Extração automática de conhecimento tácito em biblioteca digital: o uso de mecanismos de inteligência artificial

Automatic extraction of tacit knowledge in a digital library: using artificial intelligence mechanisms

André Luiz de Castro Leal
Doutor em informática pela Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio; Professor na Universidade Federal Rural do Rio de Janeiro, UFRRJ, Seropédica, RJ, Brasil.
ORCID: https://orcid.org/0000-0002-8206-0992
E-mail: andrecastr@gmail.com

Sanderson Nascimento Milagres Filho
Graduando em Sistemas de Informação pela Universidade Federal Rural do Rio de Janeiro, UFRRJ, Seropédica, RJ, Brasil.
ORCID: https://orcid.org/0009-0006-2674-3349
E-mail: sanderson.milagres.filho@gmail.com

Lorena Vasconcellos Oliveira Magalhaes
Graduanda em Sistemas de Informação pela Universidade Federal Rural do Rio de Janeiro, UFRRJ, Seropédica, RJ, Brasil.
E-mail: lorenavasconcellos@ufrrj.br

Weslei de Carvalho Vianna
Graduando em Sistemas de Informação pela Universidade Federal Rural do Rio de Janeiro, UFRRJ, Seropédica, RJ, Brasil.
ORCID: https://orcid.org/0009-0003-7459-0402
E-mail: wesleivianna235@gmail.com

Gizelle Kupac Vianna
Doutora em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro, UFRJ; Professora da Universidade Federal Rural do Rio de Janeiro, UFRRJ, Seropédica, RJ, Brasil.
ORCID: https://orcid.org/0000-0001-8988-3329
E-mail: gkupac@gmail.com

Resumo

Objetivo: O presente estudo apresenta os resultados do esforço de pesquisa aplicada na construção de soluções de agentes inteligentes baseados em processamento de linguagem natural, mineração de texto e aprendizado de máquina, como também, construção de apresentação de dados visuais em grafos e listas, para extração e exibição de conteúdos de conhecimento tácito presente em artigos científicos base de dados uma biblioteca digital. **Metodologia:** O trabalho trata de pesquisa qualitativa e interpretativa a partir da interação entre pesquisadores e análise de resultados de dados extraídos. Metodologicamente baseia-se em *Design Science Research* visto que ele é metateórico. Epistemologicamente, o estudo fundamenta-se em *Design Science*, que cria conhecimento abordando como as pesquisas de caráter científico-tecnológico tratam concepção de artefatos para soluções de problemas. A prova de conceito efetuada a partir de técnicas de inteligência artificial para extração de conhecimento em dados de biblioteca de artigos científicos digitais trata de pesquisa de base de domínio aplicado. **Resultados:** Pesquisas efetuadas por agentes inteligentes computacionais apresentaram resultados satisfatórios, uma vez que, encontrados termos mais comumente utilizados nos artigos científicos escritos pelos autores, a mineração textual e algoritmos específicos agruparam autores (*clustering* por aprendizado não supervisionado de máquina) de mesma afinidade de pesquisa de acordo com assuntos de interesse. **Conclusão:** Dessa forma, a partir dessa pesquisa, uma biblioteca digital previamente programada com estruturas tradicionais de entrada e apresentação de dados, passou a contar com uma conjunto agentes inteligentes executam processos que vão além de apresentação resultados de filtros triviais encontrados em bibliotecas, como pesquisas por autor, evento, assunto, palavras-chave.

Palavras-chave: biblioteca digital; inteligência artificial; aprendizado de máquina; mineração de texto; clusterização; grafos de conhecimento.

Abstract

Objective: This study reports the results of applied research to build intelligent-agent solutions based on natural language processing, text mining, and machine learning, along with visual data presentations (graphs and lists), to extract and display tacit knowledge contained in scientific articles within a digital library database. **Methodology:** The study is qualitative and interpretive, arising from interactions among researchers and the analysis of extracted data results. Methodologically, it is grounded in Design Science Research, as this approach is metatheoretical. Epistemologically, the study is based on Design Science, which generates knowledge by examining how science-and-technology research designs artifacts to solve problems. The proof of concept—using AI techniques to extract knowledge from data in a digital library of scientific articles—constitutes applied, domain-based research. **Results:** Searches conducted by computational intelligent agents yielded satisfactory outcomes: after identifying the terms most commonly used by the authors in their scientific articles, text mining and specific algorithms grouped authors with similar research interests (clustering through unsupervised machine learning). **Conclusion:** As a result of this research, a digital library previously built with traditional data input and display structures now includes a set of intelligent agents that perform processes beyond the trivial filters typically offered—such as searches by author, event, subject, or keywords.

Keywords: digital library; artificial intelligence; machine learning; text mining; clustering; knowledge graphs.

1. Introdução

Bibliotecas Digitais (BDs) para armazenamento de artigos científicos constitui assunto atual. O acesso universal à internet e a prática da pesquisa em repositórios digitais que armazenam estudos científicos têm sido feito por pesquisadores dos mais diversos domínios do conhecimento. Esses repositórios compõem websites especializados em armazenar o material produzido por autores de diversas instituições ao redor do mundo (Ioannidis, 2001).

Algumas BDs possuem o seu conteúdo restrito, em que o acesso aos documentos só pode ser feito através da autenticação de um membro associado, como é o caso do portal de periódicos da CAPES (Periódicos CAPES, 2024). Outra categoria de BDs é aquela de conteúdo pago, como a ACM Digital Library (ACM, 2025), onde algumas informações sucintas dos materiais, como *abstract*, títulos, autores, são disponibilizadas gratuitamente, mas, para ter acesso completo ao material, é necessário dispor de recursos financeiros.

Conhecidos também são os mecanismos de busca como o Scholar Google (Scholar Google, 2025) e a DBLP Computer Science Bibliography (Ley, 2002) que respondem com listas de títulos, a partir das buscas realizadas pelos usuários a partir de strings, ou trechos textuais, de busca.

