UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

Bacharelado em Sistemas de Informação

Érica Silva Rodrigues

DESCOBERTA DE CONHECIMENTO EM SITES DE TECNOLOGIA POR MEIO DE MINERAÇÃO DE TEXTOS E CLUSTERIZAÇÃO

Diamantina - MG

2025

Érica Silva Rodrigues

DESCOBERTA DE CONHECIMENTO EM SITES DE TECNOLOGIA POR MEIO DE MINERAÇÃO DE TEXTOS E CLUSTERIZAÇÃO

Trabalho de Conclusão de Curso apresentado à Faculdade de Ciências Exatas da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como parte dos requisitos para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Marcus Vinícius Carvalho Guelpeli.

Diamantina - MG

2025

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por Sua graça e fidelidade, que me sustentaram e permitiram a conclusão desta jornada. Expresso minha sincera gratidão aos meus pais, Sandra Rodrigues e Santos Quirino, que constituíram a base de apoio e sustentação em todos os momentos. Aos meus irmãos, Tiago e Diogo, pelo incentivo constante, e ao meu marido, Lucas Kaique, por estar sempre ao meu lado, oferecendo apoio e compreensão. Agradeço ao meu orientador, Marcus Vinicius Guelpeli, pela paciência, dedicação e valiosas contribuições ao longo do desenvolvimento deste trabalho, as quais foram essenciais para meu aprendizado e crescimento acadêmico. Estendo meus agradecimentos a toda a minha família e à irmandade da igreja de Diamantina, cujo apoio foi de grande importância para a concretização deste trabalho. Agradeço, ainda, aos amigos da faculdade, que fizeram parte dessa trajetória e contribuíram de maneira significativa para a realização desta etapa

RESUMO

O crescimento acelerado das publicações em sites de tecnologia torna cada vez mais difícil compreender quais temas se destacam nesses ambientes digitais e como os conteúdos se relacionam entre si. Diante desse cenário, surge a necessidade de métodos capazes de organizar grandes volumes de documentos e revelar estruturas semânticas que não são facilmente percebidas pela análise manual. Este estudo teve como objetivo descobrir conhecimento em textos publicados em portais de tecnologia, empregando técnicas de corpusmineração de textos e métodos de clusterização para analisar a distribuição e a organização temática de um *corpus* composto por 1.864 textos coletados de cinco sites especializados no período de 2023 a 2025. Para isso, foram realizadas simulações no Cassiopeia, utilizando as métricas de Coesão e Acoplamento para avaliar a qualidade dos agrupamentos formados. Os resultados obtidos indicam a presença de similaridades significativas entre os documentos, permitindo observar como os textos se distribuem entre diferentes temas e categorias. A análise evidencia o potencial das técnicas empregadas para apoiar a organização e interpretação de dados textuais, contribuindo para estudos voltados à compreensão de padrões informacionais em domínios especializados.

Palavras- Chaves: mineração de textos; *clustering*; cassiopeia; coesão e acoplamento; tecnologia;

ABSTRACT

The rapid growth of publications on technology websites makes it increasingly difficult to identify which themes stand out in these digital environments and how the content relates to each other. In this context, there is a growing need for methods capable of organizing large volumes of documents and revealing semantic structures that are not easily detected through manual analysis. This study aimed to discover knowledge in texts published on technology portals by employing text mining techniques and clustering methods to analyze the thematic distribution and organization of a *corpus* composed of 1,864 texts collected from five specialized websites between 2023 and 2025. To achieve this, simulations were conducted in Cassiopeia, using Cohesion and Coupling metrics to evaluate the quality of the formed *clusters*. The results indicate significant similarities among the documents, providing *insights* into how the texts are distributed across different themes and categories. The analysis highlights the potential of the employed techniques to support the organization and interpretation of textual data, contributing to studies focused on understanding informational patterns in specialized domains.

Keywords: Text mining, *clustering,* Cassiopeia, cohesion and coupling, technology.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A quantidade de notícias publicadas diariamente em sites de tecnologia cresce em ritmo acelerado, tornando cada vez mais difícil acompanhar e compreender tudo o que é produzido. Nesse contexto, a mineração de textos e, em especial, a técnica de clusterização, surge como uma abordagem eficiente para organizar grandes volumes de textos e revelar padrões temáticos de forma automatizada.

1.1 MOTIVAÇÃO

A motivação para este estudo surge da necessidade de entender, com maior clareza, como diferentes sites de tecnologia estruturam seus conteúdos e quais temas aparecem com mais frequência em suas publicações. Além da grande quantidade de textos, há uma diversidade de categorias, estilos e abordagens editoriais que tornam o *corpus* heterogêneo e complexo. Diante disso, tornou-se relevante aplicar técnicas capazes de organizar esse conjunto de maneira objetiva, permitindo identificar padrões semânticos e tendências editoriais que contribuam para a descoberta de conhecimento neste domínio.

1.2 PROBLEMA

Apesar da abundância de conteúdos, ainda é difícil identificar automaticamente como esses textos se agrupam, quais temas predominam e como diferentes portais estruturam suas publicações. Falta um método sistemático que permita analisar, comparar e extrair conhecimento desses documentos de forma eficiente.

1.3 HIPÓTESE

A hipótese deste estudo é que técnicas de mineração de textos, aplicadas ao *corpus* e avaliadas pelas métricas do Cassiopeia — Coesão e Acoplamento — são capazes de identificar padrões temáticos consistentes e organizar textos semelhantes em agrupamentos relevantes.

1.4 CONTRIBUIÇÃO

Este trabalho demonstra como a clusterização pode ser utilizada para descobrir conhecimento em um conjunto de 1.864 textos coletados dos sites Canaltech, Drifto, MeioBit, TechTudo e Tecnoblog. A análise inclui 200 simulações realizadas no Cassiopeia, permitindo

avaliar a qualidade dos agrupamentos formados e evidenciar tendências e padrões presentes no *corpus.*

1.5 PROPOSTA

Aplicar mineração de texto ao *corpus*, realizar simulações no Cassiopeia e identificar agrupamentos temáticos com base nas métricas de qualidade. Inicialmente, os textos foram organizados e analisados no *FineCount,* e posteriormente submetidos a 100 simulações de Coesão e Acoplamento. O objetivo é revelar como os textos se distribuem, quais temas se destacam e como essas técnicas contribuem para a descoberta de conhecimento em sites de tecnologia.

