UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE RUSSAS

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

JOÃO VICTOR FONSECA SOMBRA

UTILIZANDO CLUSTERIZAÇÃO PARA IDENTIFICAR PADRÕES EM PUBLICAÇÕES DE ARTIGO NA AMÉRICA LATINA

RUSSAS

2023

JOÃO VICTOR FONSECA SOMBRA

UTILIZANDO CLUSTERIZAÇÃO PARA IDENTIFICAR PADRÕES EM PUBLICAÇÕES DE ARTIGO NA AMÉRICA LATINA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Orientadora: Prof. Dra. Tatiane Fernandes Figueiredo.

RUSSAS

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S676u Sombra, João Victor Fonseca.
Utilizando clusterização para identificar padrões em publicações de artigo na América Latina / João Victor Fonseca Sombra. – 2023.
28 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Engenharia de Software, Russas, 2023.
Orientação: Profa. Dra. Tatiane Fernandes Figueiredo.

1. Agrupamento. 2. MiniBatch K-Means. 3. Artigos Científicos. I. Título.

CDD 005.1

JOÃO VICTOR FONSECA SOMBRA

UTILIZANDO CLUSTERIZAÇÃO PARA IDENTIFICAR PADRÕES EM PUBLICAÇÕES DE ARTIGO NA AMÉRICA LATINA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Aprovada em: 14 de Dezembro de 2023

BANCA EXAMINADORA

Prof. Dra. Tatiane Fernandes Figueiredo (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Eurinardo Rodrigues Costa
Universidade Federal do Ceará - UFC

Prof. Dr. Bonfim Amaro Junior
Universidade Estadual do Ceará - UECE

AGRADECIMENTOS

Primeiramente, expresso minha gratidão a Deus por me guiar, proteger e me dar forças para chegar até aqui e me tornar quem sou.

Agradeço a toda minha família, que sempre me auxiliou e me incentivou a seguir meu caminho, nunca me abandonou ou me deixou sozinho. Quero agradecer aos meus pais, Malba Tahan e Josilânia, por me ensinarem a caminhar e por me auxiliarem sempre que necessário. Obrigado por me darem todo o ensino, apoio e proteção, mesmo nos momentos mais difíceis confiaram em mim e me deram forças para continuar. Gostaria também de agradecer em memória do meu tio Jarbas Ramos, que embora não esteja mais conosco, continua a me inspirar com sua coragem e determinação, obrigado tio.

Agradeço também aos meus primos, Clarisse, Jefferson, Letícia, Levi, Luigi e Orlando, que sempre me proporcionaram momentos de alegria e descontração.

Agradeço aos meus amigos, Natália, Lucas, Dhioleno, Adryan, Darcio, Hanna, João Pedro, Anne, Milene, aos meus amigos artistas, Wilkinis e Thiago, e aos meus amigos mais recentes Yan, Camila, Marina, Mayronn e Victor. A todos vocês, muito obrigado. Obrigado por me ajudarem nesta caminhada difícil, vocês me ajudaram e estiveram ao meu lado diariamente, nos dias felizes e tristes, em meio a brigas e risos. Muito Obrigado a todos vocês por me ouvirem e por sempre me fazerem companhia.

À minha orientadora, professora Tatiane Fernandes Figueiredo, quero agradecer pelo seu apoio e confiança desde o meu terceiro semestre na instituição. Ser seu monitor de Estrutura de Dados e trabalhar ao seu lado em tantos projetos foi uma experiência enriquecedora e de profunda inspiração. Muito obrigado pelos conselhos, pela amizade e pelo apoio, obrigado pela sua orientação neste trabalho, nos estudos e na vida, e espero profundamente continuar sendo orientado por você no mestrado, pois sua orientação é de grande inspiração para minha jornada.

Agradeço aos professores Bonfim Amaro e Eurinardo Rodrigues, por seu tempo e esforço para avaliar meu trabalho. Seus comentários e críticas foram e serão essenciais para o aprimoramento deste projeto.

Agradeço também aos demais professores por compartilharem seus conhecimentos e experiências, todos vocês foram fundamentais para minha formação acadêmica.

Por fim, mas não menos importante, agradeço a todos que de alguma forma contribuíram para que eu chegasse onde cheguei e para a realização deste trabalho. Cada um de vocês desempenhou um papel fundamental na minha caminhada e por isso sou muito grato.

"O prêmio pela sua dedicação diária, quem diria, é ver o amor nos olhos de quem você mais queria."
(Kuffel, Caciano)

RESUMO

Com o avanço da tecnologia e o fácil acesso às informações, o número de publicações de artigos científicos cresceu exponencialmente. No entanto, devido à enorme quantidade de informação, identificar padrões relevantes tornou-se um desafio. Este trabalho propõe uma abordagem de agrupamento utilizando o método *MiniBatch K-Means* aplicado em dados de artigos publicados nos últimos 20 anos na América Latina, bem como uma exploração de seus resultados. A metodologia aplicada resultou na criação de 50 grupos distintos. Cada um desses grupos é composto por artigos relacionados a temas específicos, alguns exemplos são a existência de um grupo contendo artigos apenas relacionados a "expressão gênica e metabolismo em proteínas e células" e outro grupo relacionado a "otimização em tempo polinomial e complexidade combinatória". Os resultados obtidos pelo método de agrupamento demonstraram a capacidade do algoritmo em identificar e organizar eficientemente artigos científicos com base em seus resumos e palavras-chave.

Palavras-chave: agrupamento; Minibatch K-means; artigos científicos.