Nesse contexto de bibliotecas digitais, pesquisadores de diversas áreas de conhecimento precisam encontrar artigos científicos correlatos ao seu domínio de estudo e, a partir do

resultado dessa consulta, realizar a leitura, interpretação, análise e seleção dos conteúdos escritos nesses artigos que sejam mais adequados ao seu interesse de estudo.

Para Wu (2023), a dificuldade de se lidar com a triagem de literatura a partir de mecanismos não automáticos é um desafio para os pesquisadores, uma vez que os conteúdos podem estar armazenados em grandes volumes de dados e com especificidades que requerem algoritmos especializados para a descoberta do conhecimento. Portanto, faz-se necessário o desenvolvimento de ferramentas baseadas em Inteligência Artificial (IA) para automatizar esse processo.

Desenvolver BDs que dispõem apenas de funcionalidades simples, como inclusão, exclusão e consultas básicas de artigos cadastrados, já não mais constituem desafios para os engenheiros de software. Em uma realidade em que as bibliotecas digitais reúnem um grande número de documentos, a necessidade de ferramentas que agregam potencial às consultas é inevitável, impulsionando, assim os avanços na construção de soluções que possam agregar valor às pesquisas feitas nessas bibliotecas.

O presente estudo, experimentação aplicada, apresenta os resultados do esforço de pesquisa aplicada à construção de soluções de agentes inteligentes baseados em processamento de linguagem natural, mineração de texto e aprendizado de máquina, como também, construção de apresentação de dados visuais em grafos e listas, para extração e exibição de conteúdos de conhecimento tácito presente em artigos científicos base de dados uma biblioteca digital. Como também organiza textualmente, após estudo da literatura científica, conceitos a respeito de técnicas de IA. Dessa forma, o trabalho apresenta os resultados alcançados no desenvolvimento e execução de soluções automáticas baseadas em IA, para auxiliar os esforços de cientistas na seleção de conteúdos extraídos de artigos científicos armazenados na BD em formato PDF. Técnicas de mineração de dados, tabulações e geração de grafos de conhecimento foram inseridos nos algoritmos de descoberta dos conhecimentos implícitos para auxiliar os pesquisadores a acessar resultados descobertos de forma automatizada.

As seções do artigo estão organizadas da seguinte forma: a seção 2 apresenta os pressupostos teóricos; a seção 3 aborda os aspectos metodológicos; a seção 4 trata a apresentação dos algoritmos de mineração de texto e a clusterização dos autores dos artigos científicos a partir de palavras extraídas na mineração; na seção 5 são apresentados resultados

visuais em grafos e listas dos conteúdos extraídos e o agrupamento de conhecimento por autores com afinidade de pesquisas científicas; e, por fim, na seção 6, estão as conclusões.

2. Pressupostos teóricos

Com relação à revisão da bibliografia, o presente trabalho não faz estudo sistemático da bibliografia, dado principalmente que trata de estudo prático de experimentação aplicada, portanto não se baseia em estudos como mapeamento sistemático ou revisão sistemática da literatura (Kitchenham; Charters, 2007) (Snyder, 2019) (Petersen; Feldt; Mujtaba; Mattsson, 2008) (Petersen; Vakkalanka; Kuzniarz, 2015). No entanto, , o estudo debruça-se em principais autores da literatura científica, além da experiência dos cientistas envolvidos, o que dá base para a construção de uma solução ainda pouco explorada em bibliotecas digitais ativas e disponíveis em internet para uso no cenário nacional.

As bibliotecas digitais representam um sistema de armazenamento e recuperação de informações especializadas, que manipulam dados digitais em diferentes formatos existentes, disponibilizados em redes distribuídas, como texto, som, imagens estáticas ou dinâmicas (Saffady, 1995). Também constituem um conjunto de recursos eletrônicos e capacidades técnicas associadas para criação, pesquisa e uso da informação (Borgman, 1999).

Essas bibliotecas são construídas, coletadas e organizadas por comunidades de usuários, e suas capacidades funcionais usadas para apoiar as necessidades de informação desse grupo. Como exemplo de comunidade, há as de especialistas e pesquisadores envolvidos em estudos científicos na área de ciência da informação, computação, medicina ou outra área de conhecimento (Ioannidis, 2001) (Borgman, 1999). Dessa forma, elas devem ter facilidade de uso e ser eficientes para transpor barreiras de distância, língua e cultura, além de permitirem o acesso por diferentes dispositivos com acesso à internet.

Para autores como Ioannidis (2001), Lesk (1997) e Greenstein (2000), as BDs devem oferecer serviços integrados que auxiliem a compreensão da coleção de dados digitais, respeitando quesitos como a completude e diversidade das fontes, prover informações heterogêneas e permitir o uso colaborativo, possuir filtros específicos e consultas que permitam a verificação de listas de resultados e tabelas, além de possibilitar o armazenamento de dados de forma estruturada ou semiestruturada.

O desenvolvimento de soluções em bibliotecas digitais modernas vai além do armazenamento transacional e consultas com filtros básicos. Nesse contexto, dentre as atribuições dos engenheiros de software, dada a BD uma plataforma de software, encontra-se estudos e projetos no âmbito de extração e análise automatizada da informação, para apresentação de conhecimentos implícitos nos conteúdos registrados nas bibliotecas.

Atualmente, esses conteúdos estão armazenados em escala de grandes volumes de dados e o conhecimento implícito nos artigos e conteúdos dessas bibliotecas necessitam de uma abordagem automática de verificação e análise, uma vez que se torna humanamente impossível a extração, correlação e apresentação de dados desse volume de informações.

Particularmente nesse contexto de análise automática de dados, a IA, técnicas de processamento de linguagem natural (PLN), mineração de texto (MT) e aprendizado de máquina (machine learning) (ML) constituem, além de pesquisa de vanguarda, uma solução viável para aplicação nesse domínio de estudo, que trata a análise de grandes volumes de dados.

