Aplicação de *Doc2Vec* e Técnicas de Clusterização em Estudos de Impacto Ambiental de Pequenas Centrais Hidrelétricas em Mato Grosso

Use of *Doc2Vec* and *Cluster*ing Techniques in Environmental Impact Studies of Small Hydroelectric Power Plants in Mato Grosso

[1]Lucas Michelotti Baldini, [2]Anderson Castro Soares de Oliveira, [3]Lia Hanna Martins Morita, [4]Ibraim Fantin da Cruz

[1]Mestre em Recursos Hídricos – Faculdade Católica de Cuiabá (michelotti.lucas@gmail.com)
[2]Doutor em Estatística e Experimentação Agropecuária – Universidade Federal de Mato Grosso (anderson.oliveira@ufmt.br)
[3]Doutora em Estatística – Universidade Federal de Mato Grosso (lia.morita@ufmt.br)
[4]Doutor em Recursos Hídricos e Saneamento Ambiental – Universidade Federal de Mato Grosso (ibraimfantin@gmail.com)

RESUMO: Com a expansão das PCHs como alternativas de menor impacto, há uma demanda crescente por avaliações específicas de seus efeitos em fauna, flora, solo e recursos hídricos. Este estudo analisa os impactos ambientais das Pequenas Centrais Hidrelétricas (PCHs) em Mato Grosso, aplicando técnicas de mineração e clusterização de textos para examinar Estudos de Impacto Ambiental (EIAs) dessas usinas. Utilizando-se a técnica de representação vetorial do *Doc2Vec* para gerar vetores semânticos dos textos, os documentos foram agrupados em três *clusters* que refletem abordagens distintas: o *Cluster* 1 foca em impactos ambientais amplos e medidas de mitigação; o *Cluster* 2 enfatiza o monitoramento da qualidade da água e controle de erosão; e o *Cluster* 3 prioriza respostas rápidas para impactos no solo e socioeconômicos. Essa análise revela como os EIAs abordam os desafios ambientais das PCHs, evidenciando a importância de políticas públicas e estratégias de mitigação adaptadas a cada contexto ecológico, com vistas a oferecer subsídios para um planejamento ambiental mais efetivo e sustentável em Mato Grosso. Assim, os resultados obtidos reforçam a relevância de abordagens técnico-científicas no tratamento dos EIAs, contribuindo para a formulação de diretrizes mais precisas e contextualizadas que promovam o desenvolvimento sustentável das PCHs no estado.

Palavras Chave: Sustentabilidade; Gestão ambiental; Impactos ecológicos; Mineração de texto.

ABSTRACT: With the expansion of Small Hydropower Plants (SHPs) as lower-impact alternatives, there is a growing demand for specific assessments of their effects on fauna, flora, soil, and water resources. This study analyzes the environmental impacts of SHPs in Mato Grosso by applying text mining and clustering techniques to examine the Environmental Impact Studies (EISs) of these plants. Using the Doc2Vec vector representation technique to generate semantic vectors from the texts, the documents were grouped into three clusters reflecting distinct approaches: Cluster 1 focuses on broad environmental impacts and mitigation measures; Cluster 2 emphasizes water quality monitoring and erosion control; and Cluster 3 prioritizes rapid responses to soil and socioeconomic impacts. This analysis reveals how EISs address the environmental challenges of SHPs, highlighting the importance of public policies and mitigation strategies tailored to each ecological context, in order to provide support for more effective and sustainable environmental planning in Mato Grosso. Thus, the results reinforce the relevance of technical-scientific approaches in handling EISs, contributing to developing more precise and contextualized guidelines that promote the sustainable development of SHPs in the state.

Keywords: Sustainability; Environmental Management; Ecological impacts; Text mining.

1. INTRODUÇÃO

O desenvolvimento econômico e a expansão de infraestruturas em larga escala impulsionaram discussões sobre os impactos ambientais, especialmente em projetos de grande porte, como os empreendimentos hidrelétricos. No Brasil, a criação da Política Nacional do Meio Ambiente com a Lei Federal nº 6.938 de 1981 (BRASIL, 1990) consolidou instrumentos de controle e avaliação de impactos ambientais, como o Estudo de Impacto Ambiental (EIA) e o Relatório de Impacto Ambiental (RIMA). Esses documentos técnicos são fundamentais para identificar e avaliar os possíveis impactos ambientais, promovendo práticas de desenvolvimento sustentável e protegendo áreas ecologicamente sensíveis (REBOUÇAS, BRAGA e TUNDISI, 2015). A regulamentação ambiental foi fortalecida pelo Decreto Federal nº 99.274 de 1990 e pelas resoluções do Conselho Nacional do Meio Ambiente (CONAMA), que instituíram o EIA-RIMA como ferramenta central no licenciamento ambiental (BRASIL, 1986; BRASIL, 1997).

Dentro desse contexto, as Pequenas Centrais Hidrelétricas (PCHs) emergem como uma alternativa sustentável devido à sua menor potência instalada, de até 30 megawatts, e ao menor impacto ambiental em comparação com grandes hidrelétricas. A Lei nº 9.074 de 1995 estabeleceu os critérios de concessão para aproveitamentos hidrelétricos de pequeno porte, favorecendo a implementação de PCHs em regiões como Mato Grosso, que possui alto potencial hídrico (CLAUBERG, HENKES e BECEGATTO, 2021; BRASIL, 1995). Entretanto, apesar de serem alternativas com menor impacto, as PCHs ainda geram impactos ambientais consideráveis, que exigem uma análise criteriosa. O EIA-RIMA detalha efeitos diretos e indiretos sobre fauna, flora, solo e recursos hídricos, e sua análise sistemática é essencial para o controle ambiental.

Diante do aumento exponencial de dados textuais não estruturados, a mineração de texto surge como ferramenta estratégica para extrair conhecimento e identificar padrões em textos complexos. Essa técnica integra algoritmos de PLN com métodos estatísticos, transformando dados brutos em informações úteis (SILGE e ROBINSON, 2017; ZIZKA, DARENA e SVOBODA, 2019). Suas aplicações incluem desde análise de satisfação até extração de informações em contextos científicos e jurídicos. No campo ambiental, destaca-se pela automatização da análise de documentos como os EIAs (FELDMAN e SANGER, 2006; MORAIS e AMBRÓSIO, 2007). Essa abordagem contribui para uma compreensão mais aprofundada dos impactos de usinas hidrelétricas.

A clusterização de texto, em particular, organiza documentos em grupos com base em suas similaridades, facilitando a análise de grandes volumes de dados e identificando padrões

e temas comuns. Este método é amplamente utilizado para descobrir informações relevantes e visualizar estruturas em coleções textuais diversificadas (STEINBACH, KARYPIS e KUMAR, 2000; HUSSEIN, ALI e MOHAMED, 2015). A transformação de palavras e textos em representações matemáticas é crucial para que máquinas compreendam e manipulem o conteúdo textual, permitindo a execução de tarefas como análise de sentimentos, tradução automática e sumarização de textos. Diversas técnicas, como *one hot encoding, bag of words* e *embeddings*, desempenham papéis fundamentais nesse processo (TURIAN, RATINOV e BENGIO, 2010; JURAFSKY e MARTIN, 2019).

