrado na Figura 1, o Damicore retornou cinco textos semelhantes, com a maioria diretamente relacionados ao tema da avaliação e qualidade do tomate, sendo o primeiro resultado particularmente relevante, abordando "Revestimentos Alternativos para Conservação Pós-Colheita do Tomate (Solanum lycopersicum L.)". Por outro lado, no exemplo ilustrado na Figura 2, o K-means produziu resultados que, embora ainda relacionados, não foram tão focados no tema central, como evidenciado pelo primeiro resultado, que abordava "O Perfil de Liderança na Unidade Senac Ipameri".

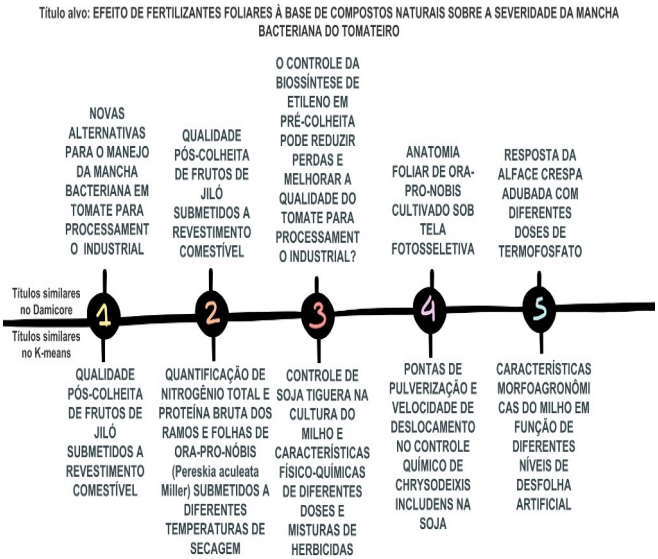

Figura 3. Comparação dos títulos semelhantes identificados pelos algoritmos Damicore e K-Means.Fonte:

Elaboração própria (2024).

No entanto, vale ressaltar que o K-Means também apresentou resultados significativos ao agrupar textos relevantes. Um exemplo notável é o trabalho intitulado “Qualidade Pós-colheita de Frutos de Jiló Submetidos a Revestimento Comestível”, que se destacou tanto nos agrupamentos gerados pelo Damicore quanto pelo K-Means, conforme ilustrado na Figura 3.

5. Conclusão

Ao alcançar este estágio do estudo, em que identificamos os textos mais parecidos através das duas técnicas de similaridade textual, o próximo passo é aprofundar a análise das métricas de agrupamento. A escolha dessas métricas é crucial, pois pode influenciar, significativamente, não apenas os resultados alcançados, mas também a interpretação e o entendimento dos dados. Diversas métricas podem desvendar diferentes aspectos da similaridade, e determinar qual se ajusta melhor aos nossos objetivos será crucial para a efetividade da pesquisa.

Além disso, é vital discutir os sucessos e desafios encontrados nas abordagens de similaridades analisadas. Esta avaliação crítica não só nos auxiliará a identificar as restrições de cada método, mas também possibilitará reflexões sobre possíveis melhorias para estudos futuros. Buscamos, assim, estabelecer uma compreensão mais sólida sobre as aplicações e restrições das técnicas de similaridade textual, contribuindo para o avanço do conhecimento na área e para a melhoria da organização e acessibilidade de dados em repositórios acadêmicos.

Apesar de a pesquisa ainda estar em andamento, as etapas seguintes são cruciais para atingir um resultado final que não apenas determine a técnica mais eficiente, mas também proporcione percepções úteis para a implementação prática dessas metodologias no contexto acadêmico. A expectativa é que as conclusões alcançadas possam fundamentar futuras pesquisas e auxiliar no progresso das técnicas de análise de similaridade textual em ambientes acadêmicos e científicos.

References

ADOMAVICIUS, G.; TUZHILIN, A. Context-Aware Recommender Systems. In: RICCI, F. *et al*. (Eds.). Recommender Systems Handbook. 2nd ed. New York: Springer, 2015. p. 217-253.

ARAÚJO DOS SANTOS, Morgana. Um estudo sobre a repercussão da eleição presidencial brasileira de 2022 no Twitter usando BERTopic. 2022. Trabalho de Conclusão de Curso (Graduação em Sistemas e Mídias Digitais) – Universidade Federal do Ceará, Fortaleza, 2022.

CILIBRASI, R.; VITANYI, P. Clustering by compression. IEEE Transactions on Information Theory, v. 51, n. 4, p. 1523-1545, 2005.

GIL, A. C. Métodos e técnicas de pesquisa social. 7. ed. São Paulo: Atlas, 2022.

GOMAA, W. H.; FAHMY, A. A. A Survey of Text Similarity Approaches. International Journal of Computer Applications, v. 68, n. 13, 2013.

GRACIANO, Helton Luiz dos Santos; RAMALHO, Rogério Aparecido Sá. SCRAPERCI: Um web scraper para coleta de dados científicos. Encontros Bibli, Florianópolis, v. 28, 2023.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

3rd ed. Upper Saddle River: Prentice Hall, 2024.

LIMA, Rui José da Rocha. Extração e análise multidimensional de dados de atletismo a partir de dados não estruturados. 2018. Dissertação (Mestrado em Engenharia de Software) – Universidade de Trás-os-Montes e Alto Douro, Vila Real, 2018.

MEDEIROS CESAR, Bruno Kim. Estudo e extensão da metodologia Damicore para tarefas de classificação. 2016. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

OLIVEIRA, Fernanda Robes de; KLEINA, Mariana; MARQUES, Marcos Augusto Mendes; GAYER, Jessika Alvares Coppi Arruda; TAMACHIRO, Thiago Shoji Obi. Clusterização de Clientes: um Modelo Utilizando Variáveis Categóricas e Numéricas. 2020.

SANCHES, Adriano; CARDOSO, Joao M. P.; DELBEM, Alexandre C. B. Identifying merge-beneficial software kernels for hardware implementation. In: 2011 International Conference on Reconfigurable Computing and FPGAs. 2011. DOI: 10.1109/ReConFig.2011.51.

SKINNER, Rafael de Araujo. Sistema de recomendação de textos acadêmicos através de clusterização com K-Means iterativo. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade Federal Fluminense, Niterói, 2019.

SU, X.; KHOSHGOFTAAR, T. M. A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009.

ZAVAGLIA, C. Ambigüidade gerada pela homonímia: Revisitação teórica, linhas limítrofes com a polissemia e proposta de critérios distintivos. D.E.L.T.A., v. 19, n. 2, p. 237-266, 2003.