INSTITUTO FEDERAL GOIANO – CAMPUS CERES
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
EDSON CANDIDO RODRIGUES FILHO

ANÁLISE DE TÉCNICAS DE SIMILARIDADE TEXTUAL NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO (RIIF Goiano)

CERES – GO
2025

EDSON CANDIDO RODRIGUES FILHO

ANÁLISE DE TÉCNICAS DE SIMILARIDADE TEXTUAL NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO (RIIF Goiano)

Trabalho de conclusão de curso apresentado ao curso de Bacharelado em Sistemas de Informação do Campus Ceres do Instituto Federal Goiano, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação sob orientação do Prof. Dr. Rafael Divino Ferreira Feitosa.

CERES – GO
2024

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema Integrado de Bibliotecas do IF Goiano - SIBi

Candido Rodrigues Filho, Edson
R696 ANÁLISE DE TÉCNICAS DE SIMILARIDADE TEXTUAL NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO (RIIF Goiano) / Edson Candido Rodrigues Filho. Uruana 2025.

16f. il.

Orientador: Prof. Dr. Rafael Divino Ferreira Feitosa.
Tcc (Bacharel) - Instituto Federal Goiano - Campus Ceres, curso de 0320203 - Bacharelado em Sistemas de Informação - Ceres (Campus Ceres).
I. Título.

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
- ☐ Dissertação (mestrado)
- ☐ Monografia (especialização)
- ☑ TCC (graduação)
- ☐ Artigo científico
- ☐ Capítulo de livro
- ☐ Livro
- ☐ Trabalho apresentado em evento
- ☐ Produto técnico e educacional - Tipo:

Nome completo do autor: Edson Candido Rodrigues Filho

Matrícula: 2018103202030010

Título do trabalho: ANÁLISE DE TÉCNICAS DE SIMILARIDADE TEXTUAL NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO (RIIF Goiano)

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☑ Não ☐ Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano: 13 /06 /2025

O documento está sujeito a registro de patente? ☐ Sim ☑ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☑ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

• Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;

• Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

• Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Ceres-GO (Local) 11 /06 /2025 (Data)

gov.br Documento assinado digitalmente
EDSON CANDIDO RODRIGUES FILHO
Data: 11/06/2025 00:34:16-0300
Verifique em https://validar.iti.gov.br

Assinatura do autor e/ou detentor dos direitos autorais

Ciente e de acordo:

gov.br Documento assinado digitalmente
RAFAEL DIVINO FERREIRA FEITOSA
Data: 12/06/2025 00:30:05-0300
Verifique em https://validar.iti.gov.br

Assinatura do(a) orientador(a)

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

ATA DE DEFESA DE TRABALHO DE CURSO

Ao(s) 5 dia(s) do mês de junho do ano de dois mil e vinte e cinco, realizou-se a defesa de Trabalho de Curso do(a) acadêmico(a) Edson Candido Rodrigues Filho, do Curso de Bacharelado em Sistemas de Informação, matrícula 2018103202030010, cujo título é "Análise de Técnicas de Similaridade Textual no Repositório Institucional do IF Goiano (RIIF Goiano)". A defesa iniciou-se às 20 horas e 56 minutos, finalizando-se às 21 horas e 44 minutos. A banca examinadora considerou o trabalho **APROVADO** com média 8,0 no trabalho escrito, média 6,7 no trabalho oral, apresentando assim média aritmética final de **7,3** pontos, estando o(a) estudante **APTO** para fins de conclusão do Trabalho de Curso.

Após atender às considerações da banca e respeitando o prazo disposto em calendário acadêmico, o(a) estudante deverá fazer a submissão da versão corrigida em formato digital (.pdf) no Repositório Institucional do IF Goiano – RIIF, acompanhado do Termo Ciência e Autorização Eletrônico (TCAE), devidamente assinado pelo autor e orientador.

Os integrantes da banca examinadora assinam a presente.

(Assinado Eletronicamente)
Prof. Dr. Rafael Divino Ferreira Feitosa
Orientador

(Assinado Eletronicamente)
Prof. Dr. Vilson Soares de Siqueira
Membro

(Assinado Eletronicamente)
Prof. Esp. Paulo Henrique Rodrigues Araujo
Membro

Documento assinado eletronicamente por:

- **Rafael Divino Ferreira Feitosa**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 05/06/2025 21:46:07.
- **Paulo Henrique Rodrigues Araujo**, **PROF ENS BAS TEC TECNOLOGICO-SUBSTITUTO** , em 05/06/2025 21:49:52.
- **Vilson Soares de Siqueira**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 05/06/2025 22:02:45.