Entre as técnicas de representação vetorial, destaca-se o Word2Vec, que cria vetores capazes de capturar relações semânticas e sintáticas entre palavras (MIKOLOV et al., 2013). Avanços como o Doc2Vec permitem representar documentos inteiros em vetores distribuídos (LE e MIKOLOV, 2014). Esse modelo utiliza os algoritmos PV-DM e PV-DBOW, que integram informações de palavras e parágrafos (LE e MIKOLOV, 2014; CASSIANO e CORDEIRO, 2018). No presente estudo, a implementação do Doc2Vec no R (WIJFFELS, 2021) viabiliza a extração e o agrupamento automático de informações em EIAs de PCHs. Tal abordagem oferece subsídios para políticas ambientais e planejamento sustentável.

As redes textuais surgem como outra ferramenta inovadora, permitindo a análise da coocorrência de palavras para estabelecer relações semânticas e sintáticas em textos de diferentes comprimentos (BAIL, 2016; QUISPE, TOHALINO e AMANCIO, 2021; LIU, *et al*., 2022). Com base no princípio da *triadic closure*, as redes textuais ajudam a extrair significado mesmo de textos curtos, possibilitando comparações e detecção de temas entre diferentes documentos. A análise de comunidades em redes complexas, como a identificação de grupos de nós interconectados, também fornece uma perspectiva valiosa para a compreensão das estruturas em grandes redes textuais (FORTUNATO, 2010; BLONDEL, *et al*., 2008).

Este estudo aplica o programa R (*R Core Team*, 2023) para realizar clusterização e mineração de texto em 15 EIAs de PCHs localizadas em Mato Grosso, empregando o *Doc2Vec* para identificar padrões nos impactos ambientais. O objetivo principal é revelar temas e fatores ambientais recorrentes de maneira automatizada, contribuindo para a análise de impacto ambiental e para a elaboração de políticas de licenciamento mais eficazes e sustentáveis, além de corroborar com a necessidade de revisão legislativa da formulação destes documentos.

2. MATERIAIS E MÉTODO

Este estudo baseia-se na coleta de Estudos de Impacto Ambiental (EIA) de 25 PCHs localizadas no estado de Mato Grosso. Para cada uma dessas PCHs, foram obtidos os documentos completos dos EIAs em formato PDF (Quadro 1), disponíveis no portal da Secretaria Estadual de Meio Ambiente (SEMA-MT). A localização espacial de cada PCH foi determinada e organizada utilizando o *udpipe* R (R Core Team, 2023), que possibilitou o mapeamento e a disposição geográfica das usinas.

Quadro 1-Organização dos Estudos de Impacto Ambiental por PCH

EIA	**PCHs**	**Município**	**Rio Represado**	**Potência Máx (MegaWats)**
EIA 1	Barra da Onça, Alto Garças	Alto Garças e Guiratinga	Rio das Garças	13 MW
EIA 2	Cabaçal I, IV, V, VI, VII e VIII	Rio Branco, São José dos Quatro Marcos, Araputanga e Reserva do Cabaçal	Cabaçal	37,9 MW
EIA 3	Cristalina	Sapezal e Campos de Julio	Juruena	16 MW
EIA 4	Cumbuco	Primavera do Leste	Cumbuco	20 MW
EIA 5	Entre Rios	Primavera do Leste	Rio das Mortes	28 MW
EIA 6	Estivadinho	Primavera do Leste	Jauru	9,9 MW
EIA 7	Formoso I, Formoso II, Formoso III	Tangará da Serra	Formoso	57,5 MW
EIA 8	Galera	Nova Lacerda e Conquista D'oeste	Galera	13 MW
EIA 9	Itiquira III	Itiquira	Itiquira	20 MW
EIA 10	Juina	Campos de Julio	Juina	29,2 MW
EIA 11	Mogno	Brasnorte	Cravari	9,3 MW
EIA 12	Octacilio Lucion	Pontes e Lacerda	Pindaituba	19 MW
EIA 13	Rancho Grande, Progresso	Indiavai	Córrego do Sangue	11,7 MW
EIA 14	Sacre 14	Brasnorte	Sacre	34,5 MW
EIA 15	Vila União	Primavera do Leste	Rio das Mortes	20,90 MW

Fonte: EIA's disponíveis no site da SEMA-MT. Elaboração: Autor 2024.

2.1 Preparação dos dados

Na fase inicial, todos os arquivos de texto dos EIAs foram listados e lidos a partir da pasta especificada, onde cada documento foi convertido em um único texto contínuo e armazenado em um *data frame* com colunas para identificação (ID) e conteúdo textual. Esse

processo permitiu centralizar os dados textuais dos EIAs e possibilitou o processamento de cada um em uma estrutura homogênea e acessível para a análise subsequente.

A redução das palavras à sua forma base (lematização) foi aplicada para garantir a consistência dos termos e facilitar a análise semântica. Esse processo utilizou um modelo *Udpipe* específico para o português, aplicado através do pacote *udpipe* (WIJFFELS, 2023), com o qual foram identificadas as formas básicas das palavras, eliminando variações morfológicas.

O pré-processamento dos textos incluiu uma série de passos para uniformizar e eliminar ruídos linguísticos:

a) Conversão das palavras para letras minúsculas, com o intuito de evitar distinções entre maiúsculas e minúsculas, garantindo uma análise mais precisa.
b) Remoção de palavras irrelevantes (*stopwords*), pontuação, números e hifens, utilizando os pacotes *stopwords* (BENOIT, MUHR e WATANABE, 2021) e tm (FEINERER e HORNIK, 2024). A eliminação desses elementos ajuda a manter o foco nos termos mais relevantes e frequentes para a análise.
c) Exclusão de espaços em branco redundantes, resultando em um corpus limpo e normalizado, o que otimiza o desempenho dos algoritmos de processamento de linguagem natural aplicados na análise subsequente.

Este pré-processamento foi crucial para garantir a precisão e consistência das informações, facilitando a identificação de temas, padrões e relações semânticas.

2.2 Mineração de Texto com *Doc2Vec*

Para capturar as similaridades semânticas entre os documentos, utilizou-se a técnica *Doc2Vec*, uma extensão do *Word2Vec* que permite representar documentos inteiros como vetores numéricos. A aplicação dessa técnica envolveu uma série de etapas para otimizar a qualidade e a eficiência dos vetores de representação.

Primeiramente, os textos foram divididos em segmentos menores (*chunks*) para facilitar o processamento e melhorar a eficiência do treinamento do modelo *Doc2Vec*. Essa divisão permitiu que o algoritmo analisasse pequenas partes do texto, mantendo a coerência semântica e reduzindo a carga computacional.

