UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Felipe Barcelos Ribeiro

ClusterPub: um sistema para clusterização de artigos científicos

Uberlândia, Brasil

2024

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Felipe Barcelos Ribeiro

ClusterPub: um sistema para clusterização de artigos científicos

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Anderson Rodrigues dos Santos

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2024

Felipe Barcelos Ribeiro

ClusterPub: um sistema para clusterização de artigos científicos

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 19 de outubro de 2024:

Prof. Dr. Anderson Rodrigues dos Santos
Orientador

Prof. Dr. Daniel Duarte Abdala

Professor Dr. Marcelo Zanchetta Nascimento

Uberlândia, Brasil
2024

Agradecimentos

Gostaria de primeiro agradecer imensamente a Deus, visto que, sem Ele não conseguiria ter dado nem mesmo o primeiro passo dessa extensa jornada.

Também gostaria de agradecer grandemente à minha família, composta, pela minha avó querida, Sandra, minha doce mãe, Tatiane e meu querido tio Nilson Júnior, porque em todos os momentos que precisei me apoiaram e suas contribuições foram vitais para que eu pudesse concluir o curso.

Outrossim, não poderia deixar de agradecer ao meu orientador Anderson pelos seus direcionamentos e comentários que me ajudaram bastante na confecção desse projeto.

Resumo

Os processos de pesquisa bibliográfica são extremamente comuns na vida acadêmica, visto que, para a confecção de alguns trabalhos, como, projetos de conclusão de curso, dissertações de mestrado e teses de doutorado, uma parte fundamental do processo de elaboração é a busca por referencial bibliográfico. Entretanto, tal processo atualmente é bastante laboroso, pois, ao realizar uma busca em repositórios acadêmicos, como, IEEE Xplore, Google Acadêmico e Pubmed, milhares de resultados são obtidos, o que faz com que o pesquisador precise executar uma tarefa manual de organização, classificação e filtragem dos artigos retornados, tarefa essa, que demanda muito tempo e esforço. Este trabalho tem como proposta a confecção de uma aplicação de linha de comando que seja capaz de processar arquivos bibliográficos, resultando na geração de dendogramas que reflitam as similaridades presentes entre os trabalhos contidos no arquivo processado, afim, de agilizar o processo de levantamento bibliográfico. Durante a confecção do corrente trabalho foi desenvolvida a ferramenta proposta, utilizando a linguagem de programação Python, juntamente com o *framework* para desenvolvimento de aplicações de linha de comando, Typer, além de terem sido usadas diversas bibliotecas, como, Scikit-Learn e Scipy, para confeccionar o algoritmo de agrupamento. Para a realização dos testes foi utilizado um arquivo bibliográfico no formato BibTex composto por 25 artigos de cada um dos seguintes temas: Inteligência Artificial, Biotecnologia, Economia Circular, Mudanças Climáticas, Sistemas Complexos, Genética, Saúde Mental e Neurociência. A combinação entre o método de ligação da média ponderada das distâncias e a medida de distância da similaridade dos cossenos pode ser considerada como a que obteve os melhores resultados de maneira geral, visto que, resultou nos valores de 0,8545, 118,0987 e 0,5394 para os índices de Davies-Bouldin, Calinski-Harabasz e de silhueta, respectivamente. Ao analisar os valores enumerados no corrente texto frente aos resultados obtidos por outros trabalhos também relacionados a modelos de clusterização textual, pode-se concluir que os resultados aferidos pelo corrente trabalho são satisfatórios, visto que, por vezes são observados valores numericamente melhores. Entretanto, não é plausível utilizar essas comparações para afirmar que a ferramenta ClusterPub é superior aos trabalhos utilizados nos comparativos citados, pois, esse trabalho não realizou testes com as bases de dados usadas pelas outras ferramentas, sendo esses artigos utilizados apenas com o intuito de obter valores de referência para as métricas analisadas. A ferramenta desenvolvida foi disponibilizada para instalação no repositório público de pacotes Python, PyPi.

Palavras-chave: Levantamento Bibliográfico, Arquivos Bibliográficos, Aplicação de Linha de Comando, Aprendizado de Máquina, Clusterização.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CLI	*Command-Line-Interface*
IA	Inteligência Artificial
RIS	*Research Information Systems*
ASCII	*American Standard Code for Information Interchange*
TF-IDF	*Term Frequency - Inverse Document Frequency*
CSV	*Comma-Separated Values*
IEEE	*Institute of Electrical and Electronic Engineers*
TCC	Trabalho de Conclusão de Curso
CCIDF	*Common Citation Inverse Document Frequency*
KSON	*Kohonen's Self-Organiz- ing Map*

Sumário

1 Introdução

De acordo com (TERRA, 2023) a oferta de cursos de pós-graduação tem crescido em média 9% ao ano, mediante a maior procura dos estudantes por especializações em suas respectivas áreas. Um dos motivos para o aumento na busca por esses cursos é o contínuo crescimento de conhecimento, além do incentivo financeiro, visto que, profissionais com especialização tendem a ter um salário maior e posições de maior destaque.

Durante toda a jornada acadêmica, alunos se deparam com a necessidade de confeccionarem trabalhos que requerem a realização de revisões e consultas bibliográficas, principalmente durante a produção de seus trabalhos de conclusão de curso, elaboração de dissertações de mestrado, teses de doutorado e escrita de artigos científicos, por serem trabalhos de natureza mais rebuscada e profunda, visto que, o aluno possui a oportunidade de escolher o tema que deseja abordar. Entretanto, nesses momentos muitos encontram dificuldades, pois como atualmente há um grande volume de artigos/trabalhos científicos já publicados, encontrar material que esteja corretamente relacionado ao tema escolhido se torna um processo trabalhoso e cansativo, por requerer que o estudante avalie manualmente um grande acervo em busca do material mais apropriado.

Além da dificuldade intrínseca ao processo de busca, também há a dificuldade imposta pelos próprios repositórios de trabalhos acadêmicos, como IEEE Xplore [1], Google Acadêmico [2] e PubMed [3], pois não fornecem uma busca mais precisa, apenas retornam os documentos que contenham as palavras presentes na busca digitada e atendam às condições de filtro inseridas, o que na grande maioria dos casos resulta em milhares de resultados compatíveis.

Com o intuito de melhorar a experiência dos pesquisadores que estão realizando seus trabalhos, é proposto o sistema CLI ClusterPub. O uso do sistema se pauta na indicação de caminhos para arquivos bibliográficos , os quais podem ser exportados pela grande maioria dos repositórios de trabalhos científicos, e que são compostos por dados, como título, resumo e autoria dos documentos retornados pela busca efetuada. O *software* produz como resultado um arquivo contendo a árvore de proximidade de artigos, a qual indica de maneira hierárquica quais são os trabalhos que estariam mais relacionados com a linha de pesquisa do usuário mediante o arquivo analizado.

O sistema ClusterPub foi desenvolvido em linguagem de programação Python com a utilização do *framework* para desenvolvimento de aplicações CLI Typer. Para a geração da árvore de relacionamentos é executado um algoritmo de agrupamento hierárquico

[1] <https://ieeexplore.ieee.org/Xplore/home.jsp>
[2] <https://scholar.google.com.br>
[3] <https://pubmed.ncbi.nlm.nih.gov>

baseado nas distâncias existentes entre os termos presentes nos trabalhos advindos do arquivo bibliográfico. O algoritmo de clusterização de artigos científicos foi escrito utilizando algumas bibliotecas, como Scikit-Learn[4] para a representação vetorial dos artigos, SciPy[5] para a execução do algoritmo de agrupamento e Matplotlib[6] para a exibição do dendograma resultante.

1.1 Objetivo

O projeto ClusterPub tem como objetivo desenvolver uma ferramenta CLI que seja capaz de auxiliar pesquisadores em seus processos de levantamento bibliográfico, através da aplicação de técnicas de clusterização sobre artigos científicos, com o intuito de explicitar os relacionamentos semânticos presentes entre os trabalhos analisados, otimizando assim os processos de pesquisa, ao ajudar na identificação de trabalhos que possuem conteúdos relacionados à área de estudo.

1.2 Exemplos de Uso

Nesta seção serão abordados possíveis cenários nos quais seria apropriado utilizar a ferramenta ClusterPub.

1.2.1 Revisão Bibliográfica para TCC

Um aluno de gradução precisa fazer uma revisão bibliográfica sobre Inteligência Artificial para a confecção do seu trabalho de conclusão de curso. Ele realiza uma busca no IEEE Xplore[7] e exporta os resultados em formato BibTeX. Com centenas de artigos, o processo manual de filtragem e organização seria demorado.

Usando a ferramenta ClusterPub, ele processa o arquivo BibTeX em alguns segundos, gerando um dendograma que agrupa os artigos com base em suas similaridades. Dessa forma, ele pode identificar rapidamente quais artigos são mais relevantes para o seu tema de pesquisa, otimizando o tempo gasto na seleção dos materiais.

1.2.2 Organização de Fontes em Pesquisa Colaborativa

Uma equipe de pesquisadores trabalhando em um projeto interdisciplinar (envolvendo saúde mental e neurociência) coleta artigos de diferentes repositórios, como IEEE

[4] <https://scikit-learn.org/stable/index.html>
[5] <https://scipy.org/>
[6] <https://matplotlib.org/>
[7] <https://ieeexplore.ieee.org/Xplore/home.jsp>

Xplore[8] e PubMed[9], exportando os arquivos em diferentes formatos (BibTeX e NBIB). Cada membro da equipe tem diferentes focos de interesse, e o número de artigos reunidos é muito grande para ser analisado manualmente.

A ferramenta ClusterPub é usada para agrupar automaticamente os artigos por similaridade temática. O resultado, exibido em forma de dendograma, ajuda a equipe a identificar subgrupos de artigos e definir quais textos são prioritários para cada subárea, facilitando a divisão das leituras e a organização da pesquisa colaborativa.

[8] <https://ieeexplore.ieee.org/Xplore/home.jsp>
[9] <https://pubmed.ncbi.nlm.nih.gov>

2 Fundamentação Teórica

Neste capítulo serão abordados os conceitos teóricos utilizados para a confecção do corrente projeto, como, medidas de distância, métodos de clusterização, técnicas de vetorização de textos, de redução de dimensionalidade e métricas de avaliação de modelos de agrupamentos, além de enumerar a linguagem de programação utilizada e explicar o formato dos arquivos bibliográficos que serão tratados.

2.1 Conceitos Teóricos

Nesta seção serão abordados os conceitos teóricos que serviram de base para o desenvolvimento do projeto ClusterPub, como, *word embeddings*, redução da dimensionalidade, técnicas de agrupamento, medidas de distância e métricas para avaliação de clusterização.