2 REFERENCIAL TEÓRICO

2.1 MINERAÇÃO DE TEXTO

A mineração de texto representa uma abordagem para o campo da descoberta de conhecimento que utiliza métodos de análise e extração para buscar informações a partir de um conjunto de textos, frases ou palavras. Seu objetivo principal é extrair conhecimento relevante que está oculto em meio a massa de dados textuais desestruturados. Como aponta Moura (2004), a mineração de textos é uma área de pesquisa tecnológica cujo foco é a busca por padrões, tendências e regularidades em textos escritos em linguagem natural. O processo envolve analisar, descobrir padrões e transformar essa matéria em conhecimento aplicável, muitas vezes com o suporte de técnicas de aprendizado de máquina. A essência da mineração de textos é usar recursos computacionais para encontrar informações que não são óbvias em uma leitura convencional. As fontes para extrair são variáveis, incluindo documentos, e-mails, sites da internet e publicações em redes sociais.

Neste contexto, desempenham um papel crucial o uso das técnicas de PLN (Processamento de linguagem natural), porque capacitam as máquinas a entender, interpretar e gerar linguagem de uma maneira mais próxima à dos humanos. A parte da execução da mineração passa por fluxo de etapas, começando pela coleta e organização dos dados textuais para facilitar o acesso e a manipulação. Na sequência, ocorre a extração de informações que utiliza métodos como a análise semântica - busca entender o significado das palavras dentro do seu contexto, enquanto a análise estatística se concentra na frequência com que as palavras aparecem e nos padrões que podem indicar. Subsequente a extração, começa a mineração propriamente dita, onde são gerados os *insights* — por exemplo, a identificação de tendências, padrões de comportamento ou a sugestão de novos caminhos para a tomada de decisão. Entre

as diversas metodologias de mineração de texto que são aplicadas durante a análise, destaca-se a clusterização que será abordada na seção seguinte.

2.1.1 CLUSTERIZAÇÃO

Clusterização é uma técnica que agrupa documentos, textos ou dados com base em características semelhantes, sem a necessidade de orientação prévia. Seu objetivo é formar grupos chamados de clusters, que sejam bastante homogêneos internamente e, ao mesmo tempo, diferentes em relação aos outros grupos, facilitando a identificação de padrões (Alves, 2007; Aggarwal, 2015). Esse processo permite que, em grandes volumes de informações, subconjuntos menores sejam criados, contribuindo para o melhor entendimento e análise dos dados. Além disso, a clusterização pode ser usada como uma etapa preliminar para outros algoritmos, ajudando a melhorar sua eficiência.

2.1.2 CASSIOPEIA

Desenvolvido por Guelpeli (2012), Cassiopeia é um modelo de agrupamento de textos baseado em sumarização. Esse modelo apresenta um novo método que reduz a alta dimensionalidade e os dados esparsos, fundamentado na curva da Lei de Zipf. Com base no ponto de corte de Luhn, que define dois limiares — um limite inferior e um limite superior — na distribuição de frequência de palavras (Luhn, 1958), o Cassiopeia propõe um corte médio para identificar as palavras mais significativas. Para isso, utiliza centroides que representam a média desses termos-chave dentro de cada agrupamento. Esses centroides permitem organizar os textos em uma hierarquia, agrupando-os pela similaridade entre as palavras (GUELPELI, 2012).

A estrutura do modelo Cassiopeia é composta por três etapas principais: préprocessamento, processamento e pós-processamento. Durante o pré-processamento, uma análise computacional inicial é realizada em cada texto. Aplica-se a técnica de *case folding* (Witten *et al., 1*994), que consiste na conversão de todos os caracteres para minúsculas ou maiúsculas. Diante disso, são removidos elementos não textuais, como imagens, tabelas e marcações, a fim de preparar os dados para as próximas fases para que o Cassiopeia funcione independentemente do idioma, uma vez que as *stopwords* podem ser mantidas. Segundo Aranha (2007), as *stopwords* são palavras que aparecem com frequência em diversos tipos de textos, mas possuem baixa relevância semântica, sendo normalmente excluídas no sistema de recuperação da informação. Na etapa de processamento, os textos são agrupados com base na similaridade textual, e cada grupo é representado por um centróide formado pelas palavras mais significativas do conjunto. A escolha dessas palavras é feita com base na frequência

média nos textos do grupo, os centroides desses agrupamentos têm a dimensionalidade dos vetores limitada a 50 posições, apresentada por Wives (2004). Devido à sua natureza dinâmica, o agrupamento requer a redistribuição de textos, da criação de novos grupos e a integração de grupos que já estão formados. Dessa forma, a reorganização ocorre de maneira hierárquica, utilizando uma abordagem top-down, e continua até que os centroides se estabilizem, ou seja, até que não ocorram mudanças significativas na inclusão de novos textos. Na fase final, o pós-processamento organiza os textos em uma hierarquia e associa a cada grupo um resumo que destaca as informações importantes. O resumo evidencia os assuntos principais dos textos originais em cada agrupamento, facilitando o entendimento e compreensão dos resultados obtidos. O Cassiopeia contempla quatro métodos principais, que se diferenciam conforme a forma de aplicação da sumarização e da clusterização: clusterização simples, sumarização simples, sumarização e clusterização com *stopwords* e sumarização e clusterização sem *stopwords*. No entanto, dado que este trabalho se concentra exclusivamente à clusterização simples, o foco será dado somente a esse método, que consiste no agrupamento direto dos textos completos, sem a realização prévia de etapas de sumarização ou de remoção de palavras irrelevantes, permitindo identificar padrões e similaridades entre os textos em sua forma integral.

2.1.2.1 MÉTRICAS DO CASSIOPEIA

Para a análise da qualidade dos agrupamentos textuais, diversas métricas podem ser utilizadas, entre elas *Precision, Recall,* Coesão e Acoplamento. Manning *et al.* (2008) descrevem *Precision* (Precisão) como a proporção de objetos corretamente atribuídos a um agrupamento em relação ao total de objetos presentes neste agrupamento. Já *Recall* (ou Revocação) é descrito como a proporção de objetos alocados corretamente em relação ao total de objetos que pertencem à classe correspondente ao agrupamento.

Por outro lado, Kunz e Black (1995) conceituam coesão como o nível de semelhança entre os componentes que compõem um mesmo agrupamento, ou seja, quanto maior a similaridade entre os membros, maior será a coesão interna do grupo. O Acoplamento, por sua vez, avalia a similaridade média entre pares de elementos em que um pertence a um determinado agrupamento e o outro não. Dessa forma, a Coesão mede a homogeneidade dentro dos grupos, enquanto o Acoplamento analisa a dissimilaridade entre grupos diferentes.