Em seguida, um modelo de *embeddings* de palavras foi treinado utilizando o pacote *Word2Vec* (WIJFFELS e WATANABE, 2023) para gerar representações vetoriais das palavras, com os seguintes parâmetros ajustados para a precisão:

ES Engineering and Science ISSN: 2358-5390 DOI: 10.18607/ES20251418843 Volume 14 Número 1

Dimensão dos vetores (dim): 150, Iterações (iter): 200, Janela de contexto (window): 10 palavras, Contagem mínima (min_count): 1, Taxa de aprendizado (lr): 0,05, Processamento paralelo utilizando 2 *threads* para otimização.

Esses parâmetros foram ajustados para capturar com precisão as relações semânticas entre as palavras. Em seguida, o modelo *Doc2Vec* foi treinado com o pacote *Doc2Vec* (WIJFFELS, 2021) utilizando os *embeddings* gerados pelo *Word2Vec*. Esse modelo empregou a abordagem PV-DBOW, uma técnica de *Doc2Vec* específica para otimizar a qualidade dos vetores de documentos. Os parâmetros do modelo *Doc2Vec* foram configurados da mesma forma que o modelo *Word2Vec*, com exceção das Iterações (iter), que no modelo *Doc2Vec* passam a ser 1000. A representação vetorial dos documentos é uma ferramenta poderosa que captura a similaridade semântica entre os textos dos EIAs, facilitando a identificação de padrões e temas comuns.

2.3 Formação de *Clusters*

Com o modelo *Doc2Vec* treinado, os vetores resultantes foram empregados para calcular a similaridade entre os documentos, identificando aqueles com conteúdo similar e facilitando a análise de agrupamento (*clustering*). O processo de formação de *clusters* seguiu uma série de etapas para garantir a precisão e a interpretação dos resultados.Primeiramente, as similaridades entre os vetores de cada documento foram calculadas, permitindo identificar o grau de proximidade semântica entre os textos. Em seguida, foi construído um grafo de similaridade utilizando os pacotes igraph (CSÁRDI e NEPUSZ, 2006) e ggraph (PEDERSEN, 2024). Neste grafo, cada nó representa um EIA, enquanto as arestas indicam a intensidade de similaridade entre os documentos, proporcionando uma representação visual das conexões semânticas entre eles.

Para identificar comunidades de documentos semelhantes dentro do grafo, aplicou-se a técnica de passeios aleatórios curtos, que permitem agrupar documentos em *clusters* com base nas conexões. Posteriormente, foram utilizadas técnicas de clusterização com o pacote *tidygraph* (PEDERSEN, 2024), que geraram grupos de documentos com alta similaridade semântica.

Os *clusters* formados foram então analisados detalhadamente, permitindo a identificação de similaridades e discrepâncias nos diagnósticos socioambientais. Essa análise possibilitou observar temas específicos como a qualidade da água, erosão do solo, ou impactos sobre a biodiversidade, que emergem de maneira diferenciada em cada *cluster*,

ES Engineering and Science ISSN: 2358-5390 DOI: 10.18607/ES20251418843 Volume 14 Número 1

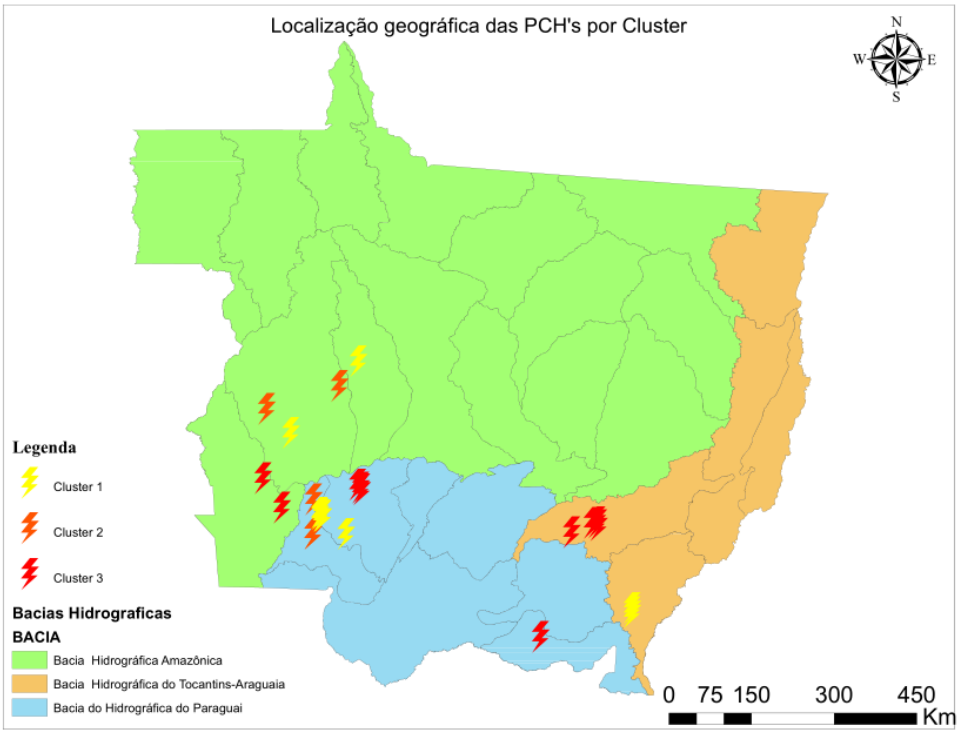
fornecendo *insights* valiosos para políticas ambientais e estratégias de mitigação adaptadas a cada contexto ecológico e geográfico.

3. RESULTADOS E DISCUSSÕES

A Figura 1 apresenta a distribuição espacial das PCHs nas principais bacias hidrográficas da região: Amazônica, Tocantins-Araguaia e Paraguai. Cada uma dessas bacias possui características ambientais únicas que influenciam o nível de impacto das PCHs e as necessidades específicas de gestão.

Na Bacia Amazônica, foram identificadas cerca de sete PCHs, inseridas em uma das maiores e mais complexas redes hidrográficas da região. Apesar de sua alta capacidade de diluição, o que minimiza o impacto percentual das PCHs, intervenções concentradas exigem atenção especial (LATRUBESSE et al., 2017; CASTELLO e MACEDO, 2016). Trechos ecologicamente sensíveis podem ser afetados mesmo em um sistema com grande volume de água. Por isso, é crucial uma gestão sustentável das PCHs nessa bacia. Tal cuidado visa preservar a biodiversidade e os ecossistemas locais, vulneráveis a alterações no fluxo hídrico.

Figura 1 - Localização geográfica das PCH's por *Cluster*.



Fonte: Autores 2024.