O PLN pressupõe o desenvolvimento de modelos e algoritmos para compreender, interpretar, gerar e manipular texto em linguagem natural, em uma nítida interação entre computadores e linguagem humana (Norvig; Russel, 2022) (Kodratoff, 1999). Estão nesse contexto de soluções tarefas como análise sintática e semântica, tradução, correção gramatical, recuperação de informações, auxílio à escrita e ao aprendizado de línguas, reconhecimento de entidades nomeadas, sumarização de texto, tradução automática, entre outras.

A mineração de texto (MT) é uma área de pesquisa que visa extrair informações úteis e relevantes de grandes quantidades de texto não estruturado, utilizando técnicas e algoritmos de PLN (Kodratoff, 1999; Witten; Moffat; Bell, 1999; Dörre; Gerstl; Seiffert, 1999). Ela possui forte ligação com as áreas de recuperação de informação e processamento de linguagem natural, que podem atuar sobre um conjunto de textos para recuperar informações que atendam ao conjunto de condições fornecido pelo usuário ou ainda para comparar documentos, buscando no conjunto disponível àqueles semelhantes a um outro documento fornecido.

A análise realizada para essas buscas é feita, geralmente, sobre estatísticas acerca das palavras de cada texto. Também existem as análises sintática e semântica, que dão importância à função das palavras em uma frase (Faceli, 2022). Como técnicas aplicadas nesses algoritmos de MT incluem a leitura automática sistematizada, a remoção das *stopwords*, a remoção de algarismos e pontuações, a análise léxica e tokenização, a normalização de palavras, o *stemming*

(estemização) e a lematização. As aplicações dos algoritmos de MT envolvem análise estatística, classificação de documentos, agrupamento, extração de padrões textuais, redundâncias e extração de conhecimentos tácitos.

Associada ao PLN e MT, o aprendizado de máquina constitui uma solução automática para auxiliar o treinamento algorítmico e categorização de informações, em *clusters*, por exemplo. Aplicada à análise textual, o objetivo do ML é representar cada texto por um vetor numérico, porém, como os dados contidos em textos não são estruturados, eles precisam ser convertidos para o formato atributo-valor antes da aplicação das técnicas convencionais de ML.

Depois que os documentos são coletados, uma etapa importante é a tokenização, que segmenta o texto em termos (tokens), que são conjuntos de caracteres alfanuméricos, separados por limitadores (normalmente caracteres de formatação, espaços em branco ou caracteres de pontuação) (Faceli, 2022). Esse processo é um passo inicial que prepara segmentos do texto para todas as demais etapas e operações de PLN.

Na aplicação da abordagem estatística, um primeiro passo comum é eliminar variações de uma mesma palavra, associando cada uma à sua raiz (*stem*), para, em seguida , construir uma representação estruturada do texto, na forma raiz-frequência (ou seja, atributo-valor). O processo de redução de um termo ao seu radical (estemização) irá variar de acordo com o idioma, sendo o algoritmo de Porter (1980) um dos mais usados para documentos escritos em língua inglesa, enquanto **o Removedor de Sufixos da Língua Portuguesa** (RSLP) é muito comum em trabalhos sobre textos em português.

As *stopwords* são as raízes de palavras que ocorrem com muita frequência em todos os textos, como preposições, conjunções e artigos, e precisam ser eliminadas, para não inserir ruídos na representação do texto. Por outro lado, palavras que ocorram muito raramente também são eliminadas, por serem pouco representativas ou discriminativas. Ao final do processo, realiza-se **a lematização**, ou seja, processo de redução de um termo à sua forma canônica.

Em sequência, a **categorização de textos** atribui categorias (rótulos ou classes) a um documento escrito em linguagem natural. Quando se utiliza técnicas supervisionadas de ML para criar um modelo de classificação, normalmente parte-se de um conjunto rotulado de segmentos de textos – trabalho feito manualmente por especialistas da área – que será utilizado para mensurar a proximidade entre um texto-alvo com conjuntos já categorizados.

3. Aspectos metodológicos

A pesquisa segue uma abordagem de perfil qualitativo, visto que se enquadra na visão de Creswell (2014), onde o assunto envolve mais a natureza interpretativa da investigação e considera a interação presencial dos pesquisadores nos relatos que eles apresentam. Utiliza-se da abordagem de *Design Science Research* como método de pesquisa, visto que este é metateórico, dentro da base epistemológica *Design Science* (DS), que cria conhecimento abordando como as pesquisas de caráter científico podem ser construídas através da concepção de artefatos, estruturando a pesquisa científica com base tecnológica para soluções de problemas (Bax, 2015; Hevner, 2010). O DS tem por objetivo trabalhar com tudo que é projetado pelo homem, e assim, ela também ficou reconhecida por ser a ciência do artificial. Tem também um caráter exploratório, visto que ela tem objetivo de tornar a oportunidade (problema) proposta mais explícita e também elaborar hipóteses mais objetivas e intuitivas.

O presente estudo é parte do esforço de se acrescentar funcionalidades baseadas em IA em uma BD construída pelos autores com funcionalidades básicas de cadastramento de eventos, autores, artigos, palavras-chave, entre outras informações, mas que passa a contar com abordagens de reconhecimento automático de conhecimento tácito contido nos artigos publicados . Portanto, é estudo que constrói algoritmos com base em técnicas de IA, organiza código fonte digital, software, e aplica técnicas de clusterização a partir de aprendizado não supervisionado, para extrair conhecimento tácito dos textos de artigos publicados por autores. Com isso, agrupa autores com perfis de pesquisa semelhantes, o que sugere futuramente outras aplicações como de construção de algoritmos de recomendação, por exemplo, trabalhos comuns, livros, editais de fomento, localização, dado que consegue agrupar dados de localização já disponíveis na biblioteca, entre outros. Dessa forma, a pesquisa constitui-se como primeiro passo para uma série de desdobramentos científicos que possam ser explorados

Foram realizadas construções de algoritmos que atenderam às seguintes demandas de: i) PLN e MT para extrair palavras significativas com maior número de ocorrências de arquivos armazenados em PDF; ii) PLN e Clustering para, a partir das palavras significativas descobertas, agrupar autores de artigos distintos em mesmo cluster; iii) apresentar resultados de autores agrupados, bem como assuntos de interesse, como listas e gráficos em tela de exibição na BD. Os algoritmos processaram mais de 200 artigos em formato PDF disponível na biblioteca, importados de uma conferência de edição 2023.