ABSTRACT

With the technology advancement and easy access to information, the number of publications scientific articles have grown exponentially. However, due to the massive amount of information, identifying relevant patterns has become a challenge. This work proposes an approach clustering using the MiniBatch K-Means method appplied in data from articles published in the last 20 years in Latin America, as well as an exploration of its results. The methodology applied resulted in the creation of 50 distinct groups. Each of these groups is composed of articles related to specific topics, some examples is the existence of a group containing articles only related to "gene expression and metabolism in proteins and cells"and another group related to "polynomial-time optimization and combinatorial complexity". The results obtained by clustering method demonstrated algorithm's ability to identify and efficiently organize scientific articles based on their abstract and keywords.

Palavras-chave: clustering; Minibatch K-means; scientific articles.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CRISP-DM	*Cross Industry Standard Process for Data Mining*
DOI	*Digital Object Identifier*
LSA	*Latent Semantic Analysis*
NLP	*Natural Language Processing*
SSE	*Sum of Squared Errors*
Wos	*Web of Science*

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da tecnologia e o fácil acesso à informação, a quantidade de publicações de artigos científicos tem crescido exponencialmente nas mais diversas áreas. Entretanto, identificar padrões relevantes em meio a essa quantidade de informações tornou-se um desafio. Buscando formas para solucionar este problema, o estudo realizado por Chang *et al.* (2022) apresenta um classificador de publicações com potencial de desenvolvimento na área do turismo. De forma análoga a isso, os autores Ayala-Bastidas *et al.* (2021) identificaram padrões e estratégias de baixo e alto impacto em publicações de pesquisadores mexicanos. Vale destacar que ambos os autores utilizaram o método de clusterização *K-Means* em suas pesquisas. Por outro lado, o autor Kang (2003) direcionou sua pesquisa para aprimorar especificamente o cálculo de similaridade entre documentos.

Este trabalho segue com objetivos e metodologia semelhantes aos dos autores mencionados, utilizando um método mais otimizado de clusterização, denominado *MiniBatch K-Means*, considerando uma base de dados com publicações de artigos científicos da América Latina. Esta pesquisa possui o objetivo de identificar padrões que ajudem a compreender os temas das publicações científicas na região e fornecer *insights* valiosos para a comunidade acadêmica.

A estrutura deste trabalho encontra-se da seguinte forma: no Capítulo 2 é apresentado o objetivo geral e os específicos; no Capítulo 3 são apresentadas as definições de conceitos e algoritmos importantes para este trabalho; no Capítulo 4 são apresentados trabalhos encontrados na literatura que são similares a este; no Capítulo 5 são apresentados os procedimentos utilizados para realizar a pesquisa e desenvolvimento desta monografia; por fim, no Capítulo 6 são apresentadas as conclusões e trabalhos futuros.

2 OBJETIVOS

2.1 Objetivo geral

Aplicar técnicas de clusterização em uma base de dados de artigos publicados na América Latina, considerando para tal dados sobre seu resumo e palavras-chave, a fim de identificar padrões nessas publicações.

2.2 Objetivos específicos

- Construir uma base de dados com as informações de artigos científicos produzidos na América Latina;
- Padronizar e limpar a base de dados em estudo;
- Implementar um algoritmo de aprendizado de máquina para identificar os padrões existentes e agrupar os artigos por similaridade;
- Realizar uma análise final a partir dos dados obtidos.

3 FUNDAMENTAÇÃO TEÓRICA

Para um bom entendimento do trabalho, são apresentados neste capítulo os conceitos fundamentais abordados na pesquisa. Na Seção 3.1 é apresentado o conceito de processamento de linguagem natural. Na Seção 3.2 é detalhado a técnica de stemização e o algoritmo *Porter Stemmer* muito utilizado para aplicação desta técnica. Na Seção 3.3 é apresentado o conceito de clusterização. Na Seção 3.4 e na Subseção 3.4.1 são apresentados os algoritmos *K-Means* e *MiniBatch K-Means*, respectivamente. Por fim, na Seção 3.5 é mostrado a abordagem *Silhouette Score*.

3.1 Processamento de linguagem natural

A técnica de *Natural Language Processing* (NLP) teve sua origem voltada a recuperação de informações a partir de dados textuais. Nesse contexto, diversas técnicas baseadas em estatística foram definidas para resolver alguns subproblemas referentes aos textos, como a *tokenization* (identificação de palavras individuais dentro de uma frase) e a decomposição morfológica (técnica de *stemming*, apresentada na próxima seção) (NADKARNI *et al.*, 2011).

Existem diversas metodologias para a aplicação prática do NLP, este trabalho utiliza uma abordagem de sequenciamento de tarefas, definido por Nadkarni *et al.* (2011) como *pipelines*. Essa metodologia, segundo os autores, consiste na aplicação de várias subtarefas executadas sequencialmente, onde a saída de um módulo se torna a entrada para o próximo. O principal objetivo dessa metodologia é resolver os subproblemas necessários e preparar os dados para as próximas etapas.

3.2 Stemização

Em diversos idiomas, as palavras apresentam derivações morfológicas, por exemplo, na lingua inglesa, a palavra "*problem*"possui derivações como "*problematic*", "*problems*"ou "*unproblematic*". A técnica de stemização (do inglês *Stemming*) é bastante útil quando é necessário extrair informações relevantes em meio a grandes conjuntos de dados textuais. Essa abordagem consegue converter uma entrada textural em um conjunto de palavras "raiz", simplificando assim a execução de algoritmos de busca e aprimorando a eficiência no processamento de linguagem natural. (WILLETT, 2006).

O algoritmo *Porter Stemmer*, proposto por Porter (1980), opera com uma série de

regras fundamentadas na estrutura morfológica das palavras, com o objetivo de identificar e eliminar possíveis derivações. Essas regras são aplicadas de maneira sequencial, até que não haja mais condições para execução. Esse processo resulta na obtenção da forma "raiz"de cada palavra, representando uma versão simplificada e padronizada.