2.1.1 *Word Embeddings*

Word Embeddings é um conjunto de técnicas que tem ajudado no desenvolvimento da área de Processamento de Linguagem Natural, sendo utilizado em diversas operações que involvem conteúdo textual, como, classificação, clusterização e geração de textos. Essas técnicas se pautam na representação de palavras/documentos como vetores numéricos multidimensionais, cujas dimensões e orientação refletem o significado semântico das palavras, juntamente com suas informações contextuais (BARNARD, 2024).

Alguns exemplos de técnicas existentes são: Hash2Vec, Matriz de Termos e Documentos e TF-IDF.

2.1.1.1 Matriz de Termos e Documentos

A Matriz de Termos e Documentos é uma técnica de *embedding* que se baseia na contagem das frequências dos termos de um vocabulário nos documentos de uma coleção individualmente, com o intuito de estabelecer um relacionamento entre os documentos e os termos (BAEZA-YATES, 2013a), sendo que quanto maior a frequência de um termo em um documento, mais relevante é esse termo para descrever o documento.

2.1.1.2 TF-IDF

TF-IDF é uma técnica de *embedding* que se pauta na ponderação dos termos presentes em uma coleção de documentos, a qual é o produto entre a frequência dos termos nos documentos de maneira isolada juntamente com a especifidade do termo, que pode

ser descrita como a frequência do termo na coleção de maneira geral, sendo que quanto maior a quantidade de documentos nos quais o termo aparece, menor a sua especificidade (BAEZA-YATES, 2013b).

2.1.1.3 Hash2Vec

Hash2Vec é uma técnica de *embedding* que se pauta na aplicação de uma função *hash* sobre os termos presentes na coleção, para subsequentemente gerar a representação em um espaço vetorial multidimensional. Como a vetorização é realizada a partir dos valores *hash* dos termos originais e suas respectivas informações de contexto, não é necessário que haja uma etapa de treinamento do algoritmo, visto que, pelo caráter determinístico das funções *hash*, termos iguais resultarão em *hashes* idênticos (GAIKWAD, 2020).

2.1.2 Redução da Dimensionalidade

Algoritmos de redução da dimensionalidade se pautam na conversão de conjuntos de dados de alta dimensionalidade, ou seja, que estão considerando um grande número de atributos, para conjuntos de baixa dimensionalidade, com o intuito de remover atributos redundantes ou irrelevantes, preservando as variáveis de interesse para o problema em análise (JIA MEILI SUN, 2022). Um desafio presente em conjuntos de dados de alta dimensionalidade é a maldição da dimensionalidade, a qual, se pauta no fato de que a cada dimensão adicionada, o volume do espaço representado cresce exponencialmente, resultando em algumas dificuldades, como, maior tempo de processamento e visualização de dados dificultosa (AWAN, 2023).

A prática de redução da dimensionalidade é comum em áreas da computação,como, Aprendizado de Máquina, Mineração de Dados e Reconhecimento de Padrões. Exemplos de algoritmos de redução da dimensionalidade são: PCA e SVD.

2.1.2.1 PCA

PCA, Análise de Componentes Principais, em português, é um algoritmo de redução da dimensionalidade que reduz o número de dimensões de conjuntos de dados mantendo o máximo de informações do conjunto original. O corrente algoritmo se pauta no resumo de informações de um grande *dataset* em um pequeno grupo de variáveis não correlatas, chamadas de Componentes Principais, os quais são combinações lineares das variáveis originais que possuem a maior variância, quando comparada com outras possíveis combinações.

Essa técnica é utilizada na área de Aprendizado de Máquina, normalmente em etapas de pré-processamento de dados, visto que, pode resultar na otimização do processamento dos dados, o que é advindo da redução da quantidade de dimensões. PCA

pode ser utilizado em tarefas de *Machine Learning*, como, Reconhecimento de Padrões, Processamento de Imagens e Processamento de Sinais (IBM, 2023).

2.1.2.2 SVD

SVD, Decomposição de Valores Singulares, em português, é uma técnica de redução da dimensionalidade, que se pauta na fatorização de um conjunto de dados representados por uma matriz, no produto de três matrizes, as quais são representadas pelas seguintes letras:U, Σ, V^T. As matrizes U, V^T são ortogonais que configuram os vetores singulares da esquerda e da direita, respectivamente, e a matriz diagonal Σ abriga os chamados valores singulares, os quais ajudam a preservar as características da matriz original (ALBRIGHT, 2004).

Também há um método derivado chamado de *TruncatedSVD*, o qual, se pauta na seleção dos K maiores valores singulares (BARUAH, 2023). Tal técnica é usada em diversos campos computacionais, como: Compressão de Dados, Redução de Ruídos e Sistemas de Recomendação.

2.1.3 Técnicas de Clusterização

Técnicas de clusterização se baseiam no agrupamento de conjuntos de dados de maneira que os elementos presentes em um determinado grupo se assemelham mais entre si do que com elementos presentes em outros grupos, sem conhecimento prévio a respeito da estrutura e classificação do conjunto de dados, se enquadrando assim, na área de aprendizado não supervisionado (K.KAMESHWARAN, 2014). Exemplos de algoritmos de clusterização são: Agrupamento Hierárquico e K-Means.

2.1.3.1 Agrupamento Hierárquico

Técnicas de agrupamento se baseiam na formação de grupos cujos pontos são semelhantes entre si, sendo os conjuntos hierárquicos formados por pontos que serão mesclados repetidamente até se tornarem um único agrupamento, o que ocorre através de uma função de ligação, a qual, indicará a similaridade entre os pontos, baseando-se em uma medida de distância, por exemplo, a Distância Euclidiana (FERREIRA et al., 2020a).

Como resultado do processo de agrupamento hierárquico é possível obter dois tipos de representação hierárquica: implícita e explícita. A representação hierárquica implícita é representada por um diagrama de Venn, composto por todos os membros do conjunto analisado, agrupados em subconjuntos, enquanto a representação hierárquica explícita é ilustrada por um dendograma, o qual a partir de seus níveis horizontais indica a hierarquia e a ordem de realização dos agrupamentos encontrados.

Para a realização do agrupamento hierárquico podem ser aplicados diferentes algoritmos, os quais tem como função identificar os grupos que serão formados, o que é feito a partir do cálculo das distâncias entre os elementos dos conjuntos (FERREIRA et al., 2020b). Os principais métodos utilizados são:

- **Método da ligação simples**: Também conhecido como o método do vizinho mais próximo, define a similaridade entre dois agrupamentos como a menor distância existente entre os pares de elementos dos dois grupos, o que faz com que tenha pouca tolerância à presença de *outliers*, visto que, tende a incluí-los em algum grupo existente (FERREIRA et al., 2020b).
- **Método da ligação completa**: Também conhecido como o método do elemento mais distante, define a similaridade entre dois grupos como a distância entre os elementos mais distantes de cada conjunto, o que pode acarretar na formação de agrupamentos com grande dissimilaridade (FERREIRA et al., 2020b).
- **Método das médias das distâncias**: Define a similaridade entre dois conjuntos como a média das distâncias presentes entre todos os pares de elementos dos dois grupos, o que faz com que esse método seja menos sensível aos *outliers* (FERREIRA et al., 2020b) e gere agrupamentos com quantidades similares. Também existe um método derivado, o qual se pauta na utilização da média ponderada das distâncias, com o objetivo de dar a todos os elementos um peso igual no cálculo das similaridades.

2.1.3.2 K-Means

K-means é um algoritmo de clusterização baseado em particionamento que tem como objetivo agrupar um conjunto de dados em um número predefinido de grupos, K, de maneira que os elementos de um grupo sejam mais similares entre si, do que com componentes de outros agrupamentos, gerando assim *clusters* com a maior distinção possível ao escolher o número de *clusters* desejado de maneira otimizada (ASHABI SHAMSUL BIN SAHIBUDDIN, 2021).

O algoritmo implementado pelo K-Means pode ser descrito pelos seguintes passos:

1. Escolha aleatória de K elementos do conjunto como pontos centrais dos agrupamentos.
2. Alocação de cada um dos elementos do conjunto ao *cluster*, cujo ponto central se encontra mais próximo, de acordo com alguma medida de distância, como a Distância Euclidiana.

3. Após a alocação de todos os elementos a um agrupamento, é calculado o ponto médio do agrupamento a partir das posições de todos os seus componentes.

4. Repete-se os passos 2 e 3 até que se encontrem os mesmos *clusters* em iterações consecutivas ou um número máximo de repetições seja atingido.

2.1.4 Medidas de Distância

Medidas de distância podem ser definidas como funções que calculam as distâncias/similaridades entre pares de elementos presentes em um conjunto, as quais são utilizadas como parâmetro para a realização de agrupamentos de dados (MERCIONI, 2019). Algumas medidas de distância utilizadas são: Distância Euclidiana, Similaridade dos Cossenos e Correlação de Pearson.

2.1.4.1 Distância Euclidiana

A distância euclidiana é um conceito advindo da geometria euclidiana que diz respeito ao cálculo da menor distância entre dois pontos em um plano multidimensional (BRAZ, 2020). Essa medida é utilizada em áreas da Ciência da Computação relacionadas a algoritmos de busca, classificação e clusterização (PIUBELLO, 2023).

A distância euclidiana é definida pela seguinte equação:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.1}$$

Os componentes da equação definida acima são:

- $\mathbf{x}$ **e** $\mathbf{y}$: pontos em um espaço multidimensional.
- $\mathbf{n}$: número de dimensões existentes em um determinado espaço multidimensional.
- $\mathbf{x}_i$: valor da i-ésima coordenada do ponto x.
- $\mathbf{y}_i$: valor da i-ésima coordenada do ponto y.

2.1.4.2 Similaridade dos Cossenos

A similaridade dos cossenos é uma medida utilizada para mensurar a similaridade entre dois vetores em um espaço multidimensional, a partir do cosseno do ângulo presente entre os mesmos, considerando não apenas os valores dos vetores, mas também as suas direções (MIESLE, 2023). Essa medida tem o seu uso indicado em tarefas relacionadas à análise textual, pois sofre pouca influência de alta dimensionalidade, característica comum em representações vetoriais de textos (MARCACINI; MOURA; REZENDE, 2009).

A similaridade dos cossenos pode ser calculada pela seguinte fórmula:

$$Cosine(x, y) = \frac{x \cdot y}{|x||y|} \tag{2.2}$$

Os elementos da fórmula acima são:

- **x e y**: vetores em um espaço multidimensional.
- **|x|**: tamanho do vetor x.
- **|y|**: tamanho do vetor y.

2.1.4.3 Correlação de Pearson

A correlação de Pearson é utilizada para mensurar a relação presente entre duas variáveis lineares, indicando a força e direção do relacionamento. Os valores resultantes da correlação variam entre -1 e 1, sendo que um resultado negativo indica uma correlação negativa, 0 demonstra que não há correlação e valores positivos referem-se a correlações positivas. Correlações negativas ocorrem quando os valores das variáveis se associam de maneira inversa, enquanto correlações positivas se dão quando os valores das variáveis se associam de maneira direta (DAWAR, 2020). Uma das principais aplicações da corrente medida na área computacional é a seleção de variáveis correlatas em conjuntos de dados.