2.2 CORPUS

O *corpus* diz respeito ao conjunto de materiais que serão usados como base para um estudo. Os dados do *corpus* podem incluir documentos, informações ou qualquer tipo de conteúdo importante que será analisado pelo pesquisador. O *corpus r*epresenta o conteúdo principal para embasar a pesquisa. Ele pode ser composto por textos, entrevistas, gravações, imagens ou outros dados que estejam relacionados ao tema em questão. A partir dele, o pesquisador consegue desenvolver suas análises, chegar a conclusões e apresentar os resultados. No campo da Análise do Discurso (AD), o *corpus* não é simplesmente "encontrado" ou dado de antemão, mas construído a partir de gestos de leitura, interpretação e compreensão do objeto de investigação, seguindo critérios teóricos e não empíricos (Orlandi, 2002). Dessa forma, definir o *corpus* já é um passo analítico importante, pois implica a seleção de recortes temáticos e a definição de propriedades discursivas relevantes para a pesquisa (Marquezan, 2009). No ambiente acadêmico, o *corpus* tem papel fundamental, pois conecta o tema e o problema da pesquisa, além de possibilitar uma investigação mais aprofundada, seja por meio de um conjunto previamente estabelecido (*corpus* estático) ou um conjunto que se desenvolve ao longo do estudo (*corpus* dinâmico). De todo modo, sua organização é sempre orientada pela teoria e pela problemática da pesquisa, em um movimento constante de idas e vindas entre ambas (Marquezan, 2009).

2.3 TÉCNICA DE NUVEM DE PALAVRA

A técnica nuvem de palavras é uma representação gráfica que destaca os termos mais comuns ou significativos em um texto. Nessa visualização, as palavras mais recorrentes aparecem de forma mais proeminente, geralmente em tamanhos maiores, facilitando a identificação rápida dos principais tópicos ou conceitos da análise. Essa técnica surgiu como uma forma simples e eficaz de analisar grandes volumes de texto, permitindo reconhecer padrões e tendências sem a leitura completa (LUNARDI; CASTRO; MONAT, 2008). Atualmente, ela é utilizada em áreas como marketing, educação e pesquisa qualitativa, por sua capacidade de converter informações textuais em representações visuais de fácil compreensão.

Na área da educação, Ramsden e Bate (2008) destacam o potencial das nuvens de palavras como ferramenta de ensino e aprendizagem. No Brasil, elas também têm sido aplicadas em pesquisas na área da saúde, tanto em estudos qualitativos quanto na formação de profissionais, com resultados relevantes (KAMI *et al.*, 2016; SOUZA *et al.,* 2018). Segundo Freitas (2023), a nuvem de palavras oferece uma visão geral das palavras-chave de um texto,

mostrando os termos mais recorrentes e, com isso, revelando os temas mais abordados. Essa representação visual facilita a identificação ágil dos elementos mais significativos em um contexto específico, de maneira direta e clara.

Entre as principais vantagens da técnica estão a facilidade de leitura, a agilidade na produção, a atratividade visual e a capacidade de proporcionar *insights* instantâneos. Apesar desses benefícios, é importante considerar algumas limitações: como o destaque excessivo de palavras comuns ou de pouco significado, além da disposição aleatória dos termos, que pode dificultar a interpretação em alguns casos.

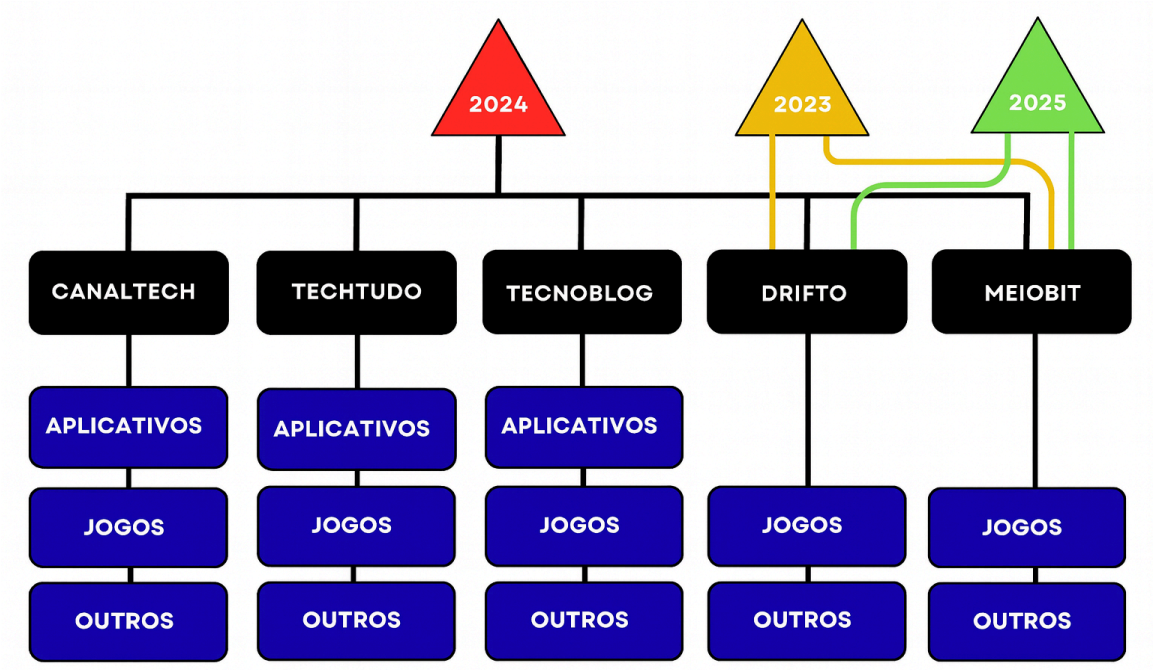
3. METODOLOGIA

A metodologia desta pesquisa teve como objetivo descobrir conhecimento em textos, com foco na análise e agrupamento de notícias de tecnologia extraídas de cinco sites especializados. O propósito principal é identificar padrões e categorizar os textos em grupos significativos, utilizando técnicas de mineração de texto

3.1 CORPUS

Nesta pesquisa, utilizou-se um *corpus* composto por 1.864 arquivos de textos extraídos de cinco sites de tecnologia — Canaltech, Drift, MeioBit, TechTudo e Tecnoblog, coletados ao longo de três anos (2023, 2024 e 2025). A Figura 1 ilustra a composição e estrutura desse *corpus.*

Figura 1. Composição do corpus da pesquisa.