Este documento foi emitido pelo SUAP em 05/06/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 714220
Código de Autenticação: 1419946614

Dedico este trabalho a todos que contribuíram e me apoiaram para a sua realização.

AGRADECIMENTOS

Agradeço a todas as influências, diretas e indiretas, que me ajudaram a evoluir durante esse percurso. O conhecimento construído, os momentos de dedicação e os aprendizados adquiridos são reflexos do esforço contínuo e da busca por aperfeiçoamento.

"Eu cheguei aqui e guiarei o caminho até os mais longínquos confins do mundo"

Gol D. Roger, One Piece

RESUMO

Os algoritmos de sugestão tornaram-se essenciais para simplificar o acesso eficaz ao saber. Esses algoritmos são frequentemente empregados em sistemas de sugestão, mecanismos de pesquisa e plataformas de conteúdo acadêmico, auxiliando na customização da experiência do usuário e na identificação de materiais pertinentes. No âmbito educacional e científico, sua implementação pode facilitar o acesso a estudos similares, incentivando a propagação do saber e o progresso das pesquisas. Diante da dificuldade de encontrar trabalhos semelhantes ao tema desejado, este artigo analisa e compara técnicas de similaridade textual no Repositório Institucional do IF Goiano (RIIF GOIANO), 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 RIIF GOIANO.

Palavras-chave: Similaridade Textual. Repositório Acadêmico. Processamento de Linguagem Natural. Algoritmos de Clusterização. Damicore. K-Means. Recuperação de Informação. Coeficiente de Silhueta. Mineração de Texto.

ABSTRACT

Suggestion algorithms have become essential for simplifying effective access to knowledge. These algorithms are frequently employed in recommendation systems, search engines, and academic content platforms, assisting in the customization of the user experience and the identification of relevant materials. In educational and scientific contexts, their implementation can facilitate access to similar studies, encouraging the dissemination of knowledge and the advancement of research. Facing the difficulty of finding works similar to the desired theme, this paper analyzes and compares text similarity techniques in the Institutional Repository of IF Goiano (RIIF GOIANO), 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 RIIF GOIANO.

Keywords: Text Similarity. Academic Repository. Natural Language Processing. Clustering Algorithms. Damicore. K-Means. Information Retrieval. Silhouette Coefficient. Text Mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

No ambiente acadêmico, é comum 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 de estudo escolhido. A escolha do tema e o estabelecimento da metodologia são fases fundamentais, podendo impactar de maneira significativa a qualidade do trabalho final (Gil, 2022). A definição do tema deve não apenas deve ressoar com os interesses pessoais do estudante, mas também considerar sua relevância e a atualidade do assunto dentro da área de estudo, de modo a garantir que a pesquisa contribua efetivamente para o avanço do conhecimento.

Depois de vencer essa fase inicial, ainda é necessário encontrar trabalhos relevantes que possam atuar como referência, mantendo uma linha de pesquisa semelhante. 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.

Nesse contexto, os sistemas de recomendação são técnicas de software que fornecem sugestões automatizadas de itens para os usuários. Eles podem ser aplicados em diversas decisões, como escolha de produtos, músicas, notícias e, no caso deste estudo, na localização de trabalhos acadêmicos. 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 disponível, 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. A capacidade 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.

Diante desse cenário, o objetivo principal deste trabalho é analisar e comparar técnicas de similaridade textual aplicadas ao Repositório Institucional do IF Goiano (RIIF GOIANO). Especificamente, são investigadas duas abordagens distintas: similaridade textual por compressão e similaridade textual por clusterização. A escolha da técnica mais adequada é fundamentada na obtenção de resultados relevantes e assertivos, visando aprimorar a organização e a acessibilidade dos dados no RIIF GOIANO e, assim, contribuir para a eficiência na recuperação de informações pelos usuários.

Optamos pelo DAMICORE e o K-Means devido à sua implementação simples, à sua robustez em ambientes de recursos computacionais moderados e à facilidade de interpretação dos resultados sem grandes aportes em infraestrutura. Essa escolha também pode servir como base para futuras análises voltadas à possível implementação dessas técnicas no repositório acadêmico.