A Bacia do Tocantins-Araguaia, no leste de Mato Grosso, abriga cerca de três PCHs, cujo impacto é relevante devido à limitada capacidade de absorção de intervenções sem afetar

ES Engineering and Science ISSN: 2358-5390 DOI: 10.18607/ES20251418843 Volume 14 Número 1

o fluxo natural dos rios (CASTELLO e MACEDO, 2016). Já a Bacia do Paraguai, com cerca de seis PCHs, é especialmente sensível, pois depende de ciclos naturais de cheias e secas para manter sua biodiversidade (ALHO e SILVA, 2018; HAMILTON, 2016). A presença das usinas amplia os riscos de impacto. Por isso, é essencial um planejamento rigoroso para preservar os regimes hidrológicos e os ecossistemas alagáveis.

Diante desse cenário, observa-se que cada bacia possui uma capacidade de suporte distinta, e a concentração de PCHs em áreas sensíveis pode intensificar os impactos locais. A análise da distribuição percentual dessas usinas pode orientar políticas públicas mais eficazes, promovendo um planejamento sustentável e licenciamento ambiental adequado às realidades regionais (MYERS e SIROIS, 2006). Além disso, os diferentes clusters revelam diversidade nos aspectos físicos das PCHs, como localização, capacidade energética e características hidrológicas. Essa heterogeneidade reforça a importância de estratégias de gestão específicas. Assim, torna-se essencial considerar as particularidades de cada bacia.

A análise do Cluster 1 revela uma ampla dispersão geográfica das PCHs, a potência estimada varia de 2,5 MW a 16 MW. As Unidades de Planejamento e Gestão (UPGs) identificadas incluem TA-3, P2, A-14 e A-13, refletindo essa variação.

Apesar da diversidade em localização e capacidade de geração de energia, o Cluster 1 mantém uma coesão em termos de tipo de empreendimento e função. A análise desses fatores físicos permite uma compreensão mais detalhada da composição das PCHs agrupadas, facilitando o planejamento e a gestão de políticas ambientais e de infraestrutura. Esse entendimento é essencial para o desenvolvimento de estratégias eficazes que promovam a sustentabilidade e minimizem os impactos ambientais dessas operações hidrelétricas.

O Cluster 2 agrupa PCHs com ampla diversidade geográfica e variação na capacidade de geração de energia, que vai de 2,5 MW a 34,5 MW. Essa variação evidencia o potencial energético do grupo. No total, o cluster apresenta uma capacidade instalada expressiva, resultado da heterogeneidade de suas usinas.

A análise dos fatores físicos facilita a compreensão da composição e das características das PCHs agrupadas, auxiliando no planejamento e na gestão de políticas ambientais e de infraestrutura. Esse entendimento é essencial para a implementação de estratégias que promovam a sustentabilidade e minimizem os impactos ambientais dessas operações hidrelétricas.

O Cluster 3 apresenta uma diversidade significativa em termos de localização e capacidade de geração de energia, mantendo, no entanto, características comuns que justificam sua inclusão no mesmo grupo.

A capacidade de geração de energia das PCHs no Cluster 3 varia de 13 MW a 28 MW. A capacidade total do Cluster é expressiva, destacando-se o conjunto das PCHs Formoso I, II e III, que somam 57,5 MW, a maior dentro do grupo. As PCHs do Cluster 3 estão associadas a uma diversidade hidrológica significativa, envolvendo rios como Cumbuco, Rio das Mortes, Formoso, Galera, Itiquira e Pindaituba.

O Cluster 3 apresenta uma combinação de diversidade geográfica e variação na capacidade de geração de energia, mantendo, no entanto, uma coesão em termos de tipo de empreendimento e função. A análise desses fatores físicos contribui para uma melhor compreensão da composição e das características das PCHs agrupadas, auxiliando na gestão e no planejamento de políticas ambientais e de infraestrutura. Esse conhecimento é essencial para a implementação de estratégias eficazes que promovam a sustentabilidade e minimizem os impactos ambientais dessas operações hidrelétricas.

A análise dos dados físicos e de planejamento revela que todos os Clusters possuem pelo menos uma PCH inserida na UPG P2 (Alto Paraguai Médio). Observa-se também que o Cluster 2 apresenta mais ligações externas do que internas, devido ao fato de contar com apenas 5 PCHs no total. Em comparação, o Cluster 1, apesar de possuir o mesmo número de documentos analisados (4 EIAs), reúne 10 PCHs, número igual ao do Cluster 3, indicando diferentes padrões de distribuição entre os grupos.

Esse detalhamento dos fatores físicos por cluster facilita a compreensão das características específicas de cada grupo, proporcionando subsídios essenciais para o planejamento e a implementação de políticas de sustentabilidade e mitigação de impactos ambientais dessas operações hidrelétricas (HAMILTON, 2016).

Analise Textual dos *Clusters*

A Figura 2 apresenta o grafo de similaridade entre os EIAs das PCHs, resultante da aplicação do modelo *Doc2Vec* e das técnicas de Clusterização. A análise de *Cluster* é uma técnica estatística usada para classificar elementos em grupos, de forma que elementos dentro de um mesmo conglomerado (*Cluster*) sejam parecidos, e os elementos em diferentes *Clusters* sejam distintos entre si (MYERS e SIROIS, 2006). Cada nó no grafo representa uma EIA enquanto as arestas representam as similaridades entre os textos dos EIA. Os nós e as arestas são agrupados em três *Clusters* distintos, destacados por cores diferentes:

a) Primeiro *Cluster* (vermelho): EIA 1, EIA 2, EIA 3 e EIA 11.
b) Segundo *Cluster* (azul): EIA 6, EIA 10, EIA 13 e EIA 14.
c) Terceiro *Cluster* (verde): EIA 4, EIA 5, EIA 7, EIA 8, EIA 9, EIA 12 e EIA 15.

ES Engineering and Science ISSN: 2358-5390 DOI: 10.18607/ES20251418843 Volume 14 Número 1

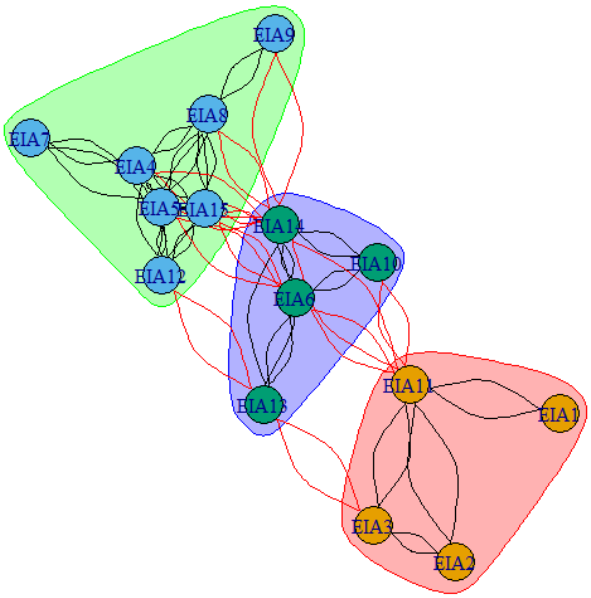
Os *clusters* destacam conjuntos de EIAs com alta similaridade interna, facilitando a identificação de padrões recorrentes nos diagnósticos socioambientais.