Fonte: Elaborado pelo autora

Conforme demonstrado na Figura 1, o *corpus* organiza-se em três dimensões fundamentais: a dimensão temporal (período de 2023 a 2025), as fontes de coleta (cinco veículos specializados em tecnologia) e as categorias temáticas (Jogos, Aplicativos e Outros). A representação visual evidencia a abrangência e distribuição dos textos coletados, destacando que: A categoria "Jogos" está presente em todos os sites analisados, configurando-se como o segmento de maior representatividade no *corpus*; A categoria "Aplicativos" foi identificada em três dos cinco veículos (Canaltech, TechTudo e Tecnoblog); A categoria "Outros" agrupa as demais temáticas não especificadas nas categorias anteriores, complementando o espectro temático da análise.

A coleta foi realizada manualmente em cada site, garantindo que os textos selecionados correspondem às categorias de interesse e estivessem disponíveis integralmente para análise. Os arquivos foram salvos no formato .txt, seguindo uma nomenclatura padronizada composta pelo ano, abreviação do site e categoria. Por exemplo: 2024_CT_J1.txt refere-se a um texto do ano de 2024, do site Canaltech, na categoria Jogos. Após a organização dos arquivos, o *corpus* foi processado no software *FineCount,* utilizado para verificação de termos e análise de frequências. A Figura 2 apresenta a interface da ferramenta com os resultados obtidos

Tabela 1. Dados extraídos do processamento no FineCount

Documento	Caracteres	Palavras	Linhas	Páginas	Taxa
1864 arquivo(s)	6985710	1379649	152648,95	8198,92	-

Fonte: Elaborado pelo autora

3.2 PROCEDIMENTOS E SIMULAÇÕES

Os arquivos de texto foram inseridos no software Cassiopeia, onde foram realizadas as simulações de agrupamento (*clustering).* As simulações tiveram como propósito avaliar a coesão interna e o acoplamento dos agrupamentos formados. As análises seguiram um modelo não supervisionado e contemplaram:

a) 100 simulações sobre o mesmo *corpus* para avaliação das métricas de Coesão;

b) 100 simulações sobre o mesmo *corpus* para avaliação do Acoplamento.

4 RESULTADOS E DISCUSSÕES

Nesta seção, são apresentados os principais resultados encontrados nas análises dos agrupamentos formados, destacando as semelhanças e diferenças identificadas na coesão e nos padrões ocultos identificados. As informações aqui apresentadas têm o objetivo de proporcionar uma visão mais clara da estrutura dos textos analisados, assim como de sua organização interna. Na métrica de Coesão e Acoplamento, foram identificados 414 *clusters.* Considerando os objetivos da pesquisa — que consistem em analisar o conteúdo textual extraído de sites especializados em tecnologia — essas métricas mostraram-se adequadas para avaliar a qualidade dos agrupamentos, permitindo observar a organização interna dos *clusters* e o nível de similaridade entre seus elementos

4.1 DISTRIBUIÇÃO DOS 414 CLUSTERS

Os 414 *clusters* identificados foram classificados conforme a quantidade de textos que os compõem. Para isso, foram definidas faixas de tamanho, permitindo observar uma concentração significativa em agrupamentos menores. Constatou-se que 366 *clusters* apresentam menos de cinco textos; 42 *clusters* abrangem entre 6 e 20 textos; 5 *clusters*

concentram entre 21 e 50 textos; e apenas 1 *cluster* ultrapassa 50 textos. Essa distribuição está representada na Tabela 2.

Tabela 2. Distribuição dos Clusters

Faixa de Tamanho do Cluster	Quantidade de Clusters
Menos de 5 textos	366
De 6 a 20 textos	42
De 21 a 50 textos	5
Mais de 50 textos	1

Fonte: Elaborado pelo autora

Foram realizada uma análise detalhada desses grupos, com o objetivo de identificar padrões e extrair *insights* relevantes em cada categoria

4.1.1 Clusters com até 5 textos

Considerando o total de 366 *clusters,* optou-se por realizar uma amostragem manual baseada em porcentagem. Foram selecionados 21 *clusters* que apresentaram índice de coesão superior a 75%, correspondendo a 42 arquivos em formato TXT, o que equivale a aproximadamente 6% do total de *clusters.* A seleção foi definida a partir das porcentagens de coesão fornecidas pelas métricas do Cassiopeia. Para a análise, foi estabelecida uma faixa de coesão que permitiu identificar diferentes níveis entre os *clusters.* Na faixa de altíssima coesão classificada como (≥ 90%), foram encontradas 5 *clusters* (23, 238, 254, 334, 409) representando 23,81% do total. Na faixa de média coesão que inclui valores de (≥ 75% e < 80%) foram encontradas 9 *clusters* (18, 20, 94, 133, 152, 190, 349, 367, 376) totalizando 42,86% do total. Em relação à predominância temporal nos 21 *clusters* analisados, o ano de 2024 foi o mais frequente, com 33 arquivos, representando 78,57% do total. Nas categorias foi predominante com 42,86% dos arquivos a categoria CT que representa o site CanalTech. O sufixo _O foi o mais presente, aparecendo em 50% dos arquivos. Para melhor compreensão, serão apresentadas visualizações por meio da técnica de nuvem de palavras,

evidenciando os motivos que levaram à formação dos agrupamentos.

4.1.1.1 Cluster 18

O *cluster* 18, com coesão de 75%, é composto pelos arquivos 202_TT_O109.txt e 2024_TT_O107.txt, ambos da categoria "Outros" do site TechTudo, referentes ao ano de 2024. A análise da nuvem de palavras evidencia a predominância dos termos "Notebook", "Tela", "Ram", "SSD", "Black Friday" e "Gamer", revelando que o conteúdo textual concentra-se em promoções de equipamentos eletrônicos direcionados ao público gamer.

Figura 2. Cluster 18.

Fonte: Elaborado pelo autor

4.1.1.2 Cluster 20

O *cluster* 20, com coesão de 77%, agrupa os arquivos 2024_MB_O9.txt e 2024_MB_-O13.txt, ambos do site MeioBit, categoria "Outros", ano 2024. A frequência dos termos "Chips", "China", "EUA", "Sanções", "Intel", "Nvidia"e "Huawei"indica que os textos abordam questões 9 relacionadas à indústria de semicondutores em contexto geopolítico, especialmente os impactos das sanções tecnológicas entre potências econômicas sobre fabricantes de componentes eletrônicos.

Figura 3. Cluster 20.

Fonte: Elaborado pelo autor

4.1.1.3 Cluster 23

O *cluster* 23 apresenta coesão de 91%, sendo composto pelos arquivos 2024_TT_-O147.txt (TechTudo) e 2024_MB_O15.txt (MeioBit), ambos de 2024 e da categoria "Outros". A predominância dos termos "Apple", "Samsung", "iPhone 15", "Oled", "Usb-C" e "Processadores" caracteriza uma discussão técnica sobre o mercado de smartphones, com ênfase em lançamentos de produtos, especificações de componentes e análise comparativa entre fabricantes.