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 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 essencial em tarefas como recuperação de informações, agrupamento de documentos e sumarização de texto. 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, em corpus e em conhecimento. Enquanto as abordagens baseadas em string operam diretamente sobre sequências de caracteres, as baseadas em corpus utilizam 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 a interpretação e organização de informações em contextos científicos e práticos.

Neste trabalho, foram selecionados dois algoritmos para aplicação de técnicas 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 textual 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), o DAMICORE integra várias etapas para analisar dados textuais de forma eficiente. 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 por meio da compressão resultante. A NCD é a medida da similaridade entre dois textos x e y com base na fórmula:

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

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, maior a similaridade entre os textos.

A relevância do DAMICORE reside em sua capacidade de operar de forma quase independente de configuração, o que facilita sua 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 conhecimento técnico aprofundado. 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).

A clusterização textual é uma abordagem fundamental em PLN e recuperação de informação, permitindo a organização de documentos semelhantes em grupos (clusters) com base em características compartilhadas. Manning et al. (2009) explicam que a similaridade entre documentos pode ser medida através de métodos de clusterização, nos quais documentos que possuem proximidade semântica ou temática são agrupados. Esse processo facilita a navegação em grandes volumes de dados e é essencial para sistemas de recuperação eficientes.

O algoritmo k-Means é uma técnica de aprendizado não supervisionado 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. O processo 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 se estabilizam.

A clusterização por K-means é um dos algoritmos mais importantes de clusterização plana. Seu objetivo principal é minimizar a distância euclidiana quadrática média entre os documentos e seus centros de cluster, onde o centro de cluster é definido como a média ou centróide $\tilde{\mu}$ dos documentos em um cluster ω:

$$\tilde{\mu}(\omega) = \frac{1}{|\omega|}\Sigma_{x\epsilon\omega}x$$

Conforme descrito por Manning et al. (2009), este algoritmo assume que os documentos são representados como vetores normalizados por comprimento em um espaço de valores reais.

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 nos dados. Em 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.

MATERIAL E MÉTODOS

Para iniciar o processo de análise de similaridade textual, foi necessário coletar uma base de dados substancial a partir do RIIF GOIANO. 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.

Enquanto o coeficiente de silhueta avalia a qualidade dos agrupamentos considerando a separação entre clusters, a medida G oferece uma perspectiva complementar ao quantificar a coerência interna das associações dentro de uma estrutura hierárquica. Essa métrica calcula a capacidade de generalização das árvores filogenéticas geradas, assumindo que cada amostra deve estar diretamente vinculada ao seu vizinho mais próximo no dendrograma. Segundo Feitosa (2020), a medida G pode ser interpretada como um indicador da acurácia dos agrupamentos, permitindo avaliar a eficácia do modelo na identificação de padrões e relações semânticas. Já Bailão (2020) destaca seu papel como uma métrica quantitativa de robustez para a comparação de agrupamentos hierárquicos, baseada no índice de congruência de par a par entre objetos.

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.

RESULTADOS E DISCUSSÃO

Os achados iniciais sugerem que o Damicore é um pouco mais eficaz que o K-Means na identificação de similaridades 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 evidenciaram 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.