De uma forma geral , a Figura 1 apresenta os passos da pesquisa, no que se refere ao estudo bibliográfico, a construção dos artefatos e os resultados.

Figura 1 – Representação do fluxo principal das atividades da pesquisa.

Fonte: Elaborado pelos autores.

Além desses passos, a pesquisa foi planejada de forma que a organização das atividades executadas e suas respectivas entregas, o produto de cada ação, orientam a utilização do método proposto. Desta forma, as etapas foram organizadas seguindo o modelo sugerido por van Aken & Romme (2009), a saber: planejamento e entendimento detalhado do problema de pesquisa; revisão bibliográfica; síntese da pesquisa; proposição do *design*, com a definição da abordagem, no caso utilizado do método, com foco na solução de problemas práticos; e avaliação dos artefatos, onde foi realizada a interatividade entre as pesquisas qualitativas e interpretativas , de cunho exploratório, e os artefatos propostos.

4. A extração do conhecimento tácito

A presente seção apresentará os resultados dos esforços de construção de algoritmos que permitiram ações para a descoberta e apresentação de resultados relacionados a descoberta de conhecimento tácito extraído da BD.

A primeira subseção apresentará os detalhes do algoritmo de MT; na segunda, serão apresentados os resultados sobre o agrupamento de autores de artigos da BD que obtiveram similaridade de pesquisa entre seus diferentes artigos; nas subseções três e quatro, serão apresentados resultados gráficos ou listas com conteúdos a partir do conhecimento extraído.

4.1 Algoritmo de mineração de textos

Os algoritmos de MT, construídos para a leitura dos artigos em formato PDF, armazenados na biblioteca foram desenvolvidos utilizando linguagem Python e as bibliotecas nltk.tokenize, nltk.corpus, nltk.stem, PyPDF2, collections.

A linguagem foi escolhida por possuir um conjunto de características necessárias para o estudo aqui realizado. Em especial, podemos citar as seguintes: ser open source, de fácil modularização, ser multiplataforma, possuir um grande conjunto de bibliotecas disponíveis e, para o caso da prova de conceito desse estudo, possuir funções e bibliotecas úteis para a função.

As bibliotecas Python desempenharam um papel essencial, ao disponibilizar módulos, métodos e funções para tarefas de leitura de arquivos PDF, tokenização, remoção de stopwords, lematização, conversão de tipos de variáveis ou objetos, geração de gráficos de palavras e exportação de dados no formato desejado. Embora construído de forma monolítica, o algoritmo apresenta seções claras específicas a cada momento de leitura de bibliotecas, transformação de dados, conversões, execuções em laço, entre outras. As instruções escritas gerenciam as bibliotecas, funções, módulos e métodos necessários para os processamentos. O trecho de código a seguir demonstra esse percurso.

```
import os
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from PyPDF2 import PdfReader
import matplotlib.pyplot as plt
from collections import Counter
import pandas as pd
import string
```

No trecho de código abaixo, foram importados os módulos necessários para o treinamento dos agentes inteligentes (punkt da biblioteca nltk), remoção de stopwords (palavras e caracteres de baixa relevância para o estudo, assim como “,”, “.”, “a”, “que”, “de”, “e”, “-”, “o”, “;”, “me”, “um”) e a construção de uma base de dados de conhecimento léxico em língua portuguesa, para a comparação e identificação de termos dentro do conjunto textual de artigos lidos.

```
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

Em seguida, apresenta-se o código utilizado para abertura e leitura dos textos dos artigos armazenados em formato PDF.

```
# Leitura dos pdfs
def pdf_reader(folder_path):
    texts = []
    for file_name in os.listdir(folder_path):
        if file_name.endswith('.pdf'):
            file_path = os.path.join(folder_path, file_name)
            with open(file_path, "rb") as file:
                pdf_reader = PdfReader(file)
                text = ""
                for page_number in range(len(pdf_reader.pages)):
                    page = pdf_reader.pages[page_number]
                    text += page.extract_text()
            texts.append(text)
    return texts
```

A MT foi realizada a partir da tokenização dos termos (processo de conversão de uma sequência de texto em partes menores, conhecidas como tokens), seguida da remoção de termos irrelevantes, da lematização (processo de redução das palavras ao seu radical) e da conversão dos tokens lematizados em strings, que são, depois, usados para comparação de termos.

```python
# Mineiração de texto
def process_text(text_pdf):
    # Tokenização
    tokens = word_tokenize(text_pdf.lower())

    # Remove stop words em português, em inglês e stopwords personalizadas
    filtered_tokens = [token for token in tokens if token not in stopwords.words('portuguese')
    and token not in stopwords.words('english')     and token not in custom_stopwords]

    # Remove pontuações
    punctuation = string.punctuation
    filtered_tokens = [token for token in filtered_tokens if token not in punctuation]

    # Remove números e palavras com menos de 2 caracteres
    filtered_tokens = [token for token in filtered_tokens if not token.isdigit()
    and (token.isalpha() and len(token) > 2)]

    # Lematização
    lemmatizer = WordNetLemmatizer()
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in filtered_tokens]

    # Converte os tokens para string
    processed_text = ' '.join(lemmatized_tokens)

    return processed_text
```

Após o processamento da MT, é realizada a extração dos termos mais citados em cada artigo, que são usados para a criação de uma tabela contendo a caracterização de cada artigo. A construção dessa tabela, por sua vez, foi pensada para facilitar a futura identificação de autores com convergência em áreas de pesquisa.

Foi estabelecido, empiricamente, a quantidade de 50 palavras mais frequentes para cada artigo. O trecho de código a seguir realiza a contabilização das palavras mais frequentes de um artigo, seguida da geração de uma tabela que apresenta o nome do arquivo que contém o artigo e as palavras mais frequentes suas respectivas frequências de ocorrência.