As arestas que conectam os nós refletem a intensidade da similaridade semântica entre os EIAs: conexões mais densas indicam uma maior correlação semântica, sugerindo que os documentos podem compartilhar dados, metodologias ou conclusões semelhantes. A identificação das comunidades no grafo facilita a detecção de agrupamentos naturais com base nas similaridades entre os textos, destacando os temas mais recorrentes e áreas de interseção nos diagnósticos (CASTELLO e MACEDO, 2016).

Essa análise de *clusters* oferece uma estrutura para explorar padrões e tendências nos EIAs, permitindo o desenvolvimento de diretrizes e recomendações para grupos de documentos com temas comuns. A comparação entre *clusters* distintos proporciona *insights* sobre variações nas abordagens e diagnósticos ambientais adotados nos estudos, o que pode auxiliar na formulação de políticas públicas e estratégias de mitigação mais eficazes (ALHO e SILVA, 2018; HAMILTON, 2016).

A estrutura do grafo revela não apenas a similaridade dentro dos *Clusters*, mas também a relação entre os diferentes *Clusters*. A análise dos passeios aleatórios curtos permitiu identificar essas comunidades dentro do grafo, facilitando a visualização das inter-relações entre os EIA. Formados os *clusters* foi realizada a localização geográfica por *cluster* conforme a figura 1.

Figura 2 - Grafo de Similaridade obtido por técnicas de clusterização e Doc2Vec entre os Estudos de Impacto Ambiental (EIA) das Pequenas Centrais Hidrelétricas (PCHs).



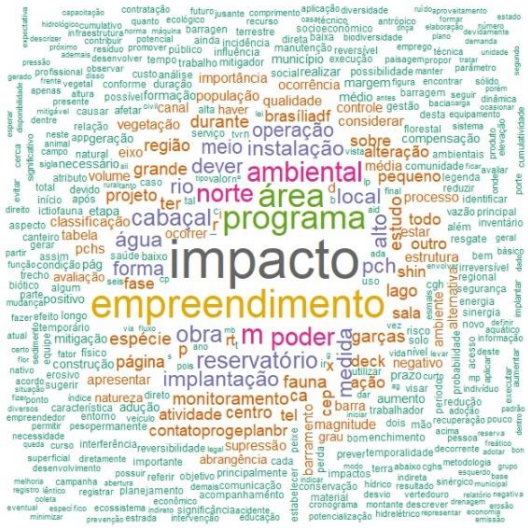
Fonte: Autores 2024.

Na Figura 3 são apresentadas as nuvens de palavras dos três *clusters*, cada uma representando os termos mais frequentes utilizados nos EIAs agrupados em cada *cluster*. Essas nuvens oferecem uma visão visual das palavras-chave e temas predominantes, refletindo diferenças nas abordagens e prioridades ambientais para cada conjunto de EIAs.

Figura 3 - Nuvem de palavras representando os termos mais frequentes nos *cluster.*

(a)

(b)



(c)

Fonte: Autores 2024.

Ao comparar as três nuvens de palavras, observa-se que o *Cluster* 1 (figura 3a) se concentra principalmente na análise dos impactos ambientais e das alterações nas condições

naturais, sugerindo uma abordagem mais geral sobre os efeitos ambientais. O *Cluster* 2 (figura 3b) destaca "programa" e "medida", sugerindo uma orientação mais prática, voltada para a implementação de ações de mitigação e controle ambiental durante a implantação dos empreendimentos. O *Cluster* 3 (figura 3c) compartilha algumas palavras-chave com o *Cluster* 2, mas com uma ênfase adicional em "norte", indicando uma possível especificidade geográfica e um enfoque regional nos estudos ambientais.

Essas distinções refletem diferentes enfoques entre os EIAs agrupados em cada *cluster*, variando entre uma preocupação mais ampla com impactos ambientais (*Cluster* 1), estratégias concretas de mitigação e controle (*Cluster* 2) e uma abordagem regionalizada, provavelmente ajustada às particularidades geográficas (*Cluster* 3). As diferenças nas palavras-chave de cada *cluster* fornecem *insights* sobre como cada grupo de EIAs aborda os desafios ambientais e sociais das PCHs, variando entre uma análise ampla dos impactos ambientais, uma orientação prática para mitigação e um enfoque geográfico específico. Essa análise é essencial para direcionar políticas públicas e estratégias de gestão ambiental ajustadas às necessidades de cada contexto.

Na Figura 4 são apresentadas as nuvens de palavras chave dos três *clusters*, cada uma representando os termos mais frequentes nos EIAs agrupados em cada *cluster*. Essas nuvens revelam os temas e palavras-chave que são mais recorrentes em cada grupo de EIAs. O método usa modelos de aprendizado de máquina para fazer a *tokenização*, a *etiquetagem morfológica* e a *análise sintática*. Para extrair palavras-chave, o *UDPipe* identifica e conta as palavras mais relevantes na frase com base na estrutura gramatical do texto (DURAN, *et al.* 2022).

No *Cluster* 1 (figura 4a), as palavras-chave como MEDIDA e MITIGADOR foram centrais, refletindo uma ênfase nas soluções e medidas propostas para mitigar os impactos ambientais. Termos como CURTO PRAZO, MÉDIO PRAZO e LONGO PRAZO surgiram como palavras-chave subsequentes, indicando uma preocupação com a implementação de soluções escalonadas ao longo do tempo.

Já no *Cluster* 2 (figura 4b), a palavra PROCESSO EROSIVO ganhou destaque ao lado de CURTO PRAZO, sugerindo uma problemática específica relacionada à instalação das PCH’s nesse grupo. A palavra BAIXO MAGNITUDE também apareceu como uma palavra-chave influente, enquanto ALTA MAGNITUDE e ALTO GRAU não tiveram grande destaque. Isso aponta que o foco no *Cluster* 2 está na solução de problemas erosivos e na gestão de impactos de menor magnitude.

ES Engineering and Science ISSN: 2358-5390 DOI: 10.18607/ES20251418843 Volume 14 Número 1

Comparativamente, no *Cluster* 3 (figura 4c), a palavra ATIVIDADE TRANSFORMADOR revela uma percepção de que as atividades associadas às PCH's têm um impacto transformador significativo. Isso indica que as PCH's nesse *Cluster* são vistas como agentes de transformação substancial no uso do solo, ecossistemas ou socioeconomia local. A presença dessa palavra-chave sugere que as PCH's no *Cluster* 3 são percebidas como tendo um impacto mais transformador do que nos outros *Clusters*.

Assim, a análise de palavras-chave nos diferentes *Clusters* destaca como cada grupo de PCH's enfrenta questões ambientais e sociais distintas, com palavras-chave específicas refletindo as preocupações predominantes em cada conjunto de estudos.

Figura 4 - Nuvem de palavras chave de bigrama por *cluster*.