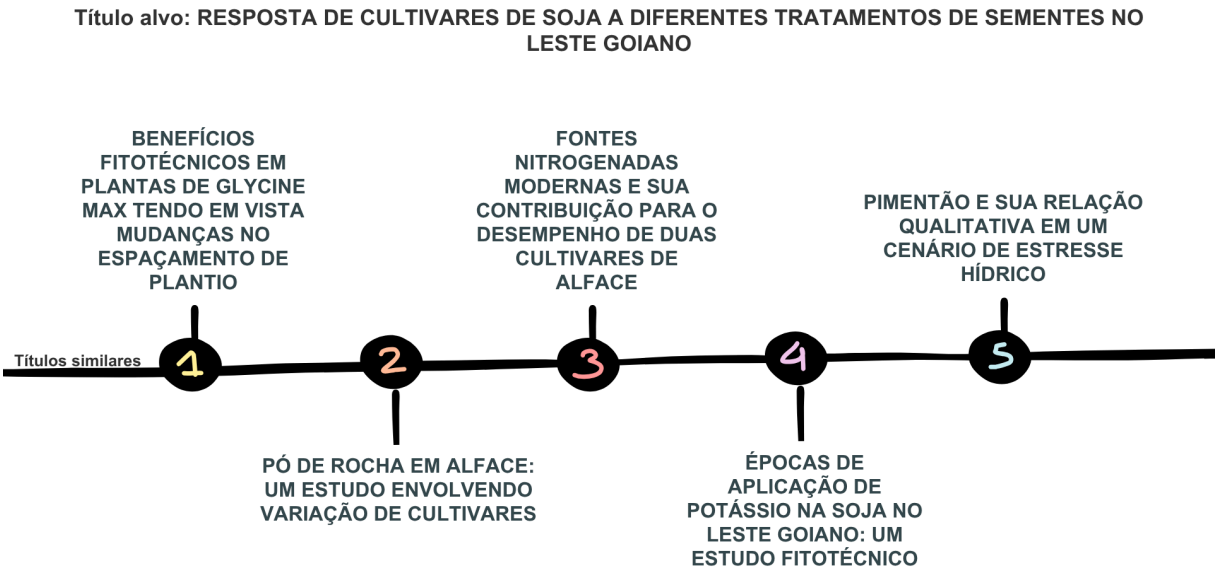

Figura 1. Exemplo de um bom caso com uso do Damicore.

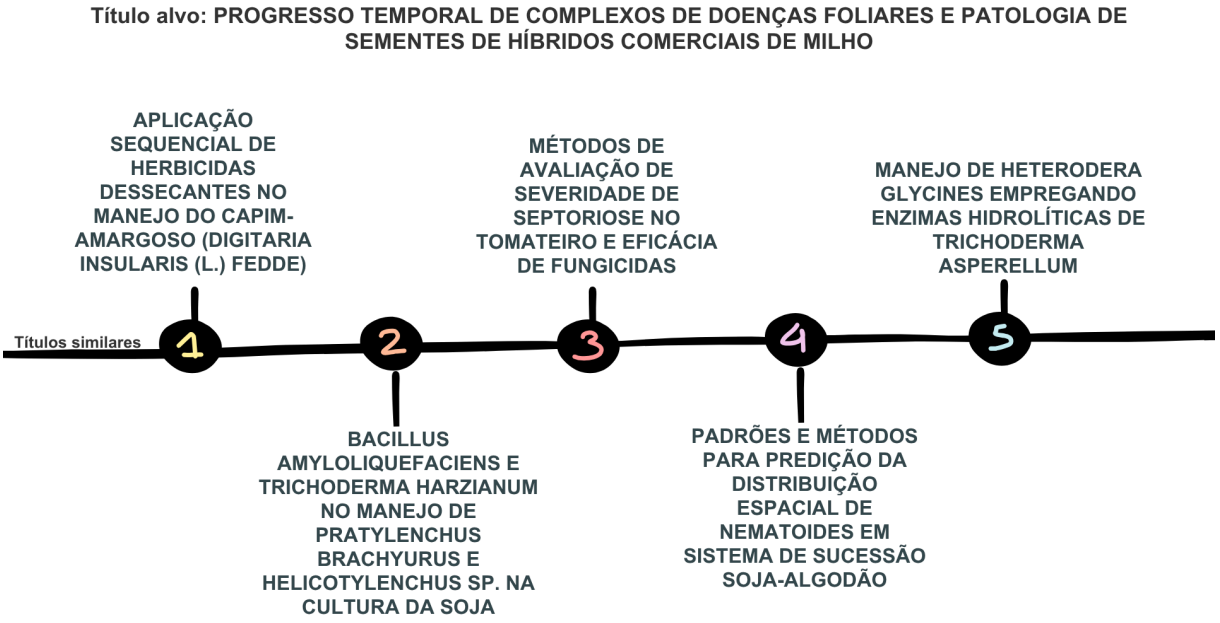

Figura 2. Exemplo de um caso ruim com uso do K-menas.

No exemplo ilustrado na Figura 1, o Damicore retornou cinco textos semelhantes, com a maioria diretamente relacionados ao tema do cultivo da soja e produtividade, sendo um desses resultados diretamente ligado à mesma cultura agrícola e região geográfica. Por outro lado, no exemplo ilustrado na Figura 2, o K-means produziu resultados que, embora ainda relacionados a tratar de pragas e doenças, não foram diretamente ligados ao milho.