Figura 4. Cluster 23.

Fonte: Elaborado pelo autor

4.1.1.4 Cluster 39

O *cluster* 39 com coesão de 81%, reúne os arquivos 2024_DFT_O201.txt (Drifto) e 2024_TT_O161.txt (TechTudo), ambos de 2024, categoria "Outros". Os termos mais frequentes — "Netflix", "Coreanas", "Séries", "Filmes", "Catálogo", "2024"e "Polêmica"— indicam uma análise do fenômeno de expansão das produções audiovisuais asiáticas, particularmente sulcoreanas, em plataformas de streaming, abordando tanto aspectos quantitativos (crescimento no catálogo) quanto qualitativos (recepção e controvérsias).

Figura 5. Cluster 39.

Fonte: Elaborado pelo autor

4.1.1.5 Cluster 44

O *cluster* 44, com coesão de 94%, reúne os arquivos 2024_TT_A3.txt e 2024_TT_A35.txt, ambos do TechTudo, categoria “Aplicativos”. A nuvem de palavras destaca termos como “Tubemate”, “Música”, “Apk”, “Baixar”, “Vídeos” e “Grátis”, indicando discussão sobre apps de download de conteúdo multimídia via fontes não oficiais. Também aparecem com força “Segurança”, “Malwares” e “Spywares”, apontando preocupações com riscos digitais associados ao uso de APKs fora das lojas oficiais.

Figura 6. Cluster 44.

Fonte: Elaborado pelo autor

4.1.1.6 Cluster 79

O *cluster 79,* com coesão de 94%, é formado pelos arquivos 2025_DFT_O14.txt e 2025_DFT_O4.txt, ambos do site Drifto, categoria "Outros". Observa-se que os arquivos possuem conteúdo idêntico. Essa característica não configura erro de coleta; ao contrário, indica que o sistema identificou corretamente a duplicidade de textos dentro da mesma fonte, justificando o elevado índice de coesão do *cluster*. Os termos predominantes — "Dark", "Jogador", "Franquia" e "Jogo"— remetem ao lançamento de um título da franquia Dark, com ênfase em características como brutalidade e experiência imersiva do jogador.

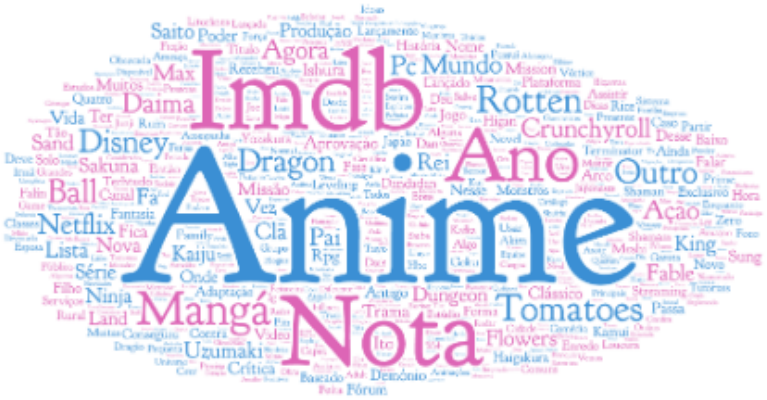
Figura 7. Cluster 79.

Fonte: Elaborado pelo autor

4.1.1.7 Cluster 94

O *cluster 94*, com coesão de 75%, agrupa os arquivos 2024_DFT_O199.txt (Drifto) e 2024_TT_O175.txt (TechTudo), ambos da categoria “Outros”. A frequência dos termos “anime”, “nota”, IMDb, “mangá” e Netflix caracteriza conteúdo focado em crítica e avaliação de produções de animação japonesa, com particular atenção às métricas de classificação em plataformas como IMDb e Rotten Tomatoes, indicando uma abordagem comparativa entre sistemas de avaliação.

Figura 8. Cluster 94.



Fonte: Elaborado pelo autor

4.1.1.8 Cluster 113

O *cluster 113*, com coesão de 75%, reúne os arquivos 2024_DFT_O82.txt (Drifto) e 2024_TT_J28.txt (TechTudo), ambos classificados na categoria “Outros”. Os termos mais frequentes — “Free Fire”, “Mira”, “Sensibilidade”, “Emulador”, “Celular”, “Melhor” e “Configuração” — indicam conteúdo voltado à otimização de desempenho no jogo Free Fire, abordando aspectos técnicos como ajustes de configuração, personalização de controles e uso de emuladores para aprimorar a experiência de jogo.

Figura 9. Cluster 113.

Fonte: Elaborado pelo autor

4.1.1.9 Cluster 133

O *cluster 133* apresenta coesão de 75% e agrupa os arquivos 2024_CT_J43.txt e 2024_- CT_J27.txt, ambos do site Canaltech, categoria "Jogos". A predominância dos termos "Game", "Awards", "Jogo", "Categoria", "Prêmio", "Troféu"e "Melhor"evidencia que os textos tratam da premiação Game Awards, cobrindo aspectos como categorias de competição, indicações e análise dos títulos concorrentes, refletindo a importância deste evento no calendário da indústria de jogos eletrônicos.

Figura 10. Cluster 133.

Fonte: Elaborado pelo autor

4.1.1.10 Cluster 152

O *cluster 152* com coesão de 75%, é composto pelos arquivos 2024_CT_O3.txt e 2024_CT_A29.txt, ambos do site Canaltech. A frequência dos termos "Caixa", "Loterias", "Site", "App", "Instabilidade", "Serviço", "Problema" e "Plataforma" revela foco em questões técnicas relacionadas a falhas e instabilidades em plataformas digitais de serviços de loteria, abordando problemas de acesso, funcionamento e experiência do usuário.

Figura 11. Cluster 152.

Fonte: Elaborado pelo autor

4.1.1.11 Cluster 190

O *cluster 190,* com coesão de 76%, agrupa os arquivos 2024_CT_J4.txt (Canaltech) e 2024_TT_J44.txt (TechTudo), ambos da categoria "Jogos". Os termos predominantes — "Jogo", "Futebol", "Game", "Soccer", "Jogador", "Bola" e "Time"— caracterizam conteúdo sobre jogos eletrônicos de futebol (soccer games), abordando aspectos como mecânica de jogo, experiência do usuário e funcionalidades das plataformas de simulação esportiva.

Figura 12. Cluster 190.