(a) (b)

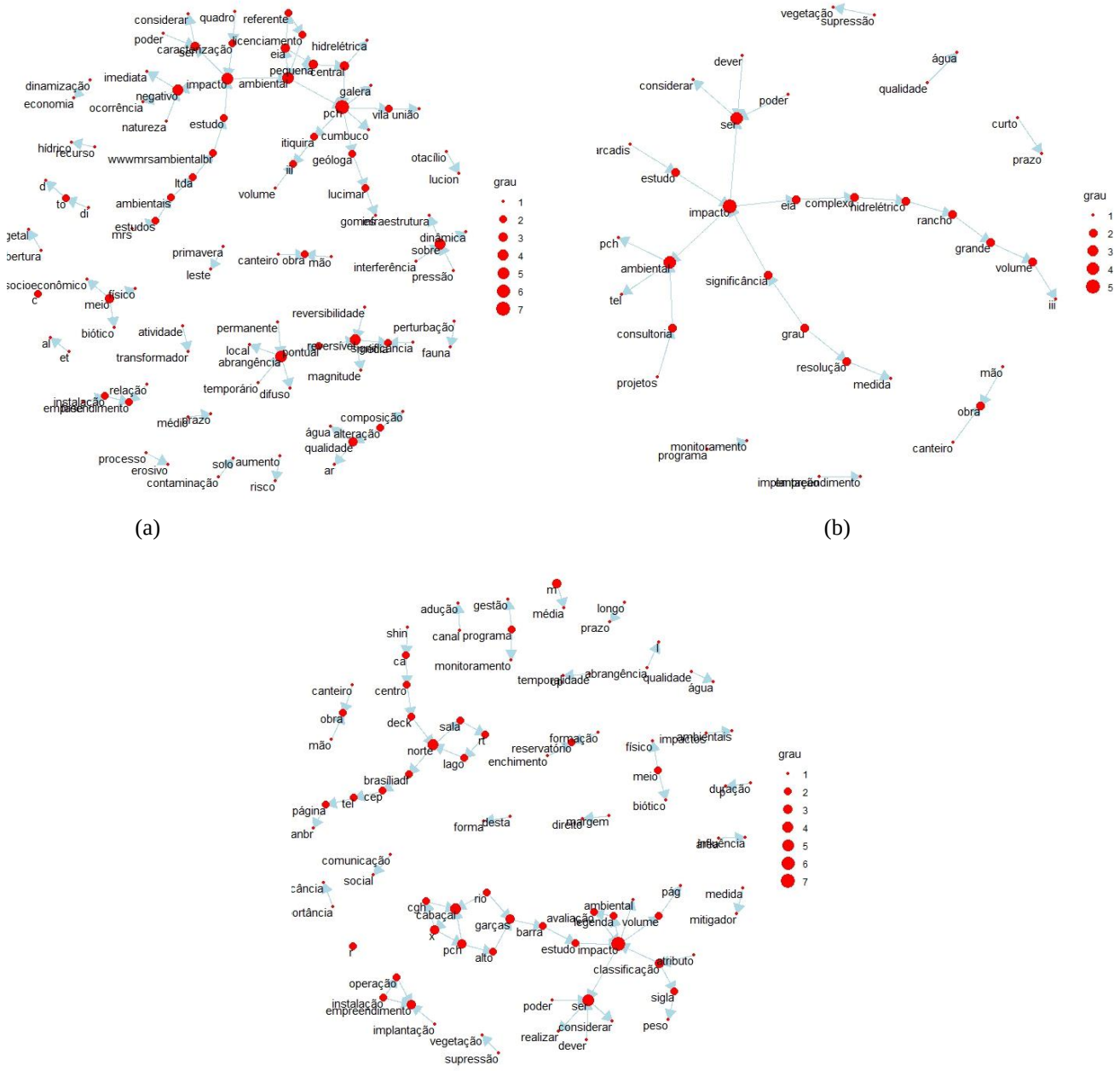
(c)

Fonte: Autores 2024.

Essas redes ajudam a visualizar a importância e inter-relação de conceitos-chave em cada grupo, fornecendo uma perspectiva mais clara sobre as prioridades e enfoques dos EIAs em cada *cluster*.

Na Figura 5, os grafos de palavras representam as conexões e frequências entre os termos mais comuns nos EIAs dos três *clusters*, ilustrando as diferenças de abordagem entre eles.

Figura 5 - Grafo por *cluster*.



(a) (b)

(c)

Fonte: Autores 2024.

Observa-se que o *Cluster* 1 (Figura 5a) destaca termos como "impacto", "ambiental", "área", "alteração" e "qualidade", indicando uma ênfase na avaliação ampla dos efeitos das PCHs sobre o ambiente natural, especialmente em aspectos como a alteração das condições ambientais e a qualidade dos recursos, com destaque para a água. O termo "reversível" sugere um enfoque na mitigação dos impactos temporários, o que está em linha com a abordagem de Hamilton (2016) sobre a necessidade de minimizar e reverter os impactos em áreas ecologicamente sensíveis, como o Pantanal.

Por outro lado, o *Cluster* 2 (Figura 5b) traz palavras centrais como "programa", "monitoramento", "significância" e "consultoria ambiental", refletindo um foco na implementação de programas de acompanhamento e controle ambiental. Esse enfoque destaca a importância da avaliação da "significância" dos impactos e do desenvolvimento de medidas específicas para gerenciá-los. A inclusão de "consultoria ambiental" indica a presença de especialistas para auxiliar na gestão sustentável, alinhando-se às recomendações de Alho e Silva (2018), que ressaltam a necessidade de monitoramento contínuo e especializado para proteger ecossistemas aquáticos.

Já o *Cluster* 3 (Figura 5c) exibe termos como "estudo", "impacto", "monitoramento", "significância" e "curto prazo". A presença de "curto prazo" e "monitoramento" indica uma preferência por ações rápidas e práticas para mitigar os impactos ambientais, refletindo uma abordagem de resposta imediata. Esse tipo de estratégia é especialmente relevante para regiões como a Amazônia, onde, segundo Castello e Macedo (2016), respostas rápidas são essenciais para reduzir os impactos das atividades humanas sobre a biodiversidade.

Comparando os três *clusters*, observa-se que o *Cluster* 1 adota uma visão ampla e preventiva sobre os impactos ambientais, o *Cluster* 2 foca em programas de controle e monitoramento contínuo, e o *Cluster* 3 prioriza respostas imediatas para a mitigação dos impactos. Essas distinções evidenciam diferentes enfoques nos EIAs, que permitem o desenvolvimento de políticas públicas e estratégias de gestão ambiental adaptadas às especificidades de cada contexto ecológico.

4. CONSIDERAÇÕES FINAIS

A análise de similaridade e clusterização dos EIAs das PCHs revelou três *clusters* com características distintas, refletindo diferentes abordagens para a gestão dos impactos ambientais, onde o *cluster* 1 enfatiza a importância de considerar os efeitos gerais das PCHs sobre o meio ambiente. O Cluster 2 indica uma abordagem detalhada e específica dos impactos ambientais, especialmente em relação à qualidade da água e aos processos erosivos

e o *cluster* 3 ressalta a percepção de que as PCHs deste *cluster* têm um impacto mais direto no uso do solo e na socioeconômica local.