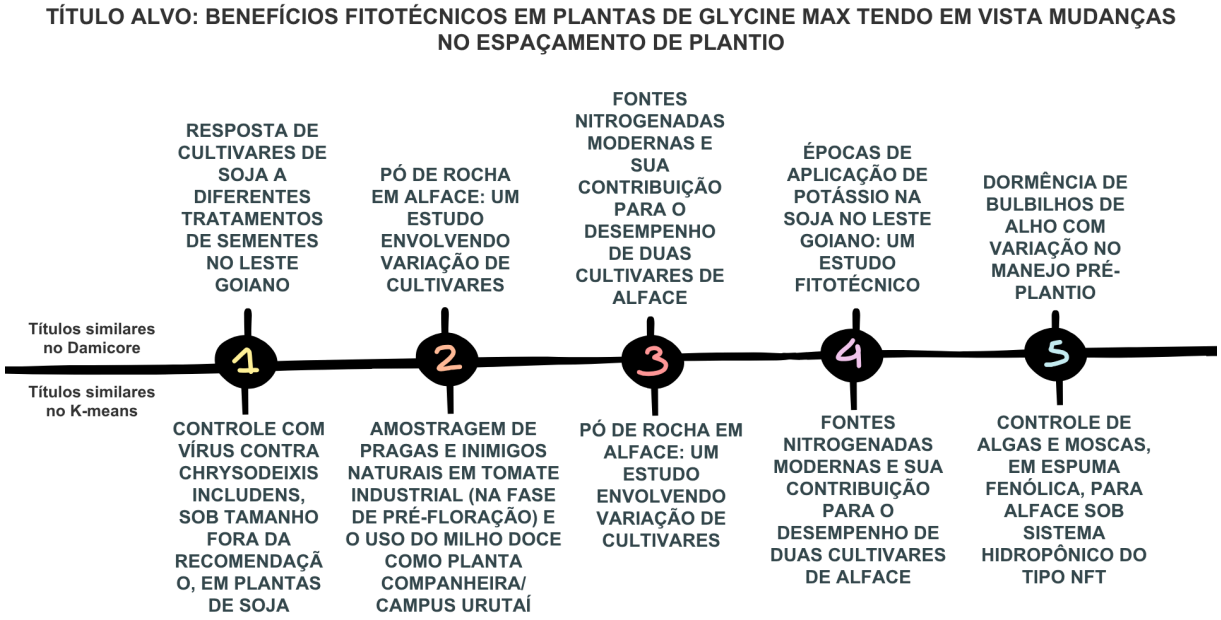

Figura 3. Comparação dos títulos semelhantes identificados pelos algoritmos Damicore e K-Means.

O K-Means por mais que apresentou resultados significativos ao agrupar textos relevantes na amostragem feita, o Damicore se mostrou um pouco superior, apresentando maior quantidade de resultado semelhante ao trabalho alvo. Um exemplo notável é o trabalho intitulado 'PÓ DE ROCHA EM ALFACE: UM ESTUDO ENVOLVENDO VARIAÇÃO DE CULTIVARES', que se destacou tanto nos agrupamentos gerados pelo Damicore quanto pelo K-Means, conforme ilustrado na Figura 3.

Além da avaliação qualitativa das categorias, utilizamos o coeficiente de silhueta como uma métrica quantitativa para avaliar a diferenciação dos agrupamentos. O desempenho mais eficiente obtido no K-Means teve um valor de 0,7077, indicando uma distinção moderada entre os conjuntos. O coeficiente de silhueta varia de -1 a +1; valores próximos a +1 indicam uma definição clara dos agrupamentos, enquanto valores próximos a 0 indicam uma sobreposição significativa entre eles. Em relação ao K-Means, o coeficiente alcançado indica que os documentos foram agrupados com boa separação entre os clusters, preservando uma estrutura coesa e representativa das similaridades observadas no conjunto textual. Esses resultados reforçam que o K-Means é uma técnica eficiente para segmentar textos com base em similaridade, obtendo uma separação clara entre agrupamentos, conforme demonstrado pelo coeficiente de Silhueta. Já o DAMICORE, evidenciou limitações significativas na separação efetiva dos grupos, com um coeficiente de Silhueta de apenas 0,0108.