Fonte: Elaborado pelo autor

4.1.1.12 Cluster 203

Com coesão de 84%, o *cluster 203* é formado pelos arquivos 2024_CT_J29.txt (Canaltech) e 2023_TB_J35.txt (Tecnoblog), ambos pertencentes à categoria “Jogos”. A frequência dos termos “Jogo”, “Grande”, “Lançamento”, “Esperado”, “Franquia”, “Título” e “Aguardado” indica que os textos abordam grandes lançamentos da indústria de jogos

eletrônicos, com ênfase nas expectativas do público consumidor, na análise de franquias consolidadas e nas projeções sobre novos títulos no mercado gamer.

Figura 13. Cluster 203.

Fonte: Elaborado pelo autor

4.1.1.13 Cluster 238

O *cluster 238* apresenta coesão de 96% e é composto por arquivos 2024_CT_O108.txt 2024_CT_J10.txt do site Canaltech, pertencentes às categorias “Outros” e “Jogos”. Observa-se que os arquivos possuem conteúdo idêntico. Essa característica não indica falha na etapa de coleta; ao contrário, evidencia a capacidade do sistema de identificar textos duplicados mesmo quando classificados em categorias diferentes dentro da mesma fonte. Tal duplicidade explica o elevado índice de coesão registrado no cluster. A análise da nuvem de palavras revela a predominância dos termos “PS”, “Sony”, “AMD”, “Console”, “Tecnologia”, “Gráfico” e “Ray Tracing”, indicando que o conteúdo aborda aspectos tecnológicos dos consoles PlayStation, com ênfase na parceria estratégica com a fabricante AMD e na implementação de recursos avançados de desempenho gráfico, especialmente a tecnologia de ray tracing.

Figura 14. Cluster 238.

Fonte: Elaborado pelo autor

4.1.1.14 Cluster 254

O *cluster 254* com coesão de 98%, agrupa os arquivos 2025_DFT_O3.txt e 2025_-DFT_O26.txt, ambos do site Drifto, categoria "Outros". Os arquivos apresentam conteúdo idêntico, explicando o alto percentual de coesão. Os termos mais frequentes — "Chuva", "Tempestade", "São Paulo", "Risco", "Fortes", "Regiões" e "Impactos"— caracterizam conteúdo de natureza meteorológica, focado em alertas sobre eventos climáticos extremos, com particular atenção aos riscos de tempestades intensas e potenciais alagamentos na região metropolitana de São Paulo.

Figura 15. Cluster 254

Fonte: Elaborado pelo autor

4.1.1.15 Cluster 203

O *cluster 334*, com coesão de 90%, é formado pelos arquivos 2024_CT_O30.txt e 2024_CT_O2.txt, ambos do site Canaltech, categoria "Outros". A frequência dos termos "Watch", "Huawei", "Bem", "Sistema", "Monitoramento", "Motorista" e "Exercícios" evidencia que os textos abordam tecnologias smartwatches e outros dispositivos de uso pessoal, com foco em funcionalidades de monitoramento de saúde, acompanhamento de atividades físicas e sistemas de rastreamento de bem-estar do usuário.

Figura 16. Cluster 203.

Fonte: Elaborado pelo autor

4.1.1.16 Cluster 349

O *cluster 349* Com coesão de 78%, reúne os arquivos 2024_TT_O18.txt (TechTudo) e 2023_DFT_O37.txt (Drifto), ambos da categoria "Outros". Os termos predominantes — "Game", "Modo", "Configurações", "Desempenho", "Celular"e "Memória"— indicam conteúdo voltado para otimização de dispositivos móveis para jogos eletrônicos, abordando aspectos técnicos como ajustes de configuração, gerenciamento de recursos de sistema, otimização de memória e configurações específicas do modo de jogo (game mode) em smartphones.

Figura 17. Cluster 349.

Fonte: Elaborado pelo autor

4.1.1.17 Cluster 359

O *cluster 359* apresenta coesão de 98% e é composto pelos arquivos 2024_TT_A75.txt e 2024_TT_A67.txt, ambos do site TechTudo, categoria "Aplicativos". A análise da nuvem de palavras revela a predominância dos termos "Horóscopo", "Planner", "App", "Celular", "Pessoal", "Previsões", "Astrologia"e "Mapa", indicando que o conteúdo aborda aplicativos móveis voltados para astrologia e planejamento pessoal, com funcionalidades que integram previsões astrológicas personalizadas e ferramentas de organização (planner) para dispositivos móveis.

Figura 18. Cluster 359.

Fonte: Elaborado pelo autor

4.1.1.18 Cluster 367

O *cluster 367* apresenta coesão de 75% e agrupa os arquivos 2024_TB_A24.txt (Tecnoblog, categoria “Aplicativos”) e 2024_CT_O75.txt (Canaltech, categoria “Outros”). Os termos mais frequentes — “WhatsApp”, “Cliente”, “Vendas”, “CRM”, “Áudio”, “Meta” e “Processo” — evidenciam que os textos abordam a utilização do aplicativo WhatsApp como ferramenta empresarial, com foco em estratégias de vendas, gestão de relacionamento com clientes (CRM) e processos de comunicação corporativa, incluindo recursos de mensagens de áudio, alinhados às políticas da Meta

Figura 19. Cluster 367.

Fonte: Elaborado pelo autor

4.1.1.19 Cluster 376

O *cluster* 376, com coesão de 75%, é formado pelos arquivos 2024_CT_A44.txt (categoria "Aplicativos") e 2024_CT_O178.txt (categoria "Outros"), ambos do site Canaltech. A nuvem de palavras revela conexão temática com o universo ficional do Homem-Aranha e da Marvel, destacando os termos "Melhor", "Peter", "Ano", "App", "Android", "Google" e "Jogo". 17 Isso indica que o conteúdo aborda aplicativos e jogos baseados na franquia Marvel para dispositivos Android, disponíveis na plataforma Google Play, possivelmente tratando de lançamentos, avaliações ou recomendações relacionadas ao personagem.

Figura 20. Cluster 376.

Fonte: Elaborado pelo autor

4.1.1.20 Cluster 409

O *cluster* 409, com coesão de 91%, o Cluster 409 é composto pelos arquivos 2024_CT_- J28.txt (Canaltech) e 2023_TB_J43.txt Tecnoblog, ambos da categoria "Jogos", porém de anos distintos. A frequência dos termos "Jogo", "PlayStation", "Exclusivo", "Sony", "Franquia", "Lançamento" e "Console" indica que os textos abordam jogos exclusivos para consoles PlayStation, com ênfase em lançamentos de franquias consolidadas e estratégias de exclusividade da Sony no mercado de jogos eletrônicos.

Figura 21 Cluster 409.