A correlação de Pearson pode ser obtida pela seguinte fórmula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.3}$$

Os componentes da expressão acima são:

- **x e y**: variáveis lineares observadas.
- **n**: número que representa a quantidade de amostras disponíveis de cada variável.
- $\mathbf{x}_i$: valor da i-ésima amostra da variável x.
- $\mathbf{y}_i$: valor da i-ésima amostra da variável y.
- $\bar{\mathbf{x}}$: média dos valores obtidos para a variável x.
- $\bar{\mathbf{y}}$: média dos valores obtidos para a variável y.

2.1.5 Métricas para Avaliação de Clusterização

Para a avaliação de modelos de clusterização existem algumas métricas que podem ser calculadas com o intuito de identificar a qualidade dos agrupamentos obtidos. Dois tipos de técnicas existentes são: Avaliação Interna e Avaliação Externa. As métricas de avaliação interna consideram apenas os dados de entrada e o resultado obtido, ou seja, sem referenciar nenhuma classificação prévia do conjunto de dados, por exemplo, Índice de Silhueta, Índice de Calinski-Harabasz e Índice de Davies-Bouldin. Já os métodos de avaliação externa focam na comparação entre os grupos formados pelo modelo de agrupamento com classificações prévias dos dados de entrada, sendo um exemplo desse tipo de técnica, o Índice de Rand (ALLA, 2021).

2.1.5.1 Índice de Silhueta

O Índice de Silhueta é uma métrica utilizada para avaliar os agrupamentos formados por um modelo de clusterização, analisando a distância de separação entre os *clusters* resultantes.

O Índice de Silhueta pode variar de -1 a 1, sendo um resultado negativo uma indicação que os dados podem ter sido associados a *clusters* de maneira errônea, enquanto valores em volta de 1 indicam que os agrupamentos estão razoavelmente separados e valores em cerca de 0 apontam para *clusters* sobrepostos, ou seja, a corrente métrica busca avaliar o quanto que os elementos de um *cluster* são similares entre si e dissimilares de outros agrupamentos (VYSALA; GOMES, 2020).

O Índice de Silhueta obtido para um agrupamento é resultado da média dos índices de silhueta calculados para todos os elementos analisados, que, pode ser obtido através da seguinte equação:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.4}$$

Os componentes da equação acima são:

- **a(i)**: distância média de um ponto i com relação a todos os elementos do seu *cluster*, a.
- **b(i)**: distância média entre um ponto i e seu *cluster* vizinho mais próximo, b.

2.1.5.2 Índice de Calinski-Harabasz

O Índice de Calinski-Harabasz é uma métrica utilizada para avaliar o resultado de modelos de agrupamento, o qual, se pauta na análise do grau de dispersão presente entre os *clusters*, e no nível de coesão de cada grupo formado. A análise dessa métrica

se pauta no fato de que, quanto maior o resultado, melhor foi a execução do modelo de clusterização, pois, indica uma maior separação entre os agrupamentos e uma maior similaridade entre os dados de um mesmo *cluster* (WANG, 2019).

O corrente índice pode ser calculado a partir da seguinte fórmula:

$$CH = \frac{Tr(B)}{Tr(W)} \cdot \frac{n-k}{k-1} \tag{2.5}$$

Os elementos da fórmula acima são:

- **n**: Número total de elementos.
- **k**: Número de *clusters* formados.
- **Tr(B)**: Covariância do nível de dispersão *inter-cluster*.
- **Tr(W)**: Covariância do nível de dispersão *intra-cluster*.

2.1.5.3 Índice de Davies-Bouldin

O Índice de Davies-Bouldin é utilizado para avaliar a qualidade de um modelo de agrupamento através da análise da coesão e da separação. A coesão se pauta na soma das distâncias entre todos os elementos de um *cluster* e o ponto central do agrupamento, enquanto, separação, se pauta nas distâncias entre os pontos centrais de cada *cluster* (MUGHNYANTI, 2020), sendo que quanto mais próximo de 0 for o valor obtido para a corrente métrica, melhor foi a execução do método de agrupamento (ASHARI ROMANTIKA BANJARNAHOR, 2022), pois, neste caso indica que há uma separação considerável entre os agrupamentos e que os elementos dos *clusters* possuem uma alta similariade entre si.

O Índice de Davies-Bouldin pode ser obtido a partir da seguinte expressão:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{S_i + S_j}{D_{ij}} \tag{2.6}$$

A expressão acima é composta pelos itens abaixo:

- **k**: número total de *clusters*.
- **Si**: medida de dispersão interna do *cluster* i.
- **Sj**: medida de dispersão interna do *cluster* j.
- **Dij**: distância entre os pontos centrais dos *clusters* i e j.

2.1.5.4 Índice de Rand

O Índice de Rand é uma métrica de avaliação externa que busca comparar a concordância entre dois *clusters*, e não necessita que os agrupamentos em análise possuam a mesma quantidade de elementos. Caso haja uma independência completa entre os grupos o valor do corrente índice será 0, mas, se existir uma associação completa o valor será 1 (KRIEGER, 1999), sendo valores próximo a 1 indicadores de que o modelo de agrupamento gerou grupos similares aos previamente definidos.

O Índice de Rand pode ser calculado a partir da seguinte equação:

$$RandIndex = \frac{a+b}{a+b+c+d} \tag{2.7}$$

Os componentes da equação acima são:

- **a**: Número de elementos corretamente agrupados e que pertencem ao mesmo grupo (verdadeiros positivos).
- **b**: Número de elementos corretamente agrupados e que pertencem a grupos diferentes (verdadeiros negativos).
- **c**: Número de elementos incorretamente agrupados e que pertencem ao mesmo grupo (falsos positivos).
- **d**: Número de elementos incorretamente agrupados e que pertencem a grupos distintos (falsos negativos).

2.1.6 Dendogramas

Dendogramas são gráficos que se assemelham a gráficos de árvores e que tem como objetivo representar uma estrutura hierárquica, sendo compostos por um nó raíz conectado a nós subordinados (IBM, 2024).

Um exemplo de dendograma pode ser visualizado na FIGURA 1 abaixo:

2.2 Tecnologias de Desenvolvimento

Nesta seção serão abordadas as tecnologias utilizadas para o desenvolvimento da ferramente ClusterPub, como a linguagem de programação e o *framework* utilizados.

2.2.1 Python

Python é uma linguagem de programação interpretada, orientada a objetos e que possui uma sintaxe de fácil aprendizagem (GEEKS, 2024c). Muito em razão de sua sim-

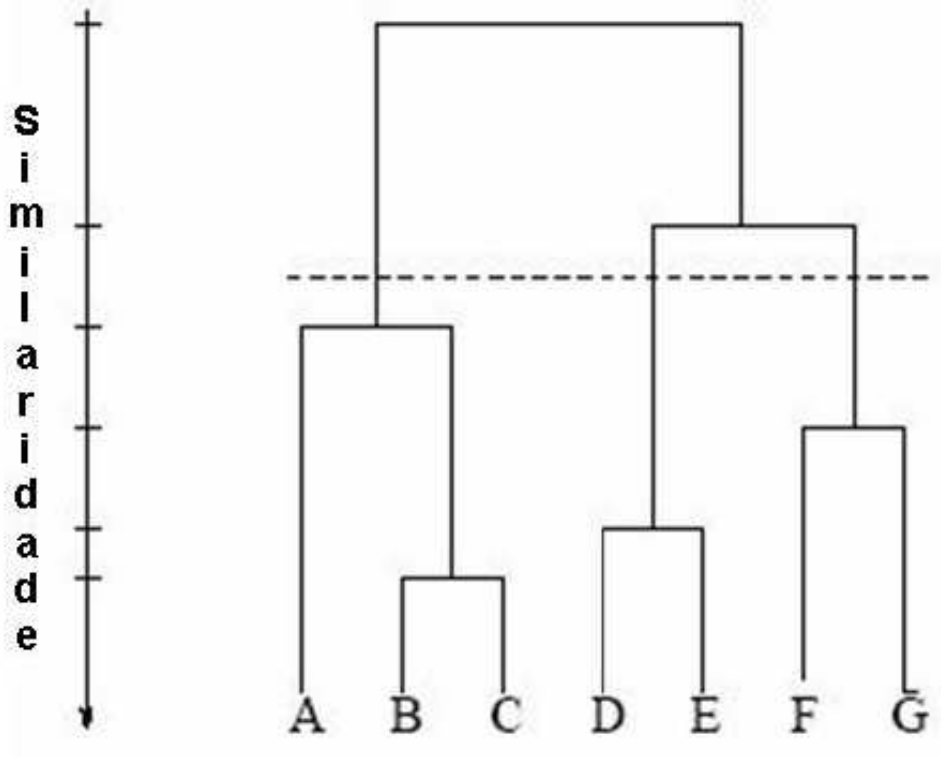

Figura 1 – Exemplo de Dendograma. Fonte: (CARVALHO, 2006)

plicidade e existência de pacotes especializados na área de inteligência artificial, essa linguagem tem se tornado a mais popular entre os programadores, o que pode ser visto na pesquisa realizada pelo IEEE Spectrum, *Top Programming Languages* 2023 [1].

Entretanto, a corrente linguagem não se destaca apenas na área de IA, mas também tem ganhado muito espaço nos campos de desenvolvimento *web*, através de *frameworks*, como Django, Flask e FastAPI e na área de aplicações CLI com bibliotecas, como Typer.

As principais bibliotecas utilizadas durante o desenvolvimento da aplicação foram:

- **bibtexparser** [2]: Biblioteca utilizada para a realização de *parsing* de arquivos bibliográficos em formato Bibtex.
- **rispy** [3]: Biblioteca utilizada para a realização de *parsing* de arquivos bibliográficos em formato RIS.
- **nbib** [4]: Biblioteca utilizada para a realização de *parsing* de arquivos bibliográficos em formato NBIB.

[1] https://spectrum.ieee.org/the-top-programming-languages-2023
[2] https://bibtexparser.readthedocs.io/en/main/index.html
[3] https://pypi.org/project/rispy/
[4] https://pypi.org/project/nbib/

- **matplotlib** [5]: Biblioteca responsável pela representação visual do dendograma resultante do agrupamento hierárquico.
- **scipy** [6]: Biblioteca utilizada para a execução do algoritmo de agrupamento hierárquico e para a produção das coordenadas referentes ao dendograma resultante.
- **scikit-learn** [7]: Biblioteca utilizada para a representação vetorial dos artigos científicos, redução da dimensionalidade e cálculo das distâncias entre os elementos presentes na representação vetorial.

2.2.2 Typer

A biblioteca de desenvolvimento de aplicações de linha de comando Typer se destaca das demais por possibilitar a conversão de simples *scripts* Python em CLIs, as quais geram automaticamente comandos de ajuda quanto às opções e argumentos dos comandos (RAMíREZ, 2024), além de também possibilitar a adição de recursos gráficos, como barras de progresso e mensagens coloridas.