Contudo, ao levarmos em conta a métrica G, que avalia a coerência entre elementos adjacentes com o mesmo rótulo, notamos uma superioridade evidente do K-Means. A métrica G avalia a congruência entre os documentos agrupados, sugerindo que valores próximos 1 indicam uma alta consistência semântica entre os agrupamentos (FEITOSA, 2020). Usando essa métrica, alcançou-se 0,9929 com o K-Means, mostrando que os agrupamentos mantiveram uma sequência semanticamente consistente. Por outro lado, o DAMICORE obteve apenas 0,0427 na métrica G, sugerindo que, mesmo com a proposta do método, a similaridade semântica não foi mantida de forma significativa ao longo da árvore.

A distância média global, que avalia a dispersão geral dos dados no espaço

de representação, apresentou uma diferença significativa para o K-Means, que atingiu 9,2851, em contraste com o DAMICORE, que atingiu 0,9687. Isso sugere que, apesar dos agrupamentos criados pelo K-Means terem uma boa distinção entre grupos, como demonstrado pela silhueta elevada, os dados estão mais dispersos globalmente, o que sugere uma maior variabilidade interna. Por outro lado, o DAMICORE produz agrupamentos mais nítidos no espaço vetorial, apesar de apresentar menor consistência semântica, conforme evidenciado pela medida G.

Técnica	Coeficiente de Silhueta	Medida G	Distância Média Global
K-Means	0,7077	0,9929	9,2851
Damicore	0,0108	0,0427	0,9687

Tabela 1. Comparação da Medida G e Silhueta entre os algoritmos Damicore e K-Means.

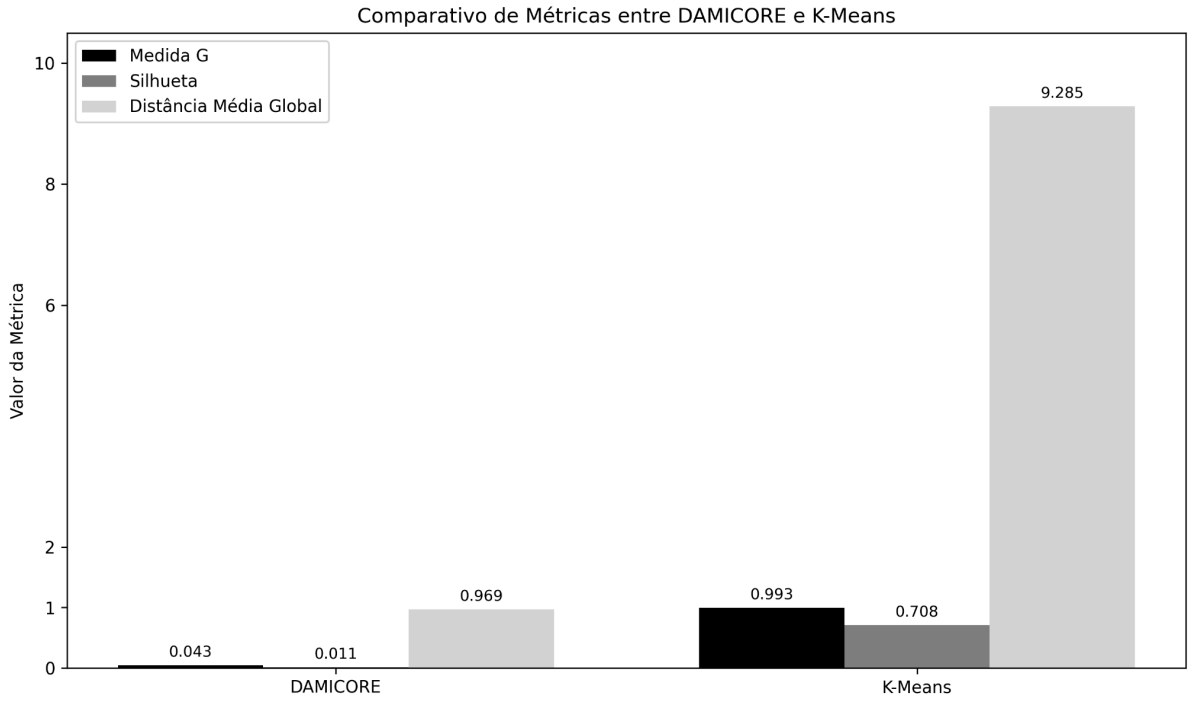

Figura 4. Comparação da Medida G e Silhueta entre os algoritmos Damicore e K-Means.