Com base nos resultados obtidos, futuros trabalhos podem aprofundar a análise dos impactos ambientais das PCHs a partir de diferentes perspectivas. Uma das possibilidades é a expansão da análise de textos por meio do uso de técnicas mais avançadas de Processamento de Linguagem Natural (PLN), como modelos de aprendizado profundo. Essas abordagens poderiam refinar a clusterização dos EIAs e identificar padrões semânticos mais complexos, contribuindo para uma melhor compreensão das tendências e lacunas nos estudos ambientais existentes.

Outra possibilidade seria a modelagem de cenários futuros, utilizando técnicas de aprendizado de máquina para prever impactos ambientais sob diferentes cenários de expansão das PCHs e mudanças climáticas. Essa abordagem permitiria simular possíveis efeitos e subsidiar a tomada de decisão para um planejamento mais sustentável.

Por fim, a análise das políticas públicas relacionadas ao licenciamento ambiental das PCHs poderia ser aprofundada para avaliar a efetividade das regulamentações existentes. Isso permitiria propor ajustes nas diretrizes de gestão e mitigação de impactos, garantindo que as medidas adotadas sejam mais adequadas às particularidades identificadas em cada cluster.

Esses resultados destacam a importância de abordagens específicas para a gestão e mitigação dos impactos ambientais das PCHs em cada bacia hidrográfica. A análise detalhada dos *clusters* e das redes de palavras oferece *insights* valiosos para a formulação de políticas ambientais e estratégias de planejamento adaptadas às necessidades e particularidades de cada contexto ecológico. Ao permitir uma compreensão mais profunda das dinâmicas e tendências dos EIAs, esses achados contribuem para o desenvolvimento de práticas que promovam a sustentabilidade e minimizem os impactos negativos dessas operações hidrelétricas no estado de Mato Grosso.

Dessa forma, essas direções de pesquisa podem contribuir significativamente para o fortalecimento do conhecimento sobre os impactos ambientais das PCHs e para o desenvolvimento de estratégias que promovam um planejamento mais sustentável no estado de Mato Grosso.

5. REFERÊNCIAS

ALHO, C. J. R.; SILVA, J. S. V. Effects of severe floods and droughts on wildlife of the Pantanal wetland (Brazil) – A review. **Animals**, v. 8, n. 3, p. 45, 2018. Disponível em: https://doi.org/10.3390/ani2040591. Acesso em: 23 abr. 2025.

ASSOCIAÇÃO BRASILEIRA DE PEQUENAS CENTRAIS HIDRELÉTRICAS (ABRAPCH). **Cenário de PCHs e CGHs no Brasil**. Disponível em: https://abrapch.org.br/o-setor/cenario-de-pchs-e-cghs-no-brasil/. Acesso em: 29 set. 2024.

BAIL, C. A. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. **Proceedings of the National Academy of Sciences**, v. 113, n. 42, p. 11823-11828, 2016. Disponível em: https://doi.org/10.1073/pnas.1607151113. Acesso em: 23 abr. 2025.

BENOIT, K.; MUHR, D.; WATANABE, K. **stopwords: Multilingual Stopword Lists**. Disponível em: https://cran.r-project.org/package=stopwords. Acesso em: 26 mar. 2024.

BLONDEL, V. D.; GUILLAUME, J. L.; LAMBIOTTE, R. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2008, n. 10, p. P10008, 2008. Disponível em: https://doi.org/10.1088/1742-5468/2008/10/P10008. Acesso em: 23 abr. 2025.

BRASIL. Conselho Nacional do Meio Ambiente – CONAMA. **Resolução nº 001, de 23 de janeiro de 1986. Dispõe sobre critérios básicos e diretrizes gerais para o Relatório de Impacto Ambiental (RIMA)**. Diário Oficial da União: seção 1, Brasília, DF, 17 fev. 1986. Disponivel em: file://G:\cnia\conam3\86\001-86.htm. Acesso em: 23 abr. 2025.

BRASIL. Conselho Nacional do Meio Ambiente – CONAMA. Resolução nº 237, de 19 de dezembro de 1997. **Dispõe sobre a revisão e complementação dos procedimentos e critérios utilizados para o licenciamento ambiental**. Diário Oficial da União: seção 1, Brasília, DF, 22 dez. 1997. Disponível em: https://www.ibama.gov.br/sophia/cnia/legislacao/MMA/RE0237-191297.PDF. Acesso em: 30 abr. 2025.

BRASIL. Decreto nº 99.274, de 6 de junho de 1990. **Regulamenta a Lei nº 6.902, de 27 de abril de 1981, e a Lei nº 6.938, de 31 de agosto de 1981, que dispõem, respectivamente, sobre a criação de Estações Ecológicas e Áreas de Proteção Ambiental e sobre a Política Nacional do Meio Ambiente, e dá outras providências**. Diário Oficial da União: seção 1, Brasília, DF, 7 jun. 1990. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto/antigos/d99274.htm. Acesso em 23 abr. 2025.

BRASIL**.** Lei nº 9.074, de 7 de julho de 1995. **Dispõe sobre a outorga e prorrogações das concessões e permissões de serviços públicos**. Diário Oficial da República Federativa do Brasil, Brasília, DF, 10 jul. 1995. Disponível em: https://legislacao.presidencia.gov.br/atos/?tipo=LEI&numero=9074&ano=1995&ato=c1ag3YU5UeJpWTf30. Acesso em: 28 mar. 2024.

CASSIANO, K. K.; CORDEIRO, D. F. Representação Semântica Vetorial para Análise de Similaridade de Documentos Textuais. In: ESCOLA REGIONAL DE INFORMÁTICA DE GOIÁS, 6., 2018, Goiânia. **Anais**... Goiânia: Instituto de Informática, UFG, 2018. p. 362.

CASTELLO, L.; MACEDO, M. N. Large-scale degradation of Amazonian freshwater ecosystems. **Global Change Biology**, v. 22, n. 3, p. 990-1007, 2016. Disponível em: https://doi.org/10.1111/gcb.13173. Acesso em: 23 abr. 2025.

CLAUBERG, A. P. C.; HENKES, J. A.; BECEGATTO, V. A. Fontes hídricas: setor energético brasileiro e o incremento das pequenas centrais hidrelétricas. **Revista Brasileira de Meio Ambiente & Sustentabilidade**, v. 1, n. 4, p. 134–174, 2021. Disponível em: https://rbmaes.emnuvens.com.br/revista/article/view/95. Acesso em: 10 abr. 2024.

CSÁRDI, G.; NEPUSZ, T**. igraph: Network Analysis and Visualization**. 2006. Disponível em: https://cran.r-project.org/package=igraph. Acesso em: 12 jun. 2024.