2.3 Arquivos Bibliográficos

Arquivos bibliográficos são arquivos que possuem dados analíticos, como data de publicação, autoria, título e resumo, referentes a itens bibliográficos (LUND, 2023).

2.3.1 Arquivos bibliográficos no formato BibTex

O formato de arquivo BibTeX é um dos padrões na área de armazenamento e compartilhamento de conteúdo bibliográfico mediante à sua facilidade de manipulação (PAPERPILE, 2022). Uma referência bibliográfica em um arquivo BibTeX é composta por três partes principais:

- **Tipo de entrada**: Indica qual o tipo de material bibliográfico está sendo referenciado. Existem 14 tipos de entrada, entre eles: livro, artigo, conferência e etc.
- **Chave de citação**: Identifica unicamente uma entrada, visto que, podem haver mais de uma entrada do mesmo tipo.
- **Pares chave-valor**: Após o tipo de entrada e a chave de citação há um conjunto de pares chave-valor que representa os dados bibliográficos armazenados.

[5] https://matplotlib.org/stable/
[6] https://scipy.org/
[7] https://scikit-learn.org/stable/

Um exemplo de entrada no formato BibTex que ilustra as partes mencionadas acima pode ser visto abaixo:

```
@Book{Entry1,
title   = "The Art of Computer Programming",
author  = "Knuth, Donald Ervin",
publisher = "Addison Wesley",
address    = "Boston, MA",
edition  = "3",
year  = "1997"
}
```

No exemplo acima podem ser visualizadas as três partes que compõem uma entrada bibliográfica no formato BibText, as quais são representadas pelos seguintes componentes:

- **@Book**: representa o tipo de entrada.
- **Entry1**: representa a chave da citação.
- **Pares chave-valor**: armazenam o conteúdo da citação.

2.3.2 Arquivos bibliográficos no formato RIS

O formato de arquivo RIS é um formato de arquivo bibliográfico padronizado baseado em *tags* (LIBRARY, 2022). As *tags* representam conjuntos de informações bibliográficas e são caracterizadas por serem compostas por duas letras e se separarem do conteúdo por um hífen. Um exemplo de arquivo bibliográfico no formato RIS pode ser visto abaixo:

```
TY  -  CONF
TI  -  Python programming training with the robot Finch
AU  -  D. Vasilev
```

2.3.3 Arquivos bibliográficos no formato NBIB

O formato de arquivo bibliográfico NBIB foi desenvolvido pela Biblioteca Nacional de Medicina dos Estados Unidos, com o intuito de substituir o formato RIS, visto que, possui suporte para algumas *tags* específicas de trabalhos armazenados no repositório PubMed [8], repositório para o qual o corrente formato foi desenvolvido. A formatação de um arquivo NBIB se pauta em um conjunto de *tags* que são compostas por letras

[8] <https://pubmed.ncbi.nlm.nih.gov/>

maiúsculas separadas do conteúdo bibliográfico por um hífen, semelhante à formatação presente em arquivos RIS (FILEINFO, 2024). Um exemplo de arquivo no formato NBIB pode ser visto abaixo:

```
PMID  -  28754806
OWN   -  NLM
STAT  -  MEDLINE
```

3 Trabalhos Relacionados

Neste capítulo serão abordados os trabalhos relacionados à área de estudo do projeto ClusterPub.

Mediante o grande número de trabalhos científicos publicados em diversas áreas, muitos projetos foram desenvolvidos com o intuito de possibilitar a extração de informações estruturadas, a partir de textos científicos planos, como mostram (FRISONI GIANLUCA MORO, 2016), ao realizarem uma grande revisão de projetos que tem como objetivo a aplicação de técnicas de Processamento de Linguagem Natural sobre artigos científicos relacionados a biomedicina, que tem como finalidade extraírem eventos biológicos descritos de maneira textual. De acordo com (SILVA BERNARDO PEREIRA NUNES, 2017) esforços na área de compartilhamento de conteúdo educacional têm sido realizados visando melhorar a interoperabilidade, acesso e recomendação de materiais acadêmicos. Entretanto, ainda existem muitos desafios que impedem uma maior integração desses materiais, como, dificuldade para processar grandes volumes de dados e a falta de padronização entre as muitas bases de dados disponíveis. Para auxiliar na disseminação de conteúdo acadêmico relacionado à área da saúde (MULUNDA PETER WAIGANJO, 2020) desenvolveram uma aplicação *web* que tem como objetivo auxiliar na realização de pesquisas educacionais, através da aplicação de técnicas de Aprendizado de Máquina, como, classificação e clusterização, sobre textos científicos, para os categorizar de acordo com os tópicos mais frequentes nas pesquisas dos usuários e na coleção de artigos analisada.

Também existem ferramentas que se pautam na análise de citações presentes em trabalhos acadêmicos, como a desenvolvida por (HE S.C. HUI, 2003), a qual, realiza um processo de indexação de trabalhos publicados em repositórios de artigos acadêmicos. Tal processo se baseia na aplicação de uma técnica de clusterização chamada KSON, derivada do algoritmo K-Means, definindo a similaridade presente entre os trabalhos através de suas citações em comum. Os agrupamentos formados são utilizados no processo de busca realizada pelo usuário, ao serem utilizados para identificar em qual *cluster* estão os artigos mais relacionados aos termos buscados. Após a determinação do agrupamento é calculada a Distância Euclidiana entre o vetor que representa a busca e os artigos presentes no *cluster* selecionado, o que é feito com o fim de definir quais são os trabalhos mais relevantes para o usuário. Outra ferramenta que se pauta no uso de referências é o sistema *web* Inciteful [1], o qual busca facilitar os processos de levantamento bibliográfico através da construção de grafos de similaridade, que são compostos por trabalhos que citam ou são citados pelos artigos de referência informados pelos usuários (INCITEFUL, 2020). O trabalho

[1] <https://inciteful.xyz/>

desenvolvido por (HUYNH KIEM HOANG, 2012) também baseia-se na recomedação de artigos científicos pela similaridade aferida através das citações presentes nos trabalhos analisados. Para calcular a proximidade entre os artigos é aplicado o algoritmo CCIDF, derivado do método TF-IDF, baseando-se na frequência das citações e não dos termos.

Há projetos que focam na criação de interfaces gráficas que auxiliem no processo de levantamento bibiliográfico, como o desenvolvido por (ANGIONI ANGELO SALATINO, 2022), o qual foca na criação de *dashboards* pelos quais é possível visualizar as principais informações a respeito de conferências realizadas na área de Ciência da Computação, como, autores, instituições e tópicos presentes em uma determinada conferência. Para a determinação dos tópicos dos trabalhos é aplicado um algoritmo de classificação não-supervisionado, o qual se pauta na análise sintática e semântica dos resumos e títulos dos trabalhos, lançando mão da técnica de representação vetorial Word2Vec para ajudar na detecção de termos relacionados aos tópicos de interesse. Outro trabalho que visa auxiliar os processos de pesquisa através da construção de gráficos é o desenvolvido por (MELO, 2005), visto que, busca criar uma representação gráfica semelhante a um mapa topológico composto pelos autores e trabalhos mais relacionados ao tema de interesse do usuário. Semelhantemente, o projeto conduzido por (TANG YANGYONG ZHU, 2019) foca na construção de mapas geográficos espaçotemporais, que representam os tópicos relacionados a uma determinada linha de pesquisa ordenados temporalmente, sendo que os temas que coocorrem no mesmo período são agrupados nas mesmas regiões espaciais do mapa.

Outra linha de trabalhos existente é a que se concentra na concepção de ferramentas baseadas em dados colaborativos, como o sistema *web* desenvolvido por (OUZZANI HOSSAM HAMMADY, 2016), o qual consiste na realização de revisões bibliográficas colaborativas e sistemáticas, que utiliza técnicas de Inteligência Artificial para calcular a relevância de artigos para linhas de pesquisa, a partir de análises feitas pelos próprios usuários.

3.1 Clustering de Artigos Científicos em uma Ferramenta Inteligente de Apoio à Pesquisa

De acordo com (MELO, 2005) o trabalho em questão tem como objetivo desenvolver uma ferramenta que auxilie pesquisadores e alunos a realizarem suas pesquisas acadêmicas. A aplicação desenvolvida se pauta na obtenção automática de trabalhos científicos relacionados ao tema de pesquisa presentes na *web*, os quais subsequentemente passarão por um pré-processamento (remoção de *stopwords* e *stemming*), para que possa ser aplicada uma técnica de *clustering*, resultando assim em uma representação gráfica, semelhante a um mapa topográfico, que indicará quais os autores e trabalhos mais re-

levantes da área de interesse. O projeto ClusterPub se relaciona com o citado trabalho, visto que, também se pauta na aplicação de técnicas de agrupamento hierárquico sobre um conjunto de artigos científicos.

3.2 Inciteful

Inciteful [2] é um sistema *web* que busca facilitar o processo de levantamento bibliográfico, a partir da construção de grafos de similaridade entre artigos para auxiliar na identificação de trabalhos correlatos relevantes. Os grafos de relacionamento são gerados a partir de *seed papers*, que seriam os artigos submetidos pelos usuários para a busca de trabalhos relacionados, podendo realizar a submissão apenas informando o título do trabalho, endereço eletrônico ou pela importação de um arquivo biliográfico no formato BibTex. No resultado estarão presentes os artigos citados pelos trabalhos submetidos, juntamente com os trabalhos que citam os *seed papers*. Sobre os resultados poderão ser aplicados filtros relacionados à data de publicação, termos presentes no título e nível de similaridade (INCITEFUL, 2020). Entretanto, os artigos associados no grafo de similaridade não são relacionados por similaridade semântica, mas, são agrupados por citarem ou serem citados pelos *seed papers*, o que pode acarretar na formação de grafos que possuem elementos com uma pequena correlação semântica, ou seja, não ajudando a encontrar trabalhos que são parecidos em termos de conteúdo.

Portanto, é possível visualizar que há uma correlação entre o sistema Inciteful com o projeto ClusterPub, visto que, ambos se pautam na identificação de correlações entre artigos científicos a partir da submissão de trabalhos acadêmicos por parte dos usuários. Entretanto, esse trabalho se diferencia da ferramenta ClusterPub pelo fato de não realizar uma análise semântica dos trabalhos e os agrupar apenas pelas conexões que possuem com os *seed papers*.