3.3 *Clustering*

A técnica de *clustering* consiste em agrupar instâncias de um conjunto de dados inicial em *K* subconjuntos (chamados *clusters*), baseando-se pelo grau de similaridade, ou seja, em cada *cluster* os dados agrupados são mais semelhantes uns com os outros, enquanto os dados que estão em *clusters* distintos são considerados mais diferentes.

Uma vez que os algoritmos de *clustering* apenas agrupam dados semelhantes, é necessário utilizar alguma medida que determine se dois dados de uma dada instância são similares ou diferentes. Para isso, muitos métodos de agrupamento utilizam medidas de distância que permitem determinar similaridade ou dissimilaridade entre pares de dados (MAIMON *et al.*, 2005). Vários critérios de avaliação têm sido desenvolvidos na literatura, sendo que um dos mais destacados a função *Sum of Squared Errors* (SSE). O SSE mede a semelhança de dois dados de uma dada instância com base na distância entre os centroides (centro de cada *cluster*) e avalia a dispersão dos elementos dentro dos subconjuntos.

Para exemplificar a medida do SSE podemos imaginar dois *clusters*, *A* e *B*. Para calcular a função SSE, começamos criando dois centroides dispersos aleatoriamente entre os dados e os pontos mais próximos são associados a cada um deles. Em seguida, calculamos a distância de cada ponto do *cluster A* ao seu centroide, essas distâncias são elevadas ao quadrado e somadas, resultando no valor de SSE do *cluster A*. Em seguida, repetimos o mesmo processo com o *cluster B*, calculando o seu SSE. Por fim, somamos o SSE dos clusters *A* e *B*, para chegarmos ao SSE total. Com resultados menores, conseguimos *clusters* mais compactos e bem definidos, indicando uma melhor qualidade do agrupamento.

3.4 *K-Means*

O algoritmo *K-Means* é um método de clusterização não supervisionado baseado em centroídes, que busca encontrar o *clustering* que minimiza a SSE. Um algoritmo de aprendizado é considerado não supervisionado quando não se tem, ou não se utiliza dados sobre os rótulos da

base. Esse algoritmo depende da escolha do número *K* de *clusters* para realizar o agrupamento dos dados, entretanto a escolha inapropriada de *K* pode resultar em agrupamentos insatisfatórios. Portanto, é indispensável a escolha de uma boa métrica para definição do número ideal de *clusters*. Nesse contexto, o *Silhouette Score* é um método utilizado para identificar o número ótimo de *clusters*, como é apresentado na Subseção 3.5. Com a utilização desta técnica, será possível obter resultados mais precisos na clusterização dos dados.

O funcionamento do algoritmo *K-Means*, segundo Nunes (2016) ocorre da seguinte forma: inicialmente, o algoritmo distribui entre os *K clusters* um conjunto de dados da instância de forma aleatória e calcula a média dos dados, a fim de determinar a posição inicial dos centroides. Em seguida, cada dado da instância é associado ao centroide mais próximo. Após a atribuição, os centroides são atualizados com a média dos pontos associados a eles. Por fim, essas duas etapas são realizadas iterativamente em busca do mínimo local, ou seja, quando os centroides não sofrem alterações após uma interação.

3.4.1 MiniBatch K-Means

Embora o *K-Means* seja amplamente utilizado por conta de seu bom desempenho de tempo, ao aplica-lo em grandes conjuntos de dados, enfrenta um ganho significativo em custo espacial e temporal. Isso ocorre devido o conjunto de dados precisar ser armazenado na memória principal (BéJAR, 2013).

Tendo essa problemática em vista, o pesquisador Sculley (2010) propôs o algoritmo *Mini-Batch K-Means* como uma alternativa. Nesta adaptação do clássico algoritmo de clusterização, são criados pequenos lotes aleatórios de dados com tamanhos fixos, denominados *Mini-Batchs*. A motivação por trás desse método é que os *Mini-Batchs* tendem a conter menos ruído, possibilitando a convergência para soluções melhores, sem sofrer um aumento significativo no custo computacional ao lidar com conjuntos de dados maiores.

3.5 Silhouette Score

Apesar do algoritmo *MiniBatch K-Means* ser bastante eficiente, a escolha de um número ideal de agrupamentos ainda é uma problemática a ser resolvida. Com o passar do tempo, diversos métodos que buscam solucionar esse problema foram surgindo, entre eles, destaca-se o método *Silhouette Score*.

Este método calcula seu coeficiente considerando a média de distância de cada dado dentro do seu próprio *cluster* (similaridade de dados) e a média da distância de cada dado até o *cluster* mais próximo (dissimilaridade de dados). O resultado deste calculo gera um *Score* que varia de -1 (indicando que os dados foram atribuídos a *clusters* incorretos) até +1 (indicando que os dados foram bem divididos e agrupados). Vale ressaltar que valores próximos a 0 sugerem que alguns agrupamentos podem estar sobrepostos (SHAHAPURE; NICHOLAS, 2020).

4 TRABALHOS RELACIONADOS

Neste capítulo, é apresentado as pesquisas relevantes para fundamentar a importância deste trabalho acadêmico. Na Subseção 4.1 é apresentado o trabalho de Ayala-Bastidas *et al.* (2021) que fala sobre estratégias de publicação no México. A Subseção 4.2 detalha o trabalho de Chang *et al.* (2022) que busca identificar tendências de pesquisa na área do turismo. Por fim, a Subseção 4.3 apresenta a pesquisa de Kang (2003) que apresenta uma abordagem alternativa para identificar a similaridade de dados

4.1 Identificando estratégias de publicação de pesquisadores do México

A busca por estratégias de publicação tem sido investigada desde a seleção de temas de pesquisa até o impacto da escolha de parceiros. O trabalho de Ayala-Bastidas *et al.* (2021) descreve os efeitos dessas escolhas e a necessidade de considerá-las no início da carreira científica. O estudo identifica estratégias de publicação por meio de observações comportamentais, direcionado especificamente a pesquisadores da área de engenharia.