```python
# Seleciona as 50 palavras mais citadas em cada pdf
def get_top_words_per_pdf(texts):
    top_words_per_pdf = []
    for text in texts:
        processed_text = process_text(text)
        word_counts = Counter(processed_text.split())
        top_words = word_counts.most_common(50)
        top_words_per_pdf.append(top_words)
    return top_words_per_pdf

# Gera um gráfico com as palavras
def create_dataframe(top_words_per_pdf):
    data = {'PDF': [], 'Palavra': [], 'Frequência': []}
    for i, top_words in enumerate(top_words_per_pdf):
        for word, freq in top_words:
            data['PDF'].append(f'PDF_{i+1}')
            data['Palavra'].append(word)
            data['Frequência'].append(freq)
    return pd.DataFrame(data)

# Exporta o gráfico como arquivo csv
def export_dataframe_to_csv(df, filename):
    df.to_csv(filename, index=False)
```

Após a geração da tabela (não apresentada nesse artigo devido ao seu tamanho), seus dados são exportados para um arquivo em formato texto (CSV), no formato apresentado nas linhas a seguir.

```
PDF_193,governança,59
PDF_193,compra,56
PDF_193,pública,52
PDF_193,público,49
PDF_193,social,47
PDF_193,pública,47
PDF_193,transparência,47
PDF_193,atuação,43
```

As linhas do arquivo apresentam o nome do artigo, o termo descoberto e a frequência de ocorrência.

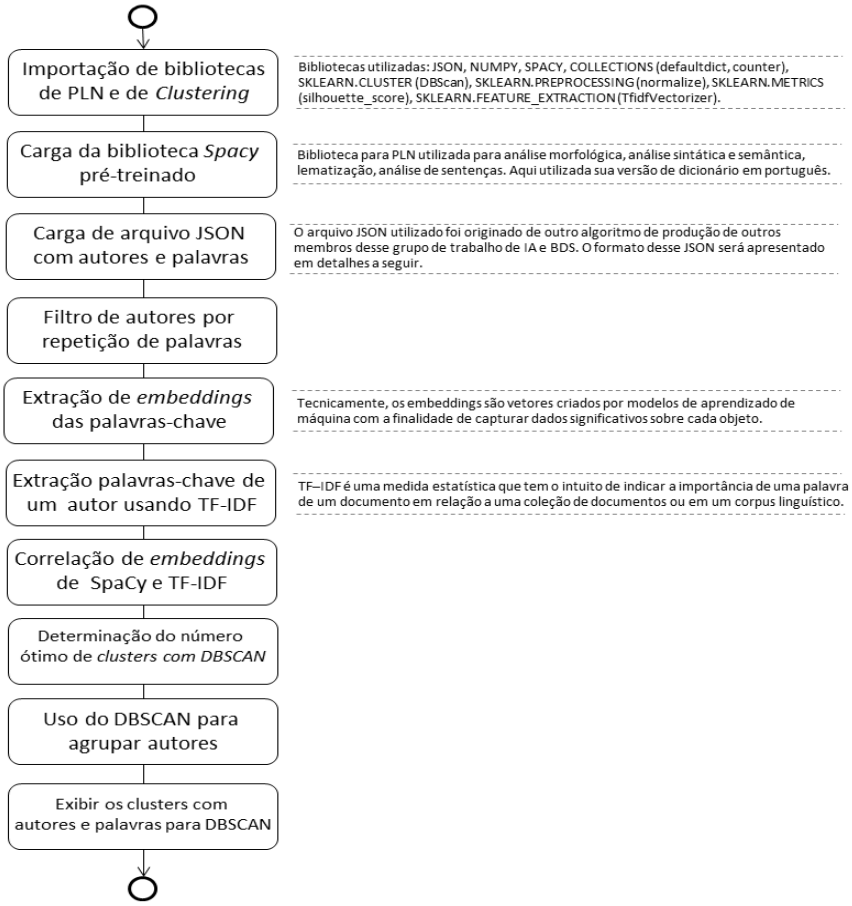
4.2. Clustering de autores

Após a execução do algoritmo de MT, a contabilização de palavras por frequência de ocorrências no texto possibilitou o agrupamento (em clusters) de autores por assuntos correlacionados. Com as palavras significativas mais utilizadas por esses autores, foi possível correlacionar autores de diferentes artigos por essas palavras.

O processo algorítmico para realizar essas tarefas, também feito a partir de bibliotecas em linguagem Python, permitiu que os clusters formados com os autores e seus descritores fossem exibidos para análise dos pesquisadores. Para cada cluster, foram exibidos os autores e os descritores mais comuns associados a eles.

A Figura 2 apresenta o processo e apresenta as principais bibliotecas, módulos ou métodos utilizados pelos pesquisadores para se trabalhar o PLN e Clustering de autores dos artigos científicos disponíveis na BD.

Figura 2 – Representação da construção do algoritmo de PLN e Clustering.



Fonte: Elaborado pelos autores.

Com a execução do algoritmo de MT, foi possível a identificação de alguns clusters de autores, ou seja, aqueles com maior potencial de possuírem pesquisas similares e portanto assuntos de pesquisa em comum. Porém para um primeiro caso, foi possível perceber que autores ficaram isolados em um cluster por não haver outros artigos identificados com a mesma caracterização de palavras. Particularmente esses clusters foram criados por autores do mesmo artigo. Os clusters abaixo representam esse fato.