3.3 Rayyan

Rayyan [3] é um sistema *web* focado na realização de revisões bibliográficas sistemáticas de maneira colaborativa, visto que, permite que vários colaboradores revisem um mesmo conjunto de trabalhos com o fim de decidir quanto a relevância de cada item para a pesquisa que está sendo realizada. A aplicação oferece uma funcionalidade de *screening*, a qual se refere à geração de um resumo de um artigo, composto pelos principais tópicos identificados, além de destacar os trechos mais relevantes do texto, o que é feito a partir de técnicas de aprendizagem de máquina. Graças às técnicas de inteligência artificial também é possível solicitar que a plataforma calcule a relevância dos artigos submetidos, baseado

[2] <https://inciteful.xyz/>
[3] <https://www.rayyan.ai/>

nas análises de relevância feitas pelos usuários (OUZZANI HOSSAM HAMMADY, 2016). Contudo, pode-se afirmar que este sistema possui a limitação de que as suas análises de relevância se pautam unicamente nas contribuições realizadas pelos usuários, e não pondera a similaridade semântica dos documentos, o que pode resultar em recomendações de artigos que não possuem conteúdos razoavelmente similares.

O sistema Rayyan se relaciona com o projeto ClusterPub, pois ambos são voltados para a gestão de artigos científicos, além de realizarem a análise da relevância dos trabalhos submetidos, mas, o sistema em questão se diferencia do trabalho ClusterPub, pois, as suas indicações de relevância são explícitas e calculadas para artigos individualmente, enquanto os resultados gerados pela ferramenta ClusterPub possuem indicações de relevância implícitas, visto que, a importância dos artigos é dada pelas distâncias presentes entre os trabalhos contidos no dendograma resultante.

4 Método

Neste capítulo serão abordadas as etapas e os métodos utilizados para a confecção do projeto ClusterPub.

O sistema CLI foi escrito utilizando a linguagem de programação Python (GEEKS, 2024c) juntamente com o *framework* para desenvolvimento de aplicações de linha de comando Typer (RAMíREZ, 2024). O *framework* é responsável por receber e tratar os comandos digitados pelos usuários e internalizá-los para a camada responsável pela clusterização dos artigos presentes no arquivo bibliográfico escolhido.

A geração das árvores de relacionamento entre os artigos analizados foi desenvolvida com a utilização de bibliotecas escritas em Python, como Scikit-Learn [1] para a representação numérica dos artigos científicos, SciPy [2] para a realização do agrupamento hierárquico e Matplotlib [3] para a visualização gráfica do resultado.

As principais etapas presentes no desenvolvimento do sistema são: leitura dos arquivos bibliográficos, pré-processamento dos textos, representação vetorial dos artigos, clusterização dos trabalhos analizados e exibição do dendograma resultante. As citadas etapas e suas respectivas ordens de execução podem ser visualizadas na FIGURA 2 abaixo:

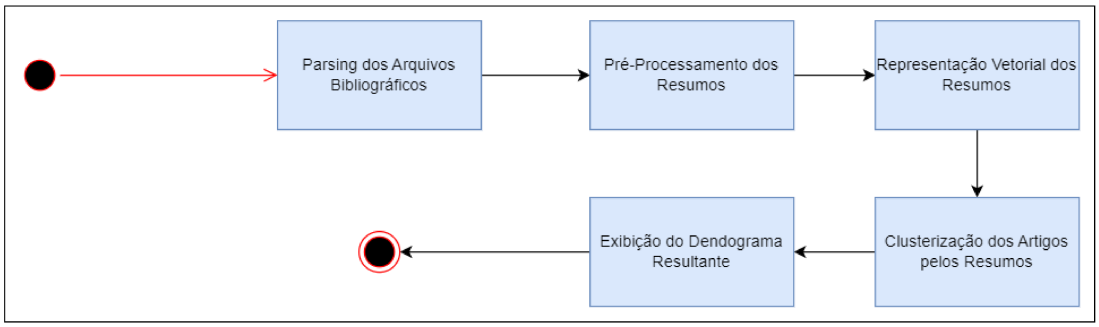

Figura 2 – Diagrama de Fluxo de Execução

4.1 Leitura dos Arquivos Bibliográficos

Mediante a formatação específica presente em cada tipo de arquivo bibliográfico foi necessária a confecção de um módulo, que fosse responsável por realizar o *parsing* do conteúdo presente no formato textual para uma estrutura de dados mais fácil de ser manipulada computacionalmente. Para tanto foram utilizadas bibliotecas escritas em lingua-

[1] <https://scikit-learn.org/stable/index.html>
[2] <https://scipy.org/>
[3] <https://matplotlib.org/>

gem Python, bibtexparser[4], rispy[5] e nbib[6], as quais processam os arquivos bibliográficos de formatação BibTex (PAPERPILE, 2022), RIS (LIBRARY, 2022) e NBIB (FILEINFO, 2024), respectivamente, e os convertem em uma lista de dicionários, os quais representam os trabalhos contidos no arquivo informado. As formatações bibliográficas aceitas são: BibTex, RIS e NBIB. Após o *parsing* do arquivo é realizada uma etapa de filtragem, que se pauta na eliminação dos dados que não serão utilizados durante o processo de clusterização, mantendo apenas os títulos e resumos dos trabalhos presentes no arquivo processado.

Para usufruir de um maior reuso de código foram utilizados os padrões de projeto *Template Method* e *Factory Method*.

O padrão *Template Method* é um padrão comportamental utilizado quando a estrutura do algoritmo é a mesma para diferentes implementações, diferindo apenas em alguns passos, os quais deverão ser especificados pelas especializações (GEEKS, 2024b), o que fez esse padrão ser adequado para ser utilizado no projeto, visto que, o único passo do processo de leitura dos arquivos que é diferente entre as formatações bibliográficas é o *parsing* do documento, função que foi implementada nas classes responsáveis por realizar o processamento de cada tipo de arquivo bibliográfico suportado.

Já o padrão *Factory Method* é um padrão criacional que auxilia na diminuição do acoplamento entre a criação de objetos e sua utilização, pois encapsula toda a lógica concernente ao processo de instanciação, além de facilitar a extensão do sistema, porque todos os objetos instanciados implementarão a mesma interface (GEEKS, 2024a), sendo utilizado no desenvolvimento desse módulo para abstrair a instanciação das classes que seriam responsáveis por realizar a leitura dos diferentes tipos de arquivos bibliográficos. A definição da classe que será instanciada para realizar o processamento do arquivo informado se dá a partir da detecção da formatação do documento, que é inferida pela extensão do mesmo, permitindo assim um mapeamento entre as extensões bibliográficas suportadas e as suas respectivas classes de *parsing*.

A estrutura de classes empregada nesse módulo pode ser vista na FIGURA 3 abaixo:

4.2 Pré-Processamento dos Textos

Após a realização do *parsing* dos arquivos bibliográficos foi executado um processo de normalização do conteúdo afim de gerar uma matriz de similaridades mais confiável. O pré-processamento realizado teve como referência o definido por (GOMES, 2023),

[4] <https://bibtexparser.readthedocs.io/en/main/index.html>
[5] <https://pypi.org/project/rispy/>
[6] <https://pypi.org/project/nbib/>

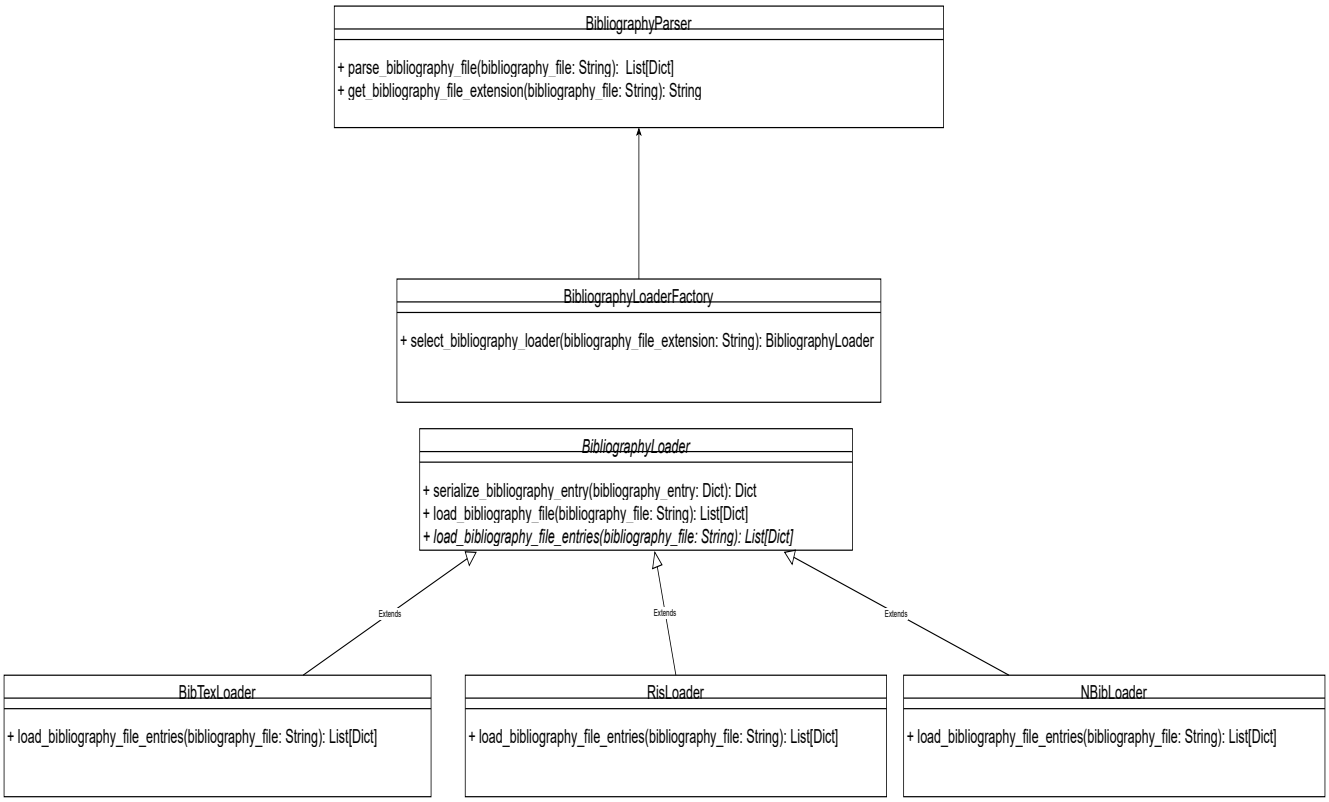

Figura 3 – Diagrama de Classes

pautando-se na capitalização dos textos por completo, juntamente com a remoção de todos os caracteres que não são alfanuméricos, como pontuação e espaços em branco, além de decodificar caracteres Unicode para ASCII. Também é realizada a conversão de caracteres numéricos para caracteres alfabéticos, com o intuito de possibilitar a detecção da similaridade entre documentos, a partir da presença de números, como constantes matemáticas e físicas. O processo de normalização de textos empregado é composto pelos seguintes passos:

1. Remoção de caracteres que não são alfanuméricos.
2. Decodificação de caracteres Unicode para ASCII.
3. Tradução de caracteres numéricos para alfabéticos.
4. Capitalização integral do texto.