Neste trabalho, os autores analisaram 3.156 pesquisadores filiados ao Sistema Nacional de Pesquisadores do México, com publicações entre os anos 2007 e 2016. No estudo as estratégias de publicações foram estabelecidas como uma combinação de indicadores de produtividade e colaboração em um período de dois anos. O sucesso de uma publicação é medido em termos de citações recebidas nos últimos três anos.

Nesta pesquisa foi utilizado o algoritmo *K-Means* para identificar padrões alinhados com um determinado nível de citação, com isso foi possível chegar a 8 estratégias de publicação com impactos distintos. A análise de Ayala-Bastidas *et al.* (2021) confirma a importância da colaboração internacional de pesquisadores e seu impacto em citações. A metodologia do trabalho pode ser usada para descobrir estratégias em outras áreas ou regiões geográficas.

4.2 Identificação de tendências de pesquisa em diferentes áreas do turismo

Nos últimos anos, empresas têm utilizado inteligência artificial para analisar enormes quantidades de dados com o objetivo de identificar padrões úteis que possam ajudá-las a inovar seus modelos de negócio. O trabalho de Chang *et al.* (2022) utiliza-se dessa premissa, que por sua vez tem o objetivo de realizar uma análise abrangente sobre publicações científicas, classificando e avaliando a relevância de artigos acadêmicos na área de turismo. A partir desta

análise, foi identificado as tendências de pesquisa em diferentes temas dentro do campo do turismo.

Nesse estudo foram classificados 5.783 artigos relacionados a turismo do banco de dados *Web of Science* (Wos) publicados entre os anos de 2010 e 2019. Para realizar a classificação o trabalho foi dividido em três etapas, sendo elas: seleção e segmentação de palavras, nessa etapa foi aplicado algoritmos para segmentar e filtrar as palavras relevantes utilizadas nos artigos. Posteriormente, aplicou-se o algoritmo *K-Means* para agrupamento dos dados tratados. Os autores utilizaram o método de análise hierárquica para determinar o valor de K. Por fim, realizou uma análise de co-palavras para entender a evolução e tendências dos artigos acadêmicos.

O trabalho de Chang *et al.* (2022) chegou a resultados satisfatórios ao identificar os principais temas na área de turismo, além daqueles com potencial de desenvolvimento. O estudo também permitiu estabelecer um processo de classificação automática para artigos relacionados ao turismo, tornando a análise desses documentos mais eficiente. Outro resultado relevante foi a identificação das palavras-chave associadas aos diferentes temas no campo do turismo, fornecendo uma visão valiosa para a compreensão dos principais temas de pesquisa.

4.3 Clusterizando documentos com base nas palavras-chave

O agrupamento de documentos em *clusters* é geralmente realizado considerando a similaridade entre dados, sendo essa similaridade medida pela frequência de termos nos documentos. Contudo, os métodos convencionais de agrupamento não consideram o conteúdo específico dos objetos de um *cluster*. Nesse contexto, Kang (2003) adotou uma abordagem analítica para aprimorar o cálculo de similaridade entre documentos, utilizando como base as palavras-chave.

Nesse artigo, o autor explorou uma base de dados comporta por 383 artigos, cada um contendo, em média, 132 palavras-chave. Seu trabalho propôs um novo método de agrupamento, baseado na ponderação das palavras-chave. Para realizar esse processo, o autor aplicou um método clássico de clusterização. Em seguida, dentro de cada *cluster*, foram selecionados os termos que possuíam valores de peso mais elevados. Por fim, esses termos foram utilizados no cálculo de similaridade entre documentos.

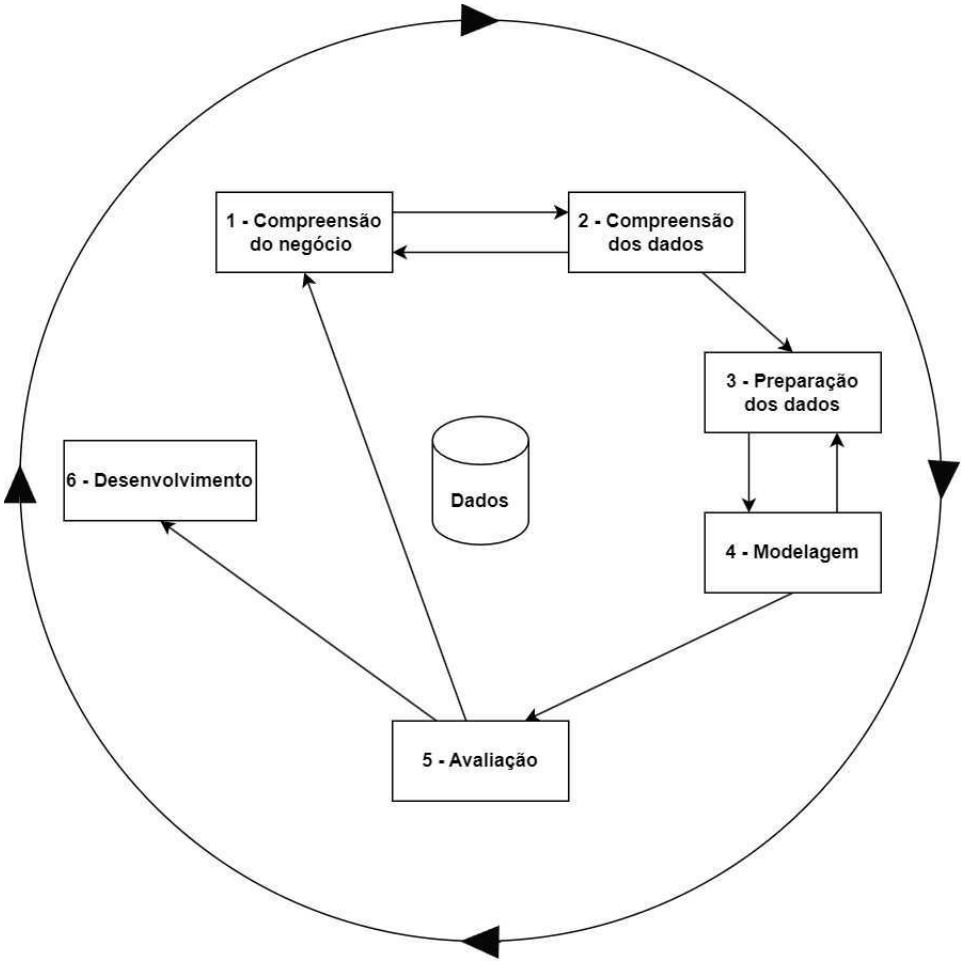
Ao implementar o modelo de ponderação de palavras-chave, Kang (2003) concluiu que, em termos de qualidade dos agrupamentos, seu novo modelo supera o método baseado