A Tabela 1 junta os achados das métricas quantitativas de avaliação, possibilitando uma comparação direta entre os métodos DAMICORE e K-Means. O K-Means demonstrou um desempenho superior em quase todas as métricas avaliadas, destacando-se a Medida G de 0,9929, que demonstra uma forte

coerência entre elementos adjacentes com o mesmo rótulo, e um coeficiente de silhueta de 0,7077, que evidencia agrupamentos claramente delimitados e distintos. Por outro lado, o DAMICORE exibiu valores significativamente inferiores, com Medida G de 0,0427 e silhueta de 0,0108, indicando uma fraca coesão interna e uma separação insuficiente entre os dois grupos. O K-Means indicou uma distância média global de 9,2851, ao passo que o DAMICORE mostrou um valor consideravelmente inferior, 0,9687. Isso sugere que as agrupações formadas pelo DAMICORE são mais sólidas e consistentes, espelhando uma estrutura semântica mais eficiente. Por outro lado, a maior dispersão no K-Means indica uma coesão interna dos grupos reduzida, mesmo com a aparente distinção semântica.

CONSIDERAÇÕES FINAIS

Ao alcançar este estágio do estudo, em que identificamos os artigos 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.

A incorporação das métricas qualitativas e quantitativas permitiu uma análise mais completa da similaridade textual, possibilitando uma avaliação mais acurada da qualidade dos agrupamentos produzidos. A análise comparativa entre a métrica G e o coeficiente de silhueta proporcionou uma perspectiva adicional sobre a estruturação e distinção de agrupamentos, destacando os benefícios de métodos fundamentados em similaridade semântica. Isso nos permite avaliar os resultados favoráveis do K-means em relação a métricas quantitativas, como o coeficiente de

silhueta e a medida G, que se destacaram. Em contrapartida, o Damicore mostrou-se positivo na avaliação qualitativa e na distância média global.

Em termos práticos, isso implica que o K-means é capaz de criar agrupamentos com estruturas matemáticas claramente definidas. Os textos dentro de cada grupo são bem "agrupados" e bem distintos dos demais. Isso é evidenciado pela silhueta alta, que sugere uma proximidade maior de cada texto com seu próprio grupo do que com os demais, e pela medida G elevada, que evidencia uma progressão consistente que reforça a disposição dos agrupamentos. Já o Damicore, ao gerar uma distância média global reduzida, indica que, em média, os documentos mais parecidos ficaram mais próximos, mesmo que o agrupamento em si não seja tão evidente quanto o do K-means. Esta "proximidade mais estreita" espelha a avaliação qualitativa: ao examinarmos manualmente os pares escolhidos pelo Damicore, notamos que ele costuma selecionar textos com maior afinidade de tema ou contexto, captando sutilezas semânticas que podem não ser percebidas por uma mera contagem de distâncias numéricas.

Em resumo, o K-means estrutura os dados de maneira mais consistente e distinta, seguindo critérios estritamente matemáticos, enquanto o Damicore, apesar de criar grupos um pouco mais desordenados, reúne os textos que realmente possuem maior relevância do ponto de vista do conteúdo. Esta correlação indica que, dependendo do propósito - seja aumentar a coesão interna do agrupamento (K-means) ou dar prioridade à afinidade semântica real entre os documentos (Damicore) - cada técnica pode se sobressair.

Esses achados não só apontaram a metodologia mais eficaz para o cenário estudado, como também fornecem subsídios valiosos para a aplicação prática dessas metodologias no ambiente acadêmico. Essas conclusões são esperadas para auxiliar estudos futuros e impulsionar o progresso das técnicas de análise de similaridade textual em repositórios de pesquisa científica.

REFERÊNCIAS

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.

BAILÃO, Adriano Soares de Oliveira. *Reconhecimento de padrões por processos adaptativos de compressão*. 2020. 160 f. Tese (Doutorado em Ciência da Computação) – Instituto de Informática, Universidade Federal de Goiás, Goiânia, 2020.

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

FEITOSA, Rafael Divino Ferreira. Classificação de cenas utilizando compressão de dados. 2020. Tese (Doutorado) – Universidade Federal de Goiás, Goiânia, 2020.

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

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2009.

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.s