Fonte: Elaborado pelo autor

4.1.1.21 Cluster 411

O *cluster* 411 apresenta coesão de 83% e agrupa os arquivos 2024_CT_O21.txt e 2024_- CT_O17.txt, ambos do site Canaltech, categoria "Outros". A nuvem de palavras evidencia a predominância dos termos "Watch", "Galaxy", "Relógio", "Atividade" e "Saúde", caracterizando conteúdo focado em tecnologias vestíveis (wearables), especificamente smartwatches da linha Galaxy Watch (Samsung), com ênfase em funcionalidades de monitoramento de atividades físicas e indicadores de saúde

Figura 22. Cluster 411.

Fonte: Elaborado pelo autor

4.1.2 Clusters com 6 a 20 textos

Foram identificados 42 *clusters* que contêm entre 6 e 20 arquivos .txt. Para investigar o comportamento desses agrupamentos, foram selecionados dois exemplos representativos: o *cluster* 114, com 19 arquivos, e o *cluster* 1, com 7 arquivos.

4.1.2 .1 Cluster 114 — 19 arquivos

O *cluster* 114 apresenta uma predominância de 47,37% do ano 2024 e categoria principal DFT com (78,95%). O Sufixo "_O"(outros) está presente em 84,21% dos arquivos tendo como temas recorrentes o entretenimento, premiações, recordes e futebol. Exemplo de arquivo: 2024_- DFT_O48.txt — Relato sobre o desempenho recente do Manchester United, que somou apenas 22 pontos em 18 jogos, aproximando-se mais da zona de rebaixamento do que da classificação para a Champions League.

4.1.2 .2 Cluster 1 — 7 arquivos

O *cluster* 1 demostrar predominância com 100% dos arquivos de texto com ano de 2024 da categoria TB. O sufixo "_O" está presente em 71,43% dos arquivos. Exemplos de arquivos: 2024_TB_O179.txt — “TIM recebe multa milionária por telemarketing ilegal”. 2024_TB_- O177.txt — “Vivo anuncia pacotes de roaming internacional”. Nessa configuração, o sufixo "_O" refere-se especificamente a conteúdos de operadoras de telefonia e serviços de telecomunicações.

A partir dessa análise, é possível comparar os *clusters* 114 e 1, identificando pontos em comum e diferenças. Ambos apresentam uso expressivo do sufixo “_O”, sugerindo que ele pode indicar uma categoria de conteúdo variável. O c*luster* 114 reúne conteúdos de mídia, conhecimento, esportes, tecnologia e entretenimento digital, incluindo marcas como Samsung, Apple e Microsoft. A análise também evidencia distinta representatividade: o c*luster* 1 concentra temas corporativos de telecomunicações, abordando serviços como roaming e pacotes de dados, com menções a Vivo, Claro e TIM. Por isso, o *cluster* 1 constitui um nicho mais específico dentro do conjunto analisado

4.1.3 CLUSTERS COM 21 A 50 TEXTOS

Foram identificados cinco *clusters* que contêm entre 20 e 50 arquivos .txt. Para compreender o comportamento editorial e temático desses agrupamentos, foram analisados todos os *clusters* individualmente.

4.1.3.1 Cluster 196 — 32 arquivos

O *cluster* 196 apresenta predominância do ano de 2024 (83,64%), com a categoria principal CT (36,36%). O sufixo mais frequente é _O (50,91%). Os temas recorrentes incluem atualizações regulatórias e tecnológicas, aplicações de inteligência artificial, saúde, mobilidade, além de uma variedade equilibrada entre conteúdos normativos, tecnológicos e relacionados a jogos. Exemplos de arquivos: a) 2024_CT_O44.txt — Mudanças previstas na CNH em 2025. B) 2024_TB_O75.txt — Smartwatches com IA para diagnóstico de apneia do sono

4.1.3.2 Cluster 101 — 35 arquivos

O *cluster* 101 também tem 2024 como ano predominante (91,43%). Suas principais categorias são TT (37,14%), TB (34,29%) e DFT (28,57%). O sufixo _O aparece em 57,14% dos arquivos. Os temas recorrentes abrangem novidades de tecnologia e consumo, destaques de mercado, inteligência artificial e eventos sazonais. Exemplos de arquivos: a) 2024_TT_O140.txt — Novidades sobre o ChatGPT) b) 2024_TT_O100.txt — Notebooks na Black Friday. c) 2024_TB_O19.txt — Ofertas do iPhone 15 Pro Max.

4.1.3.3 Cluster 117 — 20 arquivos

O *cluster* 117 apresenta 2024 como ano predominante (60,00%) e tem como categoria principal DFT (55,00%). O sufixo _J aparece em 45% dos arquivos. Entre os temas recorrentes, o desenvolvimento e tendências em jogos digitais, aplicações de inteligência artificial no entretenimento, conteúdos variados entre tecnologia, apps e operadoras. Exemplos de arquivos: a) 2024_DFT_J59.txt — Parceria entre Sonic e Angry Birds, que une os universos das franquias em cinco jogos para dispositivos móveis. B) 2024_DFT_J9.txt — Como a EA Games prevê que a IA poderá transformar o futuro dos jogos

4.1.3.4 Cluster 391 — 22 arquivos

O *cluster* 391 também apresenta predominância de 2024 (82,61%). As categorias principais são TT (52,38%), referente ao site TecTudo, e CT (23,81%), do CanalTech. Já o sufixo _O é novamente o mais frequente com _O(66,67%). Seus temas recorrentes focam em funcionalidades e atualizações de aplicativos populares, recursos de mensagem e inteligência artificial, e variedade entre tecnologia móvel, comunicação e inovação. Exemplos de arquivos: a) 2024_TT_O125.txt — WhatsApp libera filtros, efeitos e planos de fundo. B) 2024_TT_O129.txt — ChatGPT: criação de gráficos, resumo de arquivos.

4.1.3.5 Cluster 348 — 42 arquivos

O *cluster* 348 apresenta predominância do ano 2024 (88,10%) e categorias TB (38,10%) do Tecnoblog e CT (30,95%) do CanalTech. O sufixo mais comum é _O (73,81%). Os temas recorrentes neste *cluster* são lançamentos e promoções de celulares e dispositivos móveis, atualizações de marcas como Samsung e Motorola, e conteúdos voltados ao consumo tecnológico. Exemplos de arquivos: a) 2024_TB_O25.txt — Promoção do Motorola Edge 50

Fusion com 37. b) 2024_TB_O5.txt — Lançamento dos modelos Samsung Galaxy S25, S25 Plus e S25 Ultra.