apenas na frequência. O algoritmo, orientado por palavras-chave, atinge os melhores resultados quando são utilizados entre 30% e 60% dos termos mais frequentes de cada *cluster*.

5 METODOLOGIA

Este capítulo descreve e aplica a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), um padrão de processo utilizado entre indústrias para a mineração de dados. Esse método é amplamente adotado para orientar projetos de processamento e análise de um grande volume de dados. A Figura 1 ilustra as fases deste modelo:

Figura 1 – Fases do CRISP-DM



Fonte: Elaborado pelo Autor (2023)

5.1 Coleta e compreensão de dados

Nesta etapa é realizada a coleta de dados por meio da plataforma *Scopus*, que oferece recursos como pesquisa avançada e exportação de artigos em diversos formados. Além disso, a plataforma permite a aplicação diversos filtros para obter resultados mais precisos. Neste trabalho foram aplicados os seguintes filtros com o objetivo de obter os artigos recentemente publicados na América latina:

- Pesquisa: *combinatorial optimi**
- Data de publicação: entre 2002 e 2022
- Idioma: Inglês
- Estado da publicação: Final
- Países: Costa Rica, Guatemala, México, Panamá, Argentina, Bolívia, Brasil, Chile, Colômbia, Equador, Paraguai, Peru, Uruguai, Venezuela, Cuba, República Dominicana, Guadalupe, Martinica, Porto Rico.

Após a conclusão da filtragem, os dados foram exportados no formato *.csv*, originando a base inicial composta por 6.466 artigos. Esta base inclui informações como: autores, nomes completos dos autores, *IDs* dos autores, título do artigo, ano de publicação, número de vezes citado, *Digital Object Identifier* (DOI), link para artigo, *abstract* e lista de palavras-chaves do artigo.

5.2 Preparação dos dados

Durante esta fase, os dados são preparados para a etapa de implementação. É importante destacar que cada passo desse processo de preparação de dados é crucial para uma melhor eficiência nas fases seguintes, proporcionando resultados mais precisos ao fim do trabalho. A preparação dos dados foi realizada nas colunas *abstract* e *keywords* e abrange uma série de atividades sequenciais:

- ***Tokenization* e Normalização**: inicialmente, é necessário realizar a normalização o texto, para isso todos os caracteres do alfabeto são convertidos para minusculo, e quaisquer os caractere que não pertença ao alfabeto, como números ou caracteres especiais, são removidos. Em seguida, é realizado a "Tokenização", que consiste em substituir todos os espaços no texto por vírgulas, transformando, assim, uma entrada de linguagem natural em um vetor de palavras.
- **Remoção de linhas vazias**: este passo é crucial para eliminar linhas da base que possam conter dados ausentes ou que tenham se tornaram nulas após a normalização. Como resultado dessa fase, a base inicial de 6.466 artigos foi reduzida para 5.086, uma redução de mais de 1.000 artigos. Este corte foi necessário devido à impossibilidade de classificar esses artigos por conta da falta de informações.
- **Remoção de *StopWords***: na linguagem natural, diversas palavras são consideradas irrelevantes para a análise de dados, e entre elas existem as *StopWords*, que inclui temos como:

i, *me*, *my*, *we*, *you*, dentre diversos outros. A remoção de palavras como essas é crucial, pois ocupam espaço desnecessário e consomem tempo de processamento. Felizmente, a remoção dessas *StopWords* é simples devido à disponibilidade de funções e bibliotecas que facilitam esse processo.

- ***Stemming***: nessa etapa, a implementação do algoritmo de *stemming*, apresentado na Seção 3.2, desempenhou um papel fundamental. A importância desse passo é a capacidade de tornar possível agrupar palavras como *computer*, *computation* e *computational* em um único grupo, tendo em vista o radical em comum *cumput*. Esta técnica fornece uma uniformização e simplificação das palavras, facilitando a identificação de padrões e a interpretação de resultados, além de proporcionar com uma melhora significativa no processo de *clustering* adotado nas próximas etapas.
- **Remoção de dados duplicados**: por fim, nesta fase, é realizado uma verificação a fim de identificar e remover possíveis linhas duplicadas dentro da base. Entretanto, como resultado, nenhuma instância repetida foi identificada, e a base manteve-se inalterada. A implementação dessa etapa é fundamental, mesmo sem a remoção de dados, pois assegura a integridade da base e elimina a possibilidade de instâncias duplicadas, proporcionando uma melhor interpretação dos resultados em etapas subsequentes.