```
Cluster 2:
  Autores: ['autor 34', 'autor 35', 'autor 36']
  Palavras: ['ensino']

Cluster 3:
  Autores: ['autor 47', 'autor 48', 'autor 49', 'autor 50', 'autor 51']
  Palavras: ['gestão']

Cluster 7:
  Autores: ['autor 134', 'autor 135', 'autor 136']
  Palavras: ['metodologia']

Cluster 11:
  Autores: ['autor 157', 'autor 158', 'autor 159']
  Palavras: ['governança', 'prática']

Cluster 15:
  Autores: ['autor 188', 'autor 189', 'autor 190']
  Palavras: ['clube']
```

Em outro conjunto de agrupamentos, foi possível observar que autores de diferentes artigos foram agrupados por semelhança de frequência de palavras, com apenas uma palavra como interseção, apesar de suas inúmeras frequências.

```
Cluster 0:
  Autores: ['autor 9', 'autor 10', 'autor 301', 'autor 302', 'autor 303']
  Palavras: ['empresa']

Cluster 1:
  Autores: ['autor 13', 'autor 14', 'autor 15', 'autor 502', 'autor 503', 'autor 504', 'autor 507']
  Palavras: ['desenvolvimento']

Cluster 41:
  Autores: ['autor 168', 'autor 455', 'autor 456']
  Palavras: ['saúde']

Cluster 44:
  Autores: ['autor 399', 'autor 358', 'autor 488']
  Palavras: ['jogo']
```

Os autores “autor 84”, “autor 177”, “autor 178” e “autor 179” foram agrupados, uma vez que seus artigos foram escritos a partir do viés de interesse em processos e protocolos organizacionais, identificados por leitura feita pelos pesquisadores, mas também evidenciados pelo algoritmo a partir da identificação dessas palavras “protocolo” e “processo”. Tal fato pode indicar que os assuntos relacionados a questões organizacionais orientadas a processos e protocolos podem interessar esses autores.