Os cinco *clusters* analisados — 196, 101, 117, 391 e 348 — revelam uma produção editorial fortemente concentrada no ano 2024, com presença dominante das categorias TT, TB, CT e DFT. O conteúdo gira em torno de tecnologia, inovação digital, consumo informativo e entretenimento, com abordagens variadas, mas complementares. O sufixo "_O"OUTROS é o mais frequente, aparecendo em todos os *clusters* como marcador de conteúdos informativos e comerciais, especialmente sobre celulares, aplicativos, IA e serviços digitais. Já o sufixo "_- J" JOGOS aparece com destaque no *cluster* 117, indicando foco em jogos digitais.

4.1.4 CLUSTERS COM MAIS DE 50 TEXTOS.

Neste grupo foi encontrado apenas o *cluster* 206, cuja predominância textual é do ano de 2024 (98,89%). As principais categorias presentes são MB (33,33%), TB (32,59%) e TT (25,19%). Quanto aos sufixos, _J aparece em 51,11% dos arquivos, enquanto _O ocorre em 40,37%. Os temas recorrentes incluem conteúdo voltado para o universo dos jogos eletrônicos, notícias sobre lançamentos, atualizações e análises do mercado. Há também alta incidência de arquivos com o sufixo “_J”, associado diretamente a conteúdos sobre jogos. Exemplos de arquivos do *cluster*:a) Relato sobre os jogos Sonic Superstars e Super Mario Bros. Wonder, destacando que a desenvolvedora Sega atribuiu o desempenho abaixo do esperado de Sonic Superstars à concorrência com Super Mario Bros. Wonder. b) Discussão sobre os jogos The Crew 2 e The Crew Motorfest, da Ubisoft, que receberam atualizações para permitir jogabilidade offline.

Dessa forma, interpreta-se que o *cluster* 206 apresenta uma forte concentração de conteúdos relacionados ao universo dos jogos, especialmente no que diz respeito a lançamentos e atualizações de títulos de grande popularidade.

5 CONCLUSÃO

A pesquisa realizada aplicou técnicas de mineração de textos e métodos de agrupamento para identificar padrões presentes em notícias provenientes de sites especializados em tecnologia. O estudo teve como objetivo compreender a organização temática desses conteúdos e verificar se a estrutura dos *clusters* permitiria evidenciar áreas de maior concentração informacional.

A análise dos agrupamentos permitiu não apenas uma visão geral do *corpus,* mas também uma investigação detalhada sobre a composição e as características de cada *cluster*. Esse processo evidenciou a predominância de temas relacionados à tecnologia de consumo, jogos digitais, dispositivos móveis e tendências do mercado, demonstrando a eficácia da metodologia adotada para revelar padrões relevantes e estruturar o conhecimento presente nos textos analisados. Os resultados indicaram que o uso de métricas quantitativas, como Coesão e Acoplamento, é eficiente para identificar similaridades, duplicidades e agrupamentos temáticos, contribuindo para a descoberta de conhecimento em grandes volumes de textos. O método empregado também se mostrou vantajoso por reduzir o tempo de organização e filtragem do *corpus,* facilitando etapas posteriores de investigação.

Como trabalhos futuros, recomenda-se ampliar o volume de documentos analisados e explorar outras métricas e técnicas de mineração de textos, a fim de aprofundar a compreensão sobre a produção de conteúdo em sites de tecnologia e identificar possíveis mudanças nas tendências ao longo do tempo.

REFERÊNCIAS

ARANHA, C. N. **Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional.** 2007. Tese (Doutorado em Informática) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

CARVALHO JÚNIOR, P. M. *et al.* **O uso de nuvens de palavras como ferramenta de análise no ensino em saúde.** *Revista Brasileira de Educação Médica*, v. 36, n. 2, p. 45-51, 2012.

DEL BUONO, R. **O que é o corpus de uma pesquisa?** 2014. Disponível em: http://www.abntouvancouver.com.br/2014/03/o-que-e-o-corpus-de-uma-pesquisa.html. Acesso em: 25 out. 2025.

EBECKEN, N. F. F.; LOPES, M. C. S.; COSTA, M. C. A. Mineração de textos. In: REZENDE, S. O. (org.). **Sistemas inteligentes: fundamentos e aplicações.** 1. ed. São Paulo: Manole, 2003. cap. 13, p. 337–370.

FREITAS, C. **Nuvem de palavras: o que é e como utilizar para melhorar a escrita.** Guia da Carreira, 2023. Disponível em: https://www.guiadacarreira.com.br/blog/nuvem-de-palavras. Acesso em: 22 out. 2025.

GUELPELI, M. V. C. **Cassiopeia: um modelo de agrupamento de textos baseado em sumarização**. 2012. Tese (Doutorado em Computação) – Universidade Federal Fluminense, Niterói, 2012.

KUNZ, T.; BLACK, J. P. **Using automatic process clustering for design recovery and distributed debugging**. IEEE Transactions on Software Engineering, v. 21, n. 6, p. 515–527, 1995.

LOPES, L. F. C. **Descoberta de conhecimento em óbitos pediátricos na região de Diamantina.** 2014. Disponível em: https://facet.ufvjm.edu.br/wp-content/uploads/decom-tcc/2014/LUIZ_FELIPE_CORDEIRO_LOPES_VERSAO1_2014_I.pdf. Acesso em: 17 out. 2025.

LUNARDI, M. S.; CASTRO, J.; MONAT, **A. Visualização dos resultados do Yahoo em nuvens de texto: uma aplicação construída a partir de web service**s. *InfoDesign – Revista Brasileira de Design da Informação*, v. 5, n. 1, p. 21–35, 2008.

LUHN, H. P. **The automatic creation of literature abstracts.** IBM Journal of Research and Development, v. 2, n. 2, p. 159–165, 1958.

MANNING, C. D.; RAGHAVAN, P.; SHUTZE, H. **Introduction to information retrieval.** Cambridge: Cambridge University Press, 2008.

MARQUEZAN, R. **A constituição do corpus de pesquisa**. *Educação Especial*, Santa Maria, v. 22, n. 33, p. 97–110, jan./abr. 2009. Disponível em: https://periodicos.ufsm.br/educacaoespecial/article/view/172/102. Acesso em: 17 set. 2025.

MOURA, M. F. **Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos**. Campinas: Embrapa Informática Agropecuária, 2004. (ISSN 1677-9274).

ORLANDI, E. P. **Análise de discurso: princípios e procedimentos.** Campinas: Pontes, 2002.

RAMSDEN, A.; BATE, **A. Using word clouds in teaching and learning**. Bath: University of Bath, 2008.

WITTEN, I. H.; MOFFAT, A.; BELL, T. C. **Managing gigabytes.** New York: Van Nostrand Reinhold, 1994.