5.3 Implementação do *MiniBatch K-Means*

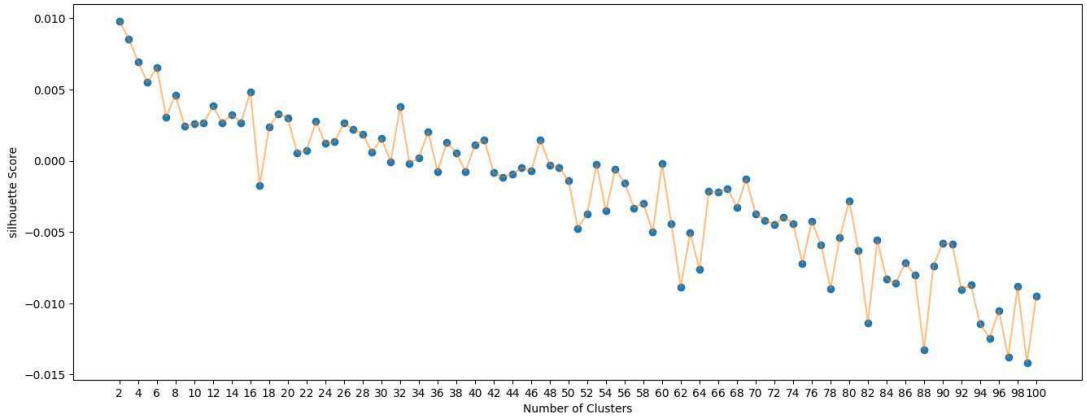
Nesta fase, é realizado a implementação do algoritmo de clusterização *MiniBatch K-Means*, conforme abordado na Seção 3.4.1. Entretanto, vale ressaltar que a execução desse algoritmo envolve três etapas: transformação textual, definição do número de agrupamentos e, por fim, clusterização.

- **Transformação textual**: o algoritmo *MiniBatch K-Means* trabalha apenas com dados numéricos, portanto, é necessário realizar uma transformação dos dados textuais para torna-los compatíveis. Nesse contexto, o algoritmo *TF-IDF Vectorizer*, é utilizado para converter o texto em uma representação numérica, baseada na frequência de cada palavra e em sua importância em relação ao conjunto total. Esse processo torna possível a aplicação eficaz de métodos de clusterização.
- **Número de agrupamentos**: para determinar o número ideal de *clusters*, foi utilizado o método *Silhouette*, exposto na Seção 3.5. Contudo, como pode ser visto na Figura 2, os resultados desse método mostraram-se bastante próximos a zero e, em alguns casos, até

mesmo negativos. Para resolver esse problema, foi necessário a aplicação do método de redução de dimensionalidade LSA. Esse modelo, segundo Foltz (1996), foi originalmente projetado para aprimorar a eficácia de métodos de recuperação de informações, permitindo comparações de similaridade semântica entre pares de informações textuais. Com a combinação desses dois métodos, os resultados foram bastante satisfatórios, conforme evidenciado na Figura 3. O número de *clusters* escolhido para a próxima fase foi 50.

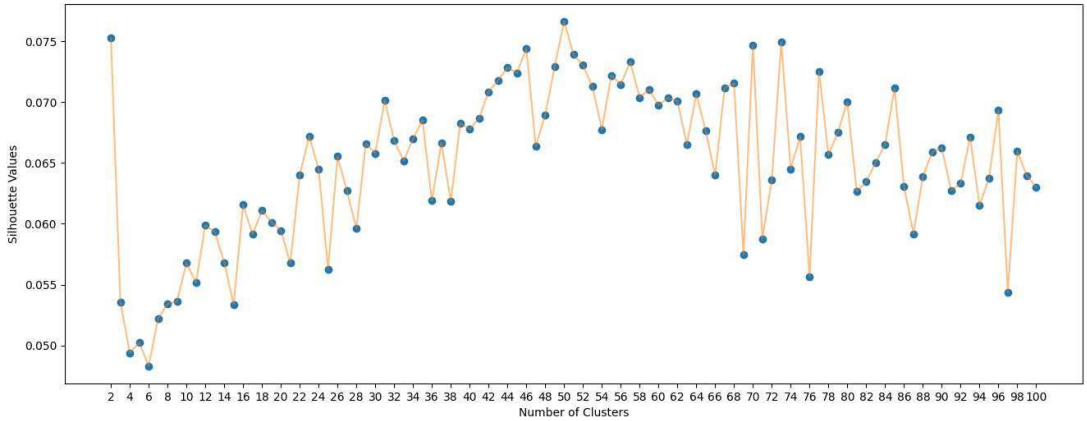
– **Clusterização**: com a base numérica devidamente preparada e o número de agrupamentos definido, basta executar o algoritmo *MiniBatch K-Means*, conforme visto na Seção 3.4.1. Ao término da execução, é realizada a adição de uma nova coluna à base de dados, chamada *Cluster Labels*. Essa coluna é responsável por armazenar a informação de qual *cluster* cada artigo pertence, fornecendo uma categorização clara de cada elemento da base.

Figura 2 – Pontuação do método *Silhouette* para *clusters* variando de 2 a 100.



Fonte: Elaborado pelo autor (2023)

Figura 3 – Pontuação dos métodos *Silhouette* e LSA para *clusters* variando de 2 a 100.