```
Cluster 12:
Autores: ['autor 84', 'autor 177', 'autor 178', 'autor 179']
Palavras: ['processo', 'protocolo']
```

O algoritmo de clusterização apresentou resultados satisfatórios no agrupamento de autores de artigos científicos e permitiu a construção de visualizações na BD, auxiliando a seus usuários a perceberem alguns detalhes de conhecimento além de filtros básicos. Nas seções seguintes, serão apresentados alguns desses resultados.

5. Apresentação dos resultados

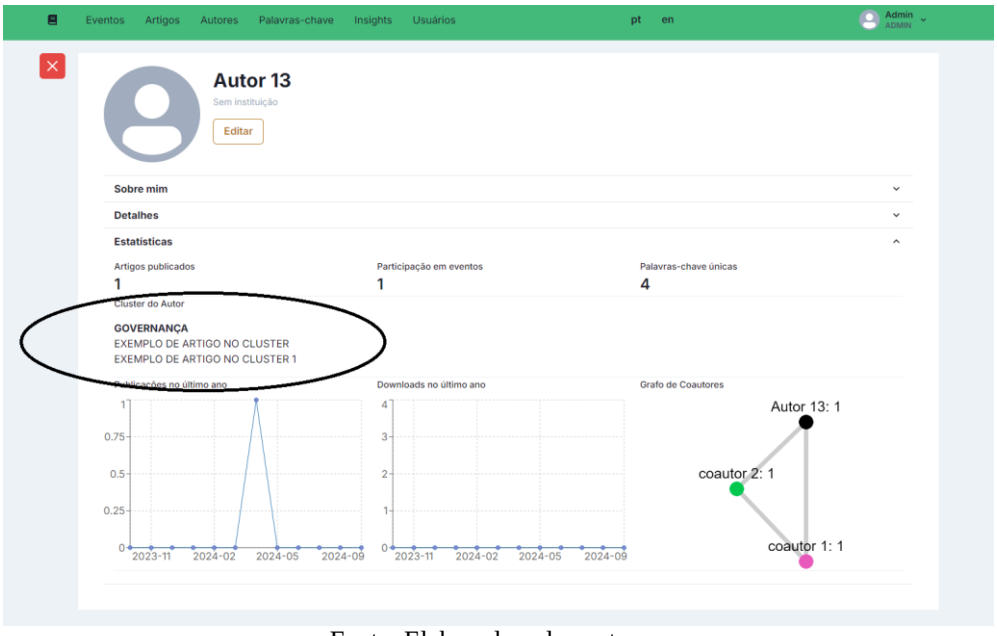
5.1. Lista de artigos por cluster de autores

O primeiro estímulo dos pesquisadores desse estudo foi utilizar o conhecimento descoberto com os clusters e recomendar aos usuários da biblioteca, uma lista de artigos de outros autores que estivessem no mesmo cluster do autor selecionado.

A Figura 3 apresenta a tela de consulta de autor na BD. Ela possui apresentações de resultados, sendo uma delas a lista de artigos por cluster comum ao autor pesquisado.

Na tela, há o autor selecionado pelo usuário, cujo o nome original foi substituído por "Autor 13", bem como a lista de recomendações de artigos de outros autores, destacada com um círculo em torno da mesma. No caso, os artigos foram selecionados do cluster "Governança". Na biblioteca, esses artigos possuem um hiperlink, possibilitando ao usuário o acesso direto a página relacionada da BD.

Figura 3 – Recomendação de Artigos por Cluster.



Fonte: Elaborado pelos autores.

Outros detalhes nessa tela não estão diretamente relacionados com os conhecimentos extraídos a partir dos algoritmos de IA elaborados nesse trabalho. Eles constituem apenas quantificadores de publicação do autor por mês/ano, número de downloads do artigo desse autor, número de artigos publicados e eventos que participou e estão registrados na BD, número

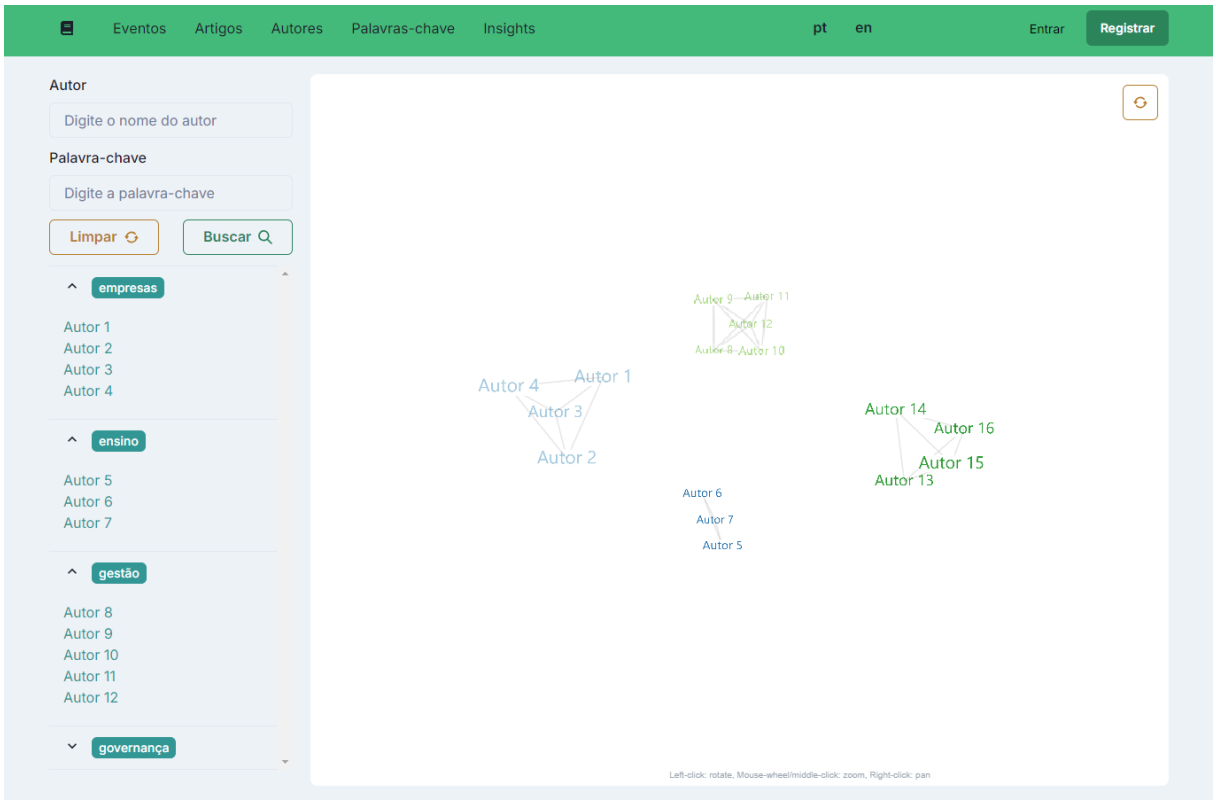
de palavras-chaves utilizadas pelo autor em seus artigos e um grafo de coautores que publicaram artigos com o “Autor 13”.

5.2. Grafos de conhecimento de clusters formados pela IA

Visando possibilitar a consulta direta aos grupos de autores, uma interface foi criada na biblioteca, possibilitando pesquisas por um autor específico de um grupo ou, então, a visualização geral. A apresentação está disponibilizada em forma de lista e em forma de grafos para todos os clusters, destacando as palavras do cluster em que os autores participam.

A Figura 4 apresenta, em seu lado esquerdo, para cada cluster encontrado pelo algoritmo de IA, a lista de autores incorporados. Já ao centro, no quadro branco, o grafo de todos os clusters. No exemplo abaixo, foi deixado o filtro de autor em branco e feita uma pequena amostra da possibilidade dos resultados. É possível perceber os vários clusters criados, com autores distintos em cada cluster.

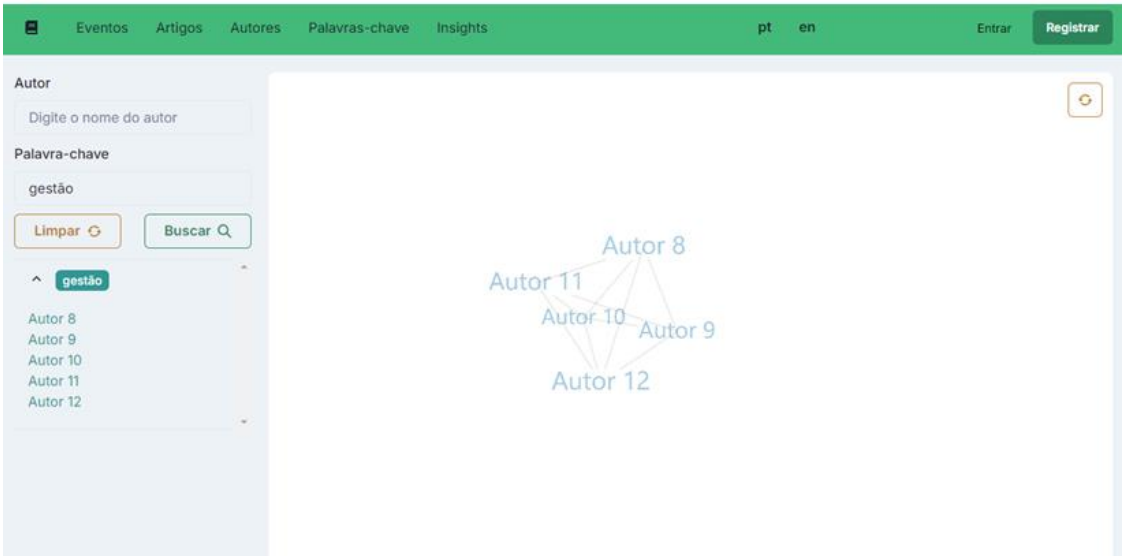
Figura 4 – Grafos de Clusters de Autores.



Fonte: Elaborado pelos autores.

Filtros de autor e palavra-chave podem ser utilizados para reduzir a consulta e ir direto aos conteúdos de interesse do usuário. A Figura 5 apresenta os resultados, com o filtro de palavra-chave indicando o termo “gestão”. O mesmo pode ser feito ao indicar o nome de um autor para pesquisa. Portanto, todos os autores apresentados na lista (à esquerda na tela) e no grafo são autores que publicaram artigos reconhecidos pela IA com domínio de estudo voltado à gestão.

Figura 5 – Grafos de Clusters de Autores com Filtro de Palavra-Chave.



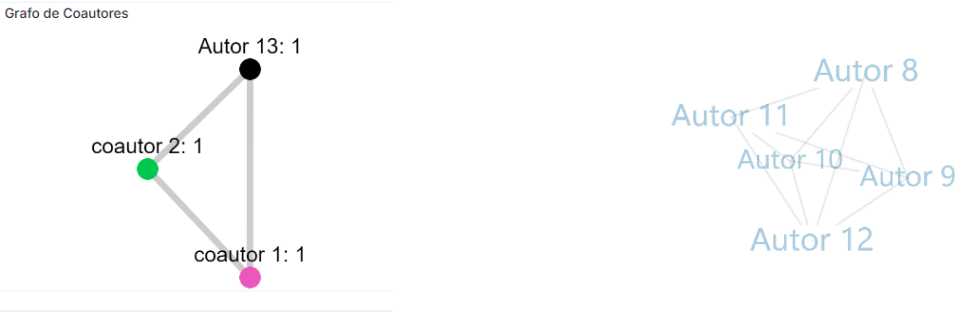
Fonte: Elaborado pelos autores.

Nos exemplos, os nomes verdadeiros dos autores foram substituídos por expressões aleatórias. A plataforma, a partir da lista de autores, permite que seja selecionado o hiperlink no nome do autor que encaminhará o usuário diretamente para a tela de consulta aos dados do autor.

5.3. Diferença de grafos disponíveis na BD

Há na biblioteca dois tipos de grafos gerados até o momento (Figura 6).

Figura 6 – Grafo Autor e Coautores x Grafo de Autores de Clusters.



Fonte: Elaborado pelos autores.