Para otimizar a execução da etapa de pré-processamento os passos acima são realizados sobre os resumos dos trabalhos de maneira paralela, a partir da utilização de multi-processamento, ou seja, distribuindo a carga de trabalho entre múltiplas unidades de processamento.

Para calcular as similaridades presentes entre os artigos são normalizados apenas os resumos, por possuírem todas as informações relevantes de seus respectivos conteúdos e terem um tamanho reduzido, o que melhora o desempenho da aplicação.

4.3 Representação Vetorial dos Artigos

Após o pré-processamento dos resumos dos artigos científicos é preciso representá-los de maneira numérica/vetorial para que os cálculos de similaridades possam ser realizados, o que é feito analisando os trios de caracteres dos resumos já normalizados. Preferiu-se realizar a análise a nível de carácter, visto que, ao processar caracteres é possível dispensar a aplicação de técnicas de pré-processamento textual que possuem foco na análise de palavras, como *stemming* [7], a qual se pauta na redução de palavras aos seus respectivos radicais.

A técnica de vetorização empregada é a implementação do algoritmo Hash2Vec (GAIKWAD, 2020) presente na biblioteca Scikit-Learn [8], escolha essa motivada pelo fato do citado algoritmo não requerer uma etapa de treinamento, o que resulta em um menor tempo de execução. Além da economia de memória computacional advinda da não utilização de um vocabulário de termos em memória, o que ocorre nas implementações dos algoritmos TF-IDF[9] (BAEZA-YATES, 2013b) e Matriz de Termos e Documentos[10] (BAEZA-YATES, 2013a), proporcionando assim uma melhor escalabilidade quanto ao tamanho da coleção de artigos.

Após a vetorização, com o intuito de otimizar e tornar o processo subsequente de agrupamento mais preciso, é aplicado o algoritmo TruncatedSVD (BARUAH, 2023) presente na biblioteca Scikit-Learn [11], visto que, segundo a documentação o supracitado algoritmo é mais adequado para lidar com matrizes esparsas, como as retornadas pela classe de vetorização HashingVectorizer, a qual, está sendo utilizada no corrente projeto. Definiu-se a manutenção dos 5 maiores valores singulares para a redução da dimensionalidade, com o intuito de diminuir o ruído dos dados sem prejudicar o tempo de execução do sistema.

---

7 <ibm.com/topics/stemming>

8 <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.HashingVectorizer.html>

9 <https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html>

10 <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>

11 <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html>

4.4 Clusterização dos Artigos Científicos

O processo de clusterização dos artigos científicos ocorre em duas etapas. Primeiramente é realizada a confecção de uma matriz de distâncias, a qual é construída a partir do cálculo da similaridade dos cossenos (MARCACINI; MOURA; REZENDE, 2009), entre todos os pontos existentes na representação vetorial dos trabalhos, utilizando a biblioteca Scikit-Learn [12].

A segunda etapa do processo de clusterização se pauta na aplicação do algoritmo de agrupamento hierárquico implementado no pacote Scipy [13], sobre a matriz de distâncias obtida no passo anterior. O método de agrupamento hierárquico está configurado para utilizar o método da média ponderada das distâncias, como função de ligação, e a similaridade dos cossenos como medida de distância.

O método de agrupamento hierárquico (FERREIRA et al., 2020a) foi escolhido devido ao fato de não necessitar de uma predefinição da quantidade de agrupamentos que serão formados, o que é adequado para o contexto do projeto ClusterPub, visto que, o usuário pode não ter conhecimento a respeito da quantidade de classes existentes no arquivo bibliográfico analisado.

4.5 Exibição do Dendograma Resultante

Para permitir a visualização dos resultados advindos do processo de clusterização, é realizada a confecção de um dendograma (IBM, 2024), o qual reflete a estrutura hierárquica dos relacionamentos semânticos existentes entre os resumos analisados. O processo para a obtenção do dendograma se pauta em dois passos. O primeiro passo consiste no cálculo das coordenadas do gráfico, as quais são determinadas pela função *dendogram* presente no módulo Scipy [14], e que recebe como parâmetro de entrada uma matriz que representa o resultado do algoritmo de agrupamento hierárquico. Finalmente, é feita a renderização e salvamento do arquivo que contém o dendograma resultante, o que é realizado através da biblioteca Matplotlib[15].

[12] <https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.pairwise.cosine_distances.html>
[13] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html>
[14] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html>
[15] <https://matplotlib.org/>

5 Resultados e Discussões

Neste capítulo serão abordados os resultados colhidos durante o desenvolvimento do projeto ClusterPub.

Os resultados se pautam na análise dos valores médios de determinadas métricas de avaliação interna de agrupamentos, as quais são: Índice de Silhueta, Índice de Davies-Bouldin e Índice de Calinski-Harabasz.

Preferiu-se lançar mão da avaliação interna, visto que, não necessita de uma classificação prévia dos dados, o que é adequado para o contexto do projeto, pois, técnicas de clusterização se enquadram na área de Aprendizado de Máquina Não-Supervisionado, ou seja, não se utilizam de categorizações feitas antecipadamente.

As métricas citadas acima serão representadas no formato de tabelas com o intuito de relacionar a medida de distância e o método de ligação utilizados, afim de encontrar a combinação que resulte nos melhores valores médios dos índices analisados. Os valores apresentados são referentes à média de 10 cálculos de um determinado índice de avaliação.

Os supracitados índices foram calculados utilizando as suas respectivas funções de cálculo presentes na biblioteca Scikit-Learn[1].

Os resultados que serão discutidos detalhadamente nas subsequentes seções foram obtidos a partir de um arquivo bibliográfico no formato BibTex composto por resultados de consultas feitas no repositório IEEE Xplore [2]. Foram selecionados 25 artigos de cada um dos temas pesquisados, os quais foram: Inteligência Artificial, Biotecnologia, Economia Circular, Mudanças Climáticas, Sistemas Complexos, Genética, Saúde Mental e Neurociência. O arquivo de teste citado pode ser conferido no seguinte endereço eletrônico: <https://github.com/barcelosf/cluster_pub/blob/master/sample_files/similarity_test.bib>.

Para o cálculo dos índices exibidos abaixo considerou-se que deveriam existir no dendograma resultante 8 *clusters*, ou seja, um agrupamento para cada um dos temas pesquisados.

No corrente capítulo também será disponibilizado um exemplo de dendograma gerado pela ferramenta ClusterPub.

[1] <https://scikit-learn.org/stable/api/sklearn.metrics.html>
[2] <https://ieeexplore.ieee.org/Xplore/home.jsp>

5.1 Análise do Índice de Silhueta

Os valores obtidos para o Índice de Silhueta podem ser observados na tabela abaixo:

Método de Ligação / Medida de Distância	Ligação Simples	Ligação Completa	Média das Distâncias	Média Ponderada das Distâncias
Distância Euclidiana	0,2784	0,3020	0,3483	0,3505
Similaridade dos Cossenos	-0,2277	0,3903	0,4917	0,5394

Tabela 1 – Valores médios calculados para o Índice de Silhueta.

Os valores do Índice de Silhueta apresentados na tabela estão associados às combinações de métodos de ligação e medidas de distância utilizadas. Os maiores valores presentes na tabela, 0,5394, 0,4917 e 0,3903, foram obtidos através da utilização dos métodos de ligação da média ponderada das distâncias, média das distâncias e ligação completa, respectivamente, juntamente com a similaridade dos cossenos, como medida de distância. Entretanto, ao combinar a similaridade dos cossenos com o método da ligação simples observa-se o valor mínimo da tabela, -0,2277.

Os valores advindos da utilização da distância euclidiana são menores que os relacionados a similaridade dos cossenos, quando usada em conjunto com os métodos de ligação da média ponderada das distâncias (0,3505), média das distãncias (0,3483) e ligação completa (0,3020), resultando em valores superiores apenas quando combinada com o método da ligação simples, obtendo o valor de 0,2784.

O Índice de Silhueta se pauta na análise da distância presente entre os elementos de um agrupamento e a proximidade entre os elementos de um *cluster* e o seu grupo vizinho mais próximo, sendo resultados negativos classificados como ruins, pois, indicam que elementos foram associados erroneamente, e, valores próximos de 1 apontam para a formação de *clusters* coesos. Portanto, ao visualizar a tabela é possível concluir que os métodos que resultaram nos melhores valores para o índice em análise foram, os métodos da média das distâncias e média ponderada das distâncias, o que ocorre em decorrência de que os *clusters* formados por essas funções de ligação são compostos por elementos que não possuem uma grande distância média entre si, pois, são agrupados dados que possuem uma distância média razoavelmente pequena.

Os piores valores foram observados a partir da utilização do método da ligação simples, o que pode ser explicado pela característica dessa função de ligação de incluir *outliers*, elementos muito afastados do conjunto de dados, nos agrupamentos mais próximos, o que aumenta a distância entre os itens de um mesmo conjunto, diminuindo assim

o valor do Índice de Silhueta.

5.2 Análise do Índice de Calinski-Harabasz

Os valores obtidos para o Índice de Calinski-Harabasz podem ser observados na tabela abaixo:

Método de Ligação / Medida de Distância	Ligação Simples	Ligação Completa	Média das Distâncias	Média Ponderada das Distâncias
Distância Euclidiana	42,5273	68,2040	101,1095	102,7183
Similaridade dos Cossenos	25,4937	91,3149	111,0379	118,0987

Tabela 2 – Valores médios calculados para o Índice de Calinski-Harabasz

Os valores do Índice de Calinski-Harabasz apresentados na tabela estão associados às combinações de métodos de ligação e medidas de distância utilizadas. Os maiores valores do índice em análise, 118,0987, 111,0379 e 91,3149 foram obtidos a partir do emprego da similaridade dos cossenos com os métodos de ligação da média ponderada das distâncias, média das distâncias e ligação completa, respectivamente. Entretanto, ao combinar a similaridade dos cossenos com o método da ligação simples é observado o valor mínimo da tabela, o qual é 25,4937.

Todos os valores advindos da utilização da distância euclidiana são menores que os relacionados à similaridade dos cossenos, quando utilizada juntamente com os métodos de ligação da média ponderada das distâncias (102,7183), média das distâncias (101,1095) e ligação completa (68,2040), ultrapassando a medida alternativa quando usada em conjunto com o método da ligação simples, obtendo o valor de 42,5273.

O Índice de Calinski-Harabasz se pauta na análise da dispersão interna dos agrupamentos e da disperão geral presente nos *clusters* formados, sendo que, quanto maior o valor obtido para o corrente índice melhor foi a execução do algoritmo de clusterização. De acordo com a tabela pode-se concluir que o método de ligação que originou os piores resultados para esse índice foi o método da ligação simples, o que pode ser explicado pelo fato de que, como esse método é sensível a *outliers*, é possível que agrupamentos contenham elementos distantes de seu ponto central, aumentando por consequência, o grau de variância/dispersão dos *clusters*, resultando em menores valores.