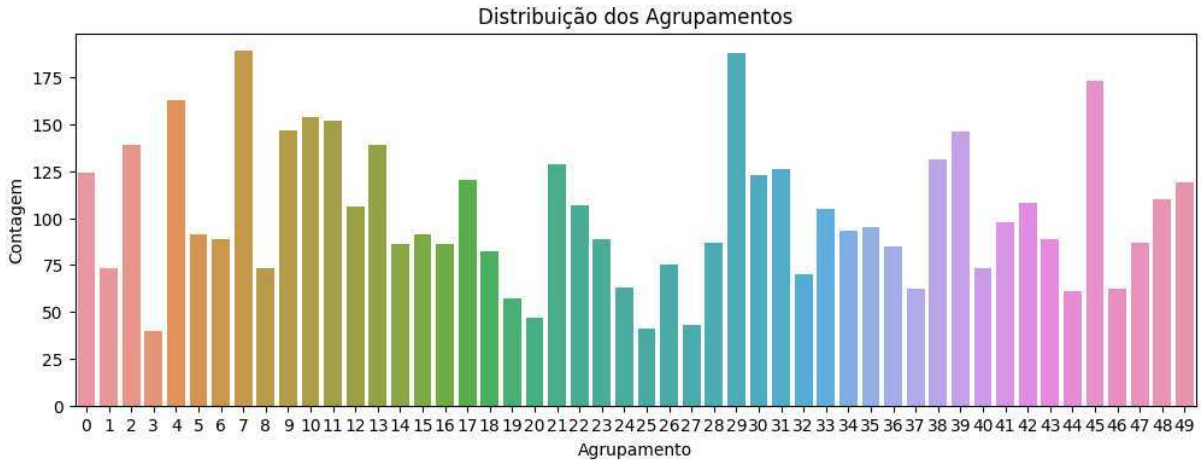


Fonte: Elaborado pelo autor (2023)

5.4 Avaliação dos resultados

Com a conclusão da etapa anterior, tornou-se possível iniciar a avaliação da qualidade dos agrupamentos. Com um total de 50 *clusters*, cada um contendo, em média, 85 artigos (distribuição ilustrada na Figura 4), surgiu a necessidade de atribuir nomes a esses *clusters*. Essa nomenclatura visa facilitar a compreensão dos tópicos de artigos armazenados em cada agrupamento.

Figura 4 – Distribuição dos artigos entre os 50 *clusters*



Fonte: Elaborado pelo autor (2023)

Para realizar a nomeação dos *clusters*, foram escolhidas as 15 palavras mais frequentes de cada agrupamento para representar seu *cluster*. Entretanto, devido ao *Abstract* possuir um volume superior de palavras em comparação com as *Keywords* e algumas dessas palavras consideradas irrelevantes assumindo a liderança, como "*use*"ou "*studi*", as *Keywords* foram ponderadas com o dobro do peso em relação as palavras do *Abstract*.

Com o objetivo de complementar ainda mais o nome dado aos *clusters*, foi utilizada a inteligência artificial *BingAI*, impulsionada pelo motor *GPT-4* desenvolvido pela *OpenAI*. Com isso, foi possível a criação de um novo campo na tabela, apresentando uma descrição de cada *cluster*. Essa descrição foi fundamentada no nome atribuído ao *cluster* e na frequência relativa de cada palavra do seu agrupamento.

Afim de ilustrar a etapa anterior, a Tabela 2 possui os campos "*Cluster*"(referente ao código do agrupamento), "Nome"e "Descrição"(gerada pela *IA*), sendo possível observar a correta relação entre os campos nome e descrição, como por exemplo, o *Cluster* 7 possui as palavras "*program*", "*problem*", "*linear*"e "*optim*"que resultaram na descrição "programação e modelagem de problemas de otimização linear".

Tabela 1 – Nome e descrição dos *clusters* 1, 7 e 13.

Cluster	Nome	Descrição
1	protein 153 articl 129 antibodi 121 studi 116 human 111 sequenc 103 analysi 97 peptid 96 cell 94 control 88 nonhuman 84 bind 82 anim 82 modecular 80 acid 80	“Estudo de Artigos sobre Proteínas e Anticorpos Humanos e Não-Humanos: Análise e Sequenciamento de Peptídeos e Células com Métodos Moleculares e Ácidos”
7	program 536 integ 464 problem 404 linear 377 optim 305 comput 247 model 221 mix 218 algorithm 201 solut 183 solv 177 schedul 176 constraint 174 method 139 formul 138	“Programação e Modelagem de Problemas de Otimização Linear: Uso de Algoritmos e Métodos Computacionais para Solucionar e Integrar Soluções de Restrições e Agendamento”
13	system 218 network 186 comput 154 optim 145 servic 128 algorithm 118 use 114 resourc 111 problem 108 model 108 perform 106 propos 100 data 98 manag 96 applic 96	"Otimização, Modelagem de Sistemas e Redes de Computação: Uso de Algoritmos e Recursos para Melhorar o Desempenho, Gestão de Serviços e Aplicações de Dados."

Fonte: Elaborado pelo autor (2023)

Ao analisar mais detalhadamente os artigos presentes no *Cluster* 7, nota-se que, embora o título do artigo não o remeta diretamente a descrição do *Cluster*, uma análise das palavras-chave e do resumo revela a coesão entre eles. A Tabela 2 apresenta as informações de "DOI"e "Título", além de "Dados Relevantes - *Keywords*", que possui algumas palavras-chave que se mostraram relacionadas ao *Cluster*, e o campo "Dados Relevantes - *Abstract*", que contém trechos que evidenciam a correta classificação do artigo no *Cluster* em questão.

Tabela 2 – Artigos do cluster 7.