DURAN, M.; NUNES, M. G. V.; LOPES, L.; PARDO, T. A. S. Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo Universal Dependencies em língua portuguesa. **Domínios de Linguagem**, v. 16, n. 4, p. 1608-1643, 2022. Disponível em: https://doi.org/10.14393/DL52-v16n4a2022-13. Acesso em: 23 abr. 2025.

FEINERER, I.; HORNIK, K**. tm: Text Mining Package**. Disponível em: https://cran.r-project.org/package=tm. Acesso em: 26 mar. 2024.

FELDMAN, R.; SANGER, J. **The text mining handbook: advanced approaches in analyzing unstructured data**. Cambridge: Cambridge University Press, 2006. Disponível em: https://archive.org/details/textmininghandbo0000feld. Acesso em: 9 fev. 2025.

FORTUNATO, S. Community detection in graphs. **Physics Reports**, v. 486, n. 3-5, p. 75-174, 2010. Disponível em: https://doi.org/10.1016/j.physrep.2009.11.002. Acesso em: 23 abr. 2025.

GOLDBERG, Y.; LEVY, O**. Neural Network Methods for Natural Language Processing**. 1. ed. San Rafael: Morgan & Claypool Publishers, 2014. 159 p.

HAMILTON, S. K**.** Hydrological controls of ecological structure and function in the Pantanal wetland (Brazil). In: JUNK, W. J.; DA SILVA, C. J.; NUNES DA CUNHA, C.; WANTZEN, K. M. (Ed.). **The Pantanal: Ecology, biodiversity, and sustainable management of a large neotropical seasonal wetland**. Sofia: Pensoft Publishers, 2016. p. 133–158.

HUSSEIN, R.; ALI, A.; MOHAMED, A. Document Clustering Based on Firefly Algorithm. **Journal of Computer Science**, v. 11, n. 3, p. 453–465, 2015. Disponível em: https://thescipub.com/abstract/jcssp.2015.453.465. Acesso em: 30 abr. 2025.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition**. 3. ed. rascunho. Stanford: Stanford University, 2025. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 30 abr. 2025.

LATRUBESSE, E., ARIMA, E., DUNNE, T. *et al.* Damming the rivers of the Amazon basin. **Nature**, v. 546, n. 7658, p. 363-369, 2017. Disponível em: https://doi.org/10.1038/nature22333. Acesso em: 23 abr. 2025.

LE, Q. V.; MIKOLOV, T**.** Distributed Representations of Sentences and Documents. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 31., 2014, Beijing. ***Anais***... Beijing: PMLR, 2014. v. 32, n. 2, p. 1188-1196.

LIU, P.; GUO, J.; ZHANG, X.; SUN, M**.** A Survey on Deep Learning for Named Entity Recognition. **IEEE Transactions on Knowledge and Data Engineering**, v. 34, n. 1, p. 50-70, 2022. Disponível em: https://doi.org/10.1109/TKDE.2020.2992548. Acesso em: 23 abr. 2025.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient Estimation of Word Representations in Vector Space. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 1., 2013, Scottsdale. **Anais**... Scottsdale: ICLR, 2013.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de textos**. Goiânia: Instituto de Informática, Universidade Federal de Goiás, 2007. (Relatório Técnico – INF_005/07). Disponível em: https://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf. Acesso em: 30 abr. 2025.

MYERS, L.; SIROIS, M. Spearman Correlation Coefficients, Differences Between. **Encyclopedia of Statistical Sciences**, 2006. Disponível em: https://doi.org/10.1002/0471667196.ess5050.pub2. Acesso em: 23 abr. 2025.

PEDERSEN, T. L. **ggraph: An Implementation of Grammar of Graphics for Graphs and Networks**. 2024. Disponível em: https://cran.r-project.org/package=ggraph. Acesso em: 12 jun. 2024.

PEDERSEN, T. L. **tidygraph: A Tidy API for Graph Manipulation**. 2024. Disponível em: https://cran.r-project.org/package=tidygraph. Acesso em: 12 jun. 2024.

PONS, P.; LATAPY, M**.** Computing Communities in Large Networks Using Random Walks. **Journal of Graph Algorithms and Applications**, v. 10, n. 2, p. 191-218, 2006. Disponível em: https://doi.org/10.7155/jgaa.00124. Acesso em: 23 abr. 2025.

QUISPE, L.; TOHALINO, J.; AMANCIO, D**.** Using Virtual Edges to Improve the Discriminability of Co-Occurrence Text Networks. **Physica A: Statistical Mechanics and its Applications**, v. 562, p. 125-344, 2021. Disponível em: https://doi.org/10.1016/j.physa.2020.125344. Acesso em: 23 abr. 2025.

R CORE TEAM**.** ***R: A language and environment for statistical computing***. Versão 4.1.2. Viena: R Foundation for Statistical Computing, 2023. Disponível em: https://www.r-project.org/. Acesso em: 10 mai. 2023.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J**.** Learning Representations by Back Propagating Errors. **Nature**, v. 323, p. 533-536, 1986. Disponível em: https://doi.org/10.1038/323533a0. Acesso em: 23 abr. 2025.

SILGE, J.; ROBINSON, D. **Text mining with R: a tidy approach**. Sebastopol: O'Reilly Media, 2017. Disponível em: https://www.tidytextmining.com. Acesso em: 30 abr. 2025.

STEINBACH, M.; KARYPIS, G.; KUMAR, V. **A comparison of document clustering techniques**. Minneapolis: University of Minnesota, Department of Computer Science, 2000. Disponível em: https://conservancy.umn.edu/handle/11299/215421. Acesso em: 30 abr. 2025.

TRAAG, V. A.; WALTMAN, L.; VAN ECK, N. J. From Louvain to Leiden: Guaranteeing Well-Connected Communities. **Scientific Reports**, v. 9, n. 1, 2019. Disponível em: https://doi.org/10.1038/s41598-019-41695-z. Acesso em: 23 abr. 2025.

TURIAN, J.; RATINOV, L.; BENGIO. Word Representations: A Simple and General Method for Semi-Supervised Learning. In: **Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics**, p. 384–394, Uppsala, Suécia, julho de 2010. Association for Computational Linguistics. Disponível em: https://aclanthology.org/P10-1040/. Acesso em: 30 abr. 2025.

WIJFFELS, J. **Doc2Vec: Distributed Representations of Sentences, Documents and Topics**. 2021. Disponível em: https://cran.r-project.org/package=Doc2Vec. Acesso em: 12 jun. 2024.

WIJFFELS, J. **Udpipe Natural Language Processing - Basic Analytical Use Cases**. 2023. Disponível em: https://cran.r-project.org/package=udpipe. Acesso em: 12 jun. 2024.

WIJFFELS, J.; WATANABE, K. **Word2Vec: Distributed Representations of Words**. 2023. Disponível em: https://cran.r-project.org/package=Word2Vec. Acesso em: 12 jun. 2024.

ŽIŽKA, J.; DAŘENA, F.; SVOBODA, A. **Text Mining with Machine Learning: Principles and Techniques**. Boca Raton: CRC Press, 2019.