Os melhores valores observados foram resultados da utilização dos métodos de ligação da média das distâncias e da média ponderada das distâncias, mas, ao utilizar a similaridade dos cossenos em conjunto com esses métodos, os resultados observados foram consideravelmente melhores do que os obtidos com o uso da distância euclidiana, sendo

uma razão para tal comportamento, o fato de que a distância euclidiana é uma medida mais sensível à magnitude dos dados, o que pode acarretar na formação de *clusters* com alta variância, por conterem elementos distantes dos centroides de seus agrupamentos, o que não ocorre com o emprego da similaridade dos cossenos, pois, considera mais a direção do que os valores reais dos elementos, resultando em conjuntos mais coesos e compactos.

Os métodos de ligação da média das distâncias e da média ponderada das distâncias apresentaram melhores resultados, porque ao considerarem as médias das distâncias presentes entre os elementos analisados, os *clusters* formados tendem a ser mais coesos e compactos, ou seja, seus componentes não estão muito distantes entre si, o que acarreta em um menor grau de dispersão interna, além de proporcionar agrupamentos que estão relativamente longe uns dos outros.

5.3 Análise do Índice de Davies-Bouldin

Os valores obtidos para o Índice de Davies-Bouldin podem ser observados na tabela abaixo:

Método de Ligação / Medida de Distância	Ligação Simples	Ligação Completa	Média das Distâncias	Média Ponderada das Distâncias
Distância Euclidiana	0,7603	1,2024	0,7710	0,7861
Similaridade dos Cossenos	0,9213	1,0669	0,8728	0,8545

Tabela 3 – Valores médios calculados para o Índice de Davies-Bouldin

Os valores do Índice de Davies-Bouldin apresentados na tabela estão associados às combinações de métodos de ligação e medidas de distância utilizadas. Os menores valores do índice em análise foram obtidos a partir da utilização da distância euclidiana, sendo o valor mínimo, 0,7603, alcançado quando combinada com o método da ligação simples, enquanto, os dois subsequentes menores valores, 0,7710 e 0,7861, foram resultados da utilização dos métodos da média das distâncias e da média ponderada das distâncias, respectivamente. Entretanto, quando a distância euclidiana foi combinada com o método da ligação completa foi obtido o maior valor da tabela, 1,2024.

Ao se utilizar a similaridade dos cossenos como medida de distância obtêm-se valores maiores do que os referentes à distância euclidana, ao mesclá-la com o método da ligação simples (0,9213), método da média das distâncias (0,8728) e método da média ponderada das distâncias (0,8545), obtendo um valor menor apenas quando usada em conjunto com o método da ligação completa (1,0669).

Ao analisar os valores obtidos para o índice em análise, quanto mais próximo de 0, melhor foi a execução do modelo de agrupamento, pois, indica uma menor dissimilaridade entre os elementos de um mesmo grupo e uma maior separação entre os *clusters* formados. Portanto, pode-se concluir que o método de ligação que obteve os piores resultados foi o método da ligação completa, porque ao utilizá-lo, independentemente da medida de distância, obteve-se os maiores valores da tabela, o que pode ser explicado pela tendência desse método de gerar agrupamentos com alta dissimilaridade, visto que, determina a proximidade entre dois grupos como a distância entre os elementos mais afastados de cada conjunto.

Outra característica relevante é o fato de que não houveram grandes diferenças quanto aos valores referentes aos métodos da média das distâncias e da média ponderada das distâncias, podendo os seus resultados serem classificados como razoáveis, porque ao combinar esses métodos com a distância euclidiana são observados resultados próximos ao valor mínimo da tabela, e a similaridade dos cossenos atinge os seus melhores resultados quando empregada em conjunto com esses métodos. Pode-se explicar tal fenômeno pelo fato de que esses métodos não utilizam elementos extremos, mais distantes ou mais próximos, para determinar a distância entre agrupamentos, o que resulta em grupos com uma dissimilaridade moderada, pois, utilizam a média, ponderada ou simples, das distâncias presentes entre os itens dos conjuntos em análise.

O melhor valor obtido para o Índice de Davies-Bouldin foi advindo do uso da distância euclidiana juntamente com o método da ligação simples, o que pode ser explicado pelo fato de que essa função de ligação tende a gerar agrupamentos com maior coesão, pois, define a similariade entre dois *clusters*, como a distância presente entre os dois itens mais próximos de cada grupo, resultando assim, em *clusters* com elementos mais similares entre si. O método da ligação simples aliado à similaridade dos cossenos resultou em um valor pior do que o relacionado à distância euclidiana, pois, como essa medida pondera mais as direções dos vetores do que os seus valores reais, é possível que tenham sido criados agrupamentos com elementos que possuem uma grande distância entre si, ou seja, que possuem uma pequena similaridade.

5.4 Exemplo de Dendograma Resultante

Um exemplo de dendograma gerado pela ferramenta ClusterPub pode ser visto no fragmento abaixo:

Com o fim de melhorar a visualização foi inserido apenas um fragmento do dendograma resultante completo, visto que, o arquivo original possui uma grande dimensionalidade. O dendograma completo está disponível no seguinte endereço eletrônico: <https://github.com/barcelosf/cluster_pub/blob/master/similarity_test_result.png>.

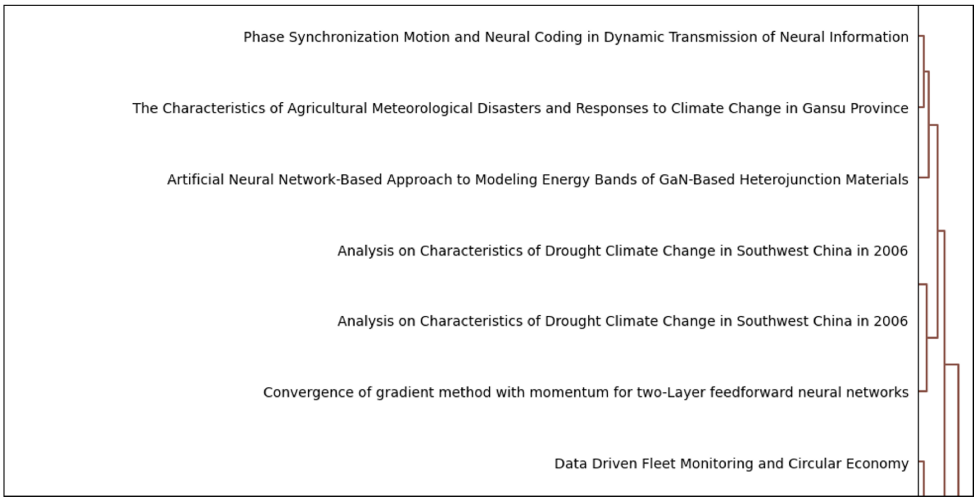

Figura 4 – Exemplo de Dendograma Resultante

O tempo de execução da aplicação para a geração do dendograma mostrado acima, foi em torno de 11 segundos aproximadamente, ao analisar um arquivo composto por 200 resumos de artigos científicos.

6 Conclusões

Ao analisar os valores obtidos para os índices detalhados no capítulo anterior, frente aos resultados presentes em outros trabalhos relacionados a modelos de clusterização, os resultados obtidos pelo corrente trabalho podem ser classificados como satisfatórios. Por exemplo, os valores calculados para os índices de Davies-Bouldin e de silhueta, são numericamente melhores que os divulgados por (YUAN JUSTIN ZOBEL, 2022), cujo melhor valor para o índice de Davies-Bouldin publicado foi de 2,58 e o melhor resultado obtido para o índice de silhueta foi de 0,41. O valor do índice de Calinski-Harabasz citado acima foi aparentemente melhor que o disponibilizado por (BHARGAVA, 2019), o qual, foi de 9,803. Entretanto, é necessário ressalvar que não é possível afirmar que a ferramenta ClusterPub é melhor que os modelos de clusterização citados acima, visto que, esses trabalhos não foram testados com a mesma base de dados utilizada para colher as métricas do corrente trabalho, sendo esses artigos utilizados apenas para obter valores de referência para as métricas analisadas.

Mediante todo o conteúdo apresentado ao longo do corrente trabalho pode-se concluir que a ferramenta desenvolvida, ClusterPub, alcançou o objetivo de gerar dendogramas referentes aos relacionamentos semânticos presentes entre os artigos contidos em um arquivo bibliográfico, de maneira satisfatória, visto que, os dendogramas gerados possuem uma resolução legível e um tempo médio de confecção baixo, além de que foram obtidos valores razoáveis para as métricas de avaliação analisadas, quando empregado o método de ligação da média ponderada das distâncias em conjunto com a similaridade dos cossenos, como medida de distância, obtendo-se 0,8545 para o índice de Davies-Bouldin, 118,0987 para a métrica de Calinski-Harabasz e 0,5394 para o índice de silhueta.

7 Trabalhos Futuros

Com o intuito de aprimorar a solução proposta pelo corrente trabalho podem ser empregadas as seguintes melhorias:

- Extender o conjunto de formatações bibliográficas aceitas, possibilitando assim, a análise de outros tipos de arquivos que também são exportados por repositórios de trabalhos cientifícos, como, arquivos CSV.
- Permitir que o usuário informe por meio de parâmetros informados na interface de linha de comando a medida de distância e o método de ligação que deseja utilizar.
- Possibilitar que o usuário apenas visualize o dendograma resultante, sem forçá-lo a salvar o arquivo contendo o resultado.
- Desenvolver uma interface gráfica que permita que o usuário navegue pelo seu sistema de diretórios para encontrar o arquivo que deseja que seja analisado. O mesmo processo também será empregado para definir o local em que o resultado será salvo.
- Adicionar um mecanismo para ajustar o tamanho da imagem referente ao dendograma resultante, de acordo com a quantidade de artigos analisados.
- Executar testes de comparação com outros modelos de agrupamento textual, com o intuito de avaliar a qualidade da ferramenta desenvolvida pelo corrente trabalho.

Referências

ALBRIGHT, R. Taming text with the svd. 01 2004. Citado na página 15.

ALLA, S. **Evaluation Metrics for Machine Learning Models**. 2021. Acessado: 2024-09-07. Disponível em: <https://blog.paperspace.com/ml-evaluation-metrics-part-2/>. Citado na página 19.

ANGIONI ANGELO SALATINO, F. O. S. The aida dashboard: A web application for assessing and comparing scientific conferences. **IEEE ACCESS**, v. 10, 2022. Citado na página 27.

ASHABI SHAMSUL BIN SAHIBUDDIN, M. S. H. A. The systematic review of k-means clustering algorithm. In: **ICNCC '20: Proceedings of the 2020 9th International Conference on Networks, Communication and Computing**. [S.l.]: ACM, 2021. p. 14. Citado na página 16.

ASHARI ROMANTIKA BANJARNAHOR, D. R. F. S. P. A. A. P. D. N. H. I. F. Application of data mining with the k-means clustering method and davies bouldin index for grouping imdb movies. **Journal of Applied Informatics and Computing**, p. 11, 2022. Citado na página 20.