"Programação e Modelagem de Problemas de Otimização Linear: Uso de Algoritmos e Métodos Computacionais para Solucionar e Integrar Soluções de Restrições e Agendamento"			
DOI	Título	Dados relevantes - Keywords	Dados relevantes - Abstract
10.1287/opre.1120.1050	Um novo algoritmo para o problema de agendamento da produção de minas a céu aberto.	Algorithms Computer programming Heuristic methods Integer programming Optimization	"O agendamento de produção consiste em decidir quais blocos devem ser extraídos" "Blocos próximos à superfície devem ser extraídos primeiro, e restrições de capacidade limitam a produção em cada período de tempo." "Neste artigo, estudamos uma formulação conhecida de programação inteira do problema, que chamamos de C-PIT." "Propomos um novo método de decomposição para resolver a relaxação de programação linear (LP) de C-PIT quando há uma única restrição de capacidade por período de tempo."
10.1109/TPWRS.2003.814858	Colocação de deslocadores de fase em sistemas de grande escala via programação linear inteira mista.	Integer programming Linear programming	"Este artigo utiliza avanços recentes em programação linear inteira mista (MILP) para realizar um estudo de design preliminar sobre a colocação combinatória [...]" "Também leva em consideração limites ativos de fluxo e geração, além de restrições dos transformadores de mudança de fase."
10.1287/opre.1080.0548	Uma abordagem exata para o problema de layout de instalações unidimensionais.	Computational results Linear programs Mixed-integer programs Optimality Problem instances Quadratic programming models Quadratic programs	"[...] esse modelo é formulado como um programa misto-inteiro equivalente" "Em seguida, são introduzidas restrições redundantes adicionais e linearizadas em um espaço superior para obter um programa linear misto 0-1 equivalente." "É mostrado que o programa linear misto 0-1 resultante é mais eficiente do que as formulações mistas-inteiras previamente publicadas." "várias instâncias do problema retiradas da literatura foram eficientemente resolvidas até a optimalidade."

Fonte: Elaborado pelo autor (2023)

6 CONCLUSÕES E TRABALHOS FUTUROS

Esta monografia apresentou o processo e os resultados do agrupamento de artigos com base em seus resumos e palavras-chave. Resultados esses que foram alcançados por meio da aplicação de técnicas de normalização de texto, tratamento de dados, redução de dimensionalidade e clusterização, visando obter um agrupamento de dados mais precisa. Para a realização deste trabalho, os dados de mais de 6.000 artigos publicados nos últimos 20 anos na América latina foram analisados.

Com o tratamento adequado da base e aplicação dos algoritmos mencionados, foi possível formar 50 *clusters* de artigos, agrupando-os por sua similaridade e com base não apenas em suas palavras-chave, mas também em seus resumos. Após uma análise desses agrupamentos, foi possível constatar que os artigos foram, em sua maioria, corretamente agrupados, avaliando seus nomes, provenientes das 15 palavras mais frequentes de cada *cluster*, e o título dos artigos. Vale destacar que, ao lidar com conjuntos de dados maiores, o algoritmo pode produzir resultados satisfatórios sem a necessidade de um aumento proporcional no número de *clusters*, pois novos artigos podem ser incorporados aos *clusters* já existentes.

Toda a base de dados empregada nesse trabalho é composta por artigos publicados exclusivamente na língua inglesa. No entanto, devido aos artigos serem provenientes da América Latina, pretende-se ampliar esta monografia para incluir, além do inglês, os idiomas oficias da região, sendo eles: o português, o espanhol e o francês.

REFERÊNCIAS

AYALA-BASTIDAS, G.; CEBALLOS, H. G.; GARZA, S. E.; CANTU-ORTIZ, F. J. Identifying researchers' publication strategies by clustering publication and impact data. **Publishing research quarterly**, Springer US, v. 37, n. 3, p. 347–363, 2021. ISSN 1053-8801.

BéJAR, J. A. K-means vs mini batch k-means: a comparison. 2013.

CHANG, I.-C.; HORNG, J.-S.; LIU, C.-H.; CHOU, S.-F.; YU, T.-Y. Exploration of topic classification in the tourism field with text mining technologymdash;a case study of the academic journal papers. **Sustainability**, MDPI AG, v. 14, n. 7, p. 4053, 2022. ISSN 2071-1050.

FOLTZ, W. P. Latent semantic analysis for text-based research. **Behavior Research Methods, Instruments Computers**, n. 28, p. 197–202, 1996.

KANG, S.-S. Keyword-based document clustering. In: **Proceedings of the sixth international workshop on Information retrieval with Asian languages**. [*S. l.*: *s. n.*], 2003. p. 132–137.

MAIMON, O.; MAIMON, O.; ROKACH, L. **Data Mining and Knowledge Discovery Handbook**. Springer, 2005. ISBN 9780387244358. Disponível em: https://books.google.com.br/books?id=jizrAIWUJ6UC.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, v. 18, n. 5, p. 544–551, 2011. ISSN 1067-5027. Disponível em: https://doi.org/10.1136/amiajnl-2011-000464.

NUNES, D. H. F. **Um breve estudo sobre o algoritmo K-means**. Dissertação (Mestrado) – Universidade de Coimbra, 2016.

PORTER, M. An algorithm for suffix stripping. **Eletronic library and information systems**, v. 14, n. 3, p. 130–137, 1980.

SCULLEY, D. Web-scale k-means clustering. In: **Proceedings of the 19th International Conference on World Wide Web**. New York, NY, USA: Association for Computing Machinery, 2010. p. 1177–1178. ISBN 9781605587998.

SHAHAPURE, K. R.; NICHOLAS, C. Cluster quality analysis using silhouette score. **2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)**, p. 747–748, 2020.

WILLETT, P. The porter stemming algorithm: Then and now. **Program electronic library and information systems**, v. 40, 07 2006.