AWAN, A. A. **The Curse of Dimensionality in Machine Learning: Challenges, Impacts, and Solutions**. 2023. Acessado: 2024-10-20. Disponível em: <https://www.datacamp.com/blog/curse-of-dimensionality-machine-learning>. Citado na página 14.

BAEZA-YATES, B. R.-N. R. **Recuperação de Informação - Conceitos e Tecnologia das Máquinas de Busca**. [S.l.]: Bookman Editora, 2013. 27-28 p. Citado 2 vezes nas páginas 13 e 33.

_____. **Recuperação de Informação - Conceitos e Tecnologia das Máquinas de Busca**. [S.l.]: Bookman Editora, 2013. 31-40 p. Citado 2 vezes nas páginas 14 e 33.

BARNARD, J. **What are word embeddings?** 2024. Acessado: 2024-07-24. Disponível em: <https://www.ibm.com/topics/word-embeddings>. Citado na página 13.

BARUAH, I. D. **Dimensionality Reduction Techniques — PCA, LCA and SVD**. 2023. Acessado: 2024-09-05. Disponível em: <https://medium.com/nerd-for-tech/dimensionality-reduction-techniques-pca-lca-and-svd-f2a56b097f7c>. Citado 2 vezes nas páginas 15 e 33.

BHARGAVA, A. Grouping of medicinal drugs used for similar symptoms by mining clusters from drug benefits reviews. In: **Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM)**. [S.l.]: Birla Institute of Technology, 2019. p. 1053–1056. Citado na página 41.

BRAZ, A. M. Análise de agrupamento (cluster) para tipologia de paisagens. **Mercator - Universidade Federal do Ceará**, v. 19, p. 6, 2020. Citado na página 17.

CARVALHO, C. A. B. D. **O USO DE TÉCNICAS DE RECUPERAÇÃO DE INFORMAÇÕES EM CRIPTOANÁLISE**. 2006. Disponível em: <https://www.researchgate.net/profile/Carlos-Andre-Batista-De-Carvalho/publication/237699601/figure/fig10/AS:298731015557130@1448234432816/Exemplo-de-um-dendrograma.png>. Citado na página 22.

DAWAR, H. **All About the Pearson Correlation Coefficient in Data Science**. 2020. Acessado: 2024-08-10. Disponível em: <https://medium.com/swlh/all-about-the-pearson-correlation-coefficient-in-data-science-84d7cb771db0>. Citado na página 18.

FERREIRA, R. R. M.; PAIM, F. A. de P.; RODRIGUES, V. G. S.; CASTRO, G. S. A. Análise de cluster não supervisionado em r: agrupamento hierárquico. **Embrapa Territorial**, v. 1, p. 12–13, 2020. Citado 2 vezes nas páginas 15 e 34.

_____. Análise de cluster não supervisionado em r: agrupamento hierárquico. **Embrapa Territorial**, v. 1, p. 18–20, 2020. Citado na página 16.

FILEINFO. **The RIS (File Format) Explained**. 2024. Acessado: 2024-06-03. Disponível em: <https://fileinfo.com/extension/nbib>. Citado 2 vezes nas páginas 25 e 31.

FRISONI GIANLUCA MORO, A. C. G. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. **IEEE Access**, v. 4, 2016. Citado na página 26.

GAIKWAD, N. **Comparison of Word2vec with Hash2vec for Machine Translation**. Dissertação (Mestrado) — San Jose State UniversitySan Jose State University, 2020. Citado 2 vezes nas páginas 14 e 33.

GEEKS, G. for. **Factory method Design Pattern**. 2024. Acessado: 2024-07-22. Disponível em: <https://www.geeksforgeeks.org/factory-method-for-designing-pattern/>. Citado na página 31.

_____. **Template Method Design Pattern**. 2024. Acessado: 2024-07-22. Disponível em: <https://www.geeksforgeeks.org/template-method-design-pattern/>. Citado na página 31.

_____. **What is Python? it's Uses and Applications**. 2024. Acessado: 2024-03-02. Disponível em: <https://www.geeksforgeeks.org/what-is-python/>. Citado 2 vezes nas páginas 21 e 30.

GOMES, M. M. A. Implementation of a noise filter for grouping in bibliographic databases using latent semantic indexing. **International Journal of Bioinformatics and Intelligent Computing**, v. 2, n. 1, p. 5–5, 2023. Citado na página 31.

HE S.C. HUI, A. F. Y. Citation-based retrieval for scholarly publications. **IEEE Distributed Systems Online**, 2003. Citado na página 26.

HUYNH KIEM HOANG, L. D. H. T. H. L. S. G. T. Scientific publication recommendations based on collaborative citation networks. In: IEEE. **2012 International Conference on Collaboration Technologies and Systems (CTS)**. [S.l.], 2012. p. 316–321. Citado na página 27.

IBM. **What is principal component analysis (PCA)?** 2023. Acessado: 2024-09-05. Disponível em: <https://www.ibm.com/topics/principal-component-analysis>. Citado na página 15.

_____. **Gráficos de Dendrogram**. 2024. Acessado: 2024-07-26. Disponível em: <https://www.ibm.com/docs/pt-br/watsonx/saas?topic=types-dendrogram-charts>. Citado 2 vezes nas páginas 21 e 34.

INCITEFUL. **Paper Discovery**. 2020. Acessado: 2024-04-04. Disponível em: <https://help.inciteful.xyz/power-users.html#paper-discovery>. Citado 2 vezes nas páginas 26 e 28.

JIA MEILI SUN, J. L. S. H. W. Feature dimensionality reduction: a review. **Complex Intelligent Systems**, v. 8, p. 2664, 2022. Citado na página 14.

K.KAMESHWARAN, K. Survey on clustering techniques in data mining. **International Journal of Computer Science and Information Technologies**, v. 5, p. 1–2, 2014. Citado na página 15.

KRIEGER, P. E. G. A. M. A generalized rand-index method for consensus clustering of separate partitions of the same data base. **Journal of Classification**, p. 66–67, 1999. Citado na página 21.

LIBRARY, M. **The RIS (File Format) Explained**. 2022. Acessado: 2024-03-02. Disponível em: <https://library.mskcc.org/blog/2022/09/the-ris-file-format-explained/#:~:text=The%20RIS%20(file%20format)%20is,tag%20supports%20a%20different%20field>. Citado 2 vezes nas páginas 24 e 31.

LUND. **Bibliographic databases**. 2023. Acessado: 2024-03-02. Disponível em: <https://libguides.lub.lu.se/c.php?g=677619&p=4829257>. Citado na página 23.

MARCACINI, R. M.; MOURA, M. F.; REZENDE, S. O. Uma abordagem para seleção de grupos significativos em agrupamento hierárquico de documentos. In: **Congresso da Sociedade Brasileira de Computação - CSBC**. [S.l.]: SBC, 2009. p. 931. Citado 2 vezes nas páginas 17 e 34.

MELO, V. V. de. **Clustering de Artigos Científicos em uma Ferramenta Inteligente de Apoio à Pesquisa**. 2005. Citado na página 27.

MERCIONI, S. H. M. A. A survey of distance metrics in clustering data mining techniques. In: ACM (Ed.). **CGSP '19: Proceedings of the 3rd International Conference on Graphics and Signal Processing**. AAAI Press, 2019. p. 45. Disponível em: <https://dl.acm.org/doi/pdf/10.1145/3338472.3338490>. Citado na página 17.

MIESLE, P. **What is Cosine Similarity: A Comprehensive Guide**. 2023. Acessado: 2024-08-09. Disponível em: <https://www.datastax.com/guides/what-is-cosine-similarity>. Citado na página 17.

MUGHNYANTI, S. E. M. Z. M. Analysis of determining centroid clustering xmeans algorithm with davies-bouldin index evaluation. In: **IOP Conference Series: Materials Science and Engineering**. [S.l.]: IOP Publishing Ltd, 2020. p. 2–3. Citado na página 20.

MULUNDA PETER WAIGANJO, L. M. C. Towards implementation of an information dissemination tool for health publications: Case of a developing country. In: IST-AFRICA INSTITUTE AND IIMC. **IST-Africa 2020 Conference Proceedings**. [S.l.], 2020. p. 1–11. Citado na página 26.

OUZZANI HOSSAM HAMMADY, Z. F. A. E. M. Rayyan—a web and mobile app for systematic reviews. **Systematic Reviews logo Systematic Reviews**, Springer Nature, v. 5, n. 210, p. 4–5, 2016. Citado 2 vezes nas páginas 27 e 29.

PAPERPILE. **A complete guide to the BibTeX format**. 2022. Acessado: 2024-03-02. Disponível em: <https://www.bibtex.com/g/bibtex-format/>. Citado 2 vezes nas páginas 23 e 31.

PIUBELLO, H. **Descubra a Relevância da Distância Euclidiana em Programação e Ciência de Dados**. 2023. Acessado: 2024-03-11. Disponível em: <https://codecrush.com.br/blog/distancia-euclidiana>. Citado na página 17.

RAMíREZ, S. **Typer**. 2024. Acessado: 2024-07-20. Disponível em: <https://typer.tiangolo.com/>. Citado 2 vezes nas páginas 23 e 30.

SILVA BERNARDO PEREIRA NUNES, S. W. M. S. S. D. Crystiam Pereira e. Linked data in education: a survey and a synthesis of actual research and future challenges. **IEEE Transactions on Learning Technologies**, 2017. Citado na página 26.

TANG YANGYONG ZHU, S. V. B. J. M. P. M. W. L. Z. J. M. D. W. B. C. Visualizing literature review theme evolution on timeline maps: Comparison across disciplines. **IEEE ACCESS**, v. 7, 2019. Citado na página 27.

TERRA. **MEC registra aumento de alunos em cursos de Pós-Graduação**. 2023. Disponível em: <https://www.terra.com.br/noticias/mec-registra-aumento-de-alunos-em-cursos-de-pos-graduacao,58241e304869d44a04f5d324d6be31b5riot57a1.html#:~:text=Do%20n%C3%BAmero%20total%2C%2076.323%20s%C3%A3o,profissional%20e%2041.964%20de%20doutorado.&text=Uma%20tend%C3%AAncia%20positiva%20%C3%A9%20observada,tem%20sido%20de%20aproximadamente%209%25>. Citado na página 10.

VYSALA, A.; GOMES, J. **Evaluating and Validating Cluster Results**. 2020. Disponível em: <https://arxiv.org/abs/2007.08034>. Citado na página 19.

WANG, Y. X. X. An improved index for clustering validation based on silhouette index and calinski-harabasz index. In: **IOP Conference Series: Materials Science and Engineering**. [S.l.]: IOP Publishing Ltd, 2019. p. 2. Citado na página 20.

YUAN JUSTIN ZOBEL, P. L. M. Measurement of clustering efectiveness for document collections. **Information Retrieval Journal**, p. 257–258, 2022. Citado na página 41.