O primeiro deles é formado pelos nomes do autor e coautores que publicaram artigos conjuntamente. Ou seja, os pesquisadores desenvolveram um algoritmo que pesquisa os artigos do autor e cria um grafo com todos os coautores com quem publicou. No segundo, o grafo gerado demandou uma sistemática algorítmica especializada, onde todo o contexto de PLN, MT e ML foi necessário para identificar, agrupar e gerar os conteúdos. Dessa forma, não bastou uma leitura por varredura textual no corpo do texto do artigo para identificar autores e coautores publicaram juntos. Houve a necessidade de sistemática de identificação de palavras comumente utilizadas por esses autores, a partir dos agentes inteligentes de IA, para se descobrir quais autores poderiam ter similaridade em seus estudos científicos.

6. Considerações finais

O presente artigo apresentou os resultados do esforço de pesquisa para implementar soluções de agentes inteligentes baseados em técnicas de IA, como PLN, MT, ML, assim como na construção da apresentação dos dados visuais em grafos e listas. O objetivo foi evoluir dos filtros comuns encontrados em bibliotecas digitais para inserção de algoritmos inteligentes destinados à descoberta de conhecimento tácito nas bases de dados com artigos científicos publicados.

Os resultados foram satisfatórios, uma vez que a categorização dos artigos foi possibilitada a partir das palavras mais comumente utilizadas nas publicações dos autores. Foi

possível agrupar esses autores de mesma afinidade de pesquisa de acordo com assunto de interesse inferido pelos algorítimos.

Esses esforços, principalmente com a escrita e evolução dos algoritmos de mineração de texto, permitiram, além da aquisição de expertise na mineração textual, a extração e o agrupamento de palavras de significativo contexto nas pesquisas dos autores. Dessa forma, trabalhos posteriores de agrupamento de autores e exibição de dados foram facilitados, o que evoluiu significativamente os conteúdos exibidos na BD desenvolvida pelo grupo de pesquisa responsável por esse estudo e que estará, em breve, disponível para usuários comuns.

Como trabalhos futuros, estão sendo desenvolvidos: i) seleção e importação dos estados e instituições de ensino de origem dos autores dos artigos para a base de dados da biblioteca; ii) a geração de grafos com informação dos estados e instituições de ensino com correlações de publicação conjunta e por afinidade de pesquisa; iii) a construção de algoritmo baseado em agentes inteligentes de mining para a descoberta da área de conhecimento de maior incidência dos artigos dos autores, sua correlação com áreas de avaliação da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior do Ministério da Educação (CAPES); iv) sistemas de recomendação.

Declaração de contribuição de autoria

Autor 1: escrita, rascunho original, metodologia, validação, conceitualização, análise formal, revisões finais.

Autor 2: escrita, rascunho original, validação, curadoria dos dados, codificação.

Autor 3: escrita, rascunho original, validação, curadoria dos dados, codificação.

Autor 4: escrita, rascunho original, validação, curadoria dos dados, codificação.

Autor 5: revisão de escrita, validação, metodologia, revisões iniciais.

Agradecimentos

Agradecimento especial à Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) pelo apoio incondicional às pesquisas docentes, nesse caso

pelas bolsas de iniciação tecnológica (IT), disponibilizadas a partir de edital 10/2023, e à Universidade Federal Rural do Rio de Janeiro (UFRRJ).

Referências

ACM DIGITAL LIBRARY. **Repository of resources**. Association for Computing Machinery ACM Inc. c2025. Disponível em: http://portal.acm.org. Acesso em: 18 abr. 2025.

BAX, M. P. Design science: filosofia da pesquisa em ciência da informação e tecnologia. **Ciência da Informação**, Brasília, v. 42, n. 2, ago. 2015. DOI: https://doi.org/10.18225/ci.inf.v42i2.1388. Disponível em: https://doi.org/10.18225/ci.inf.v42i2.1388. Acesso em: 24 abr. 2025.

BORGMAN, D. L. What are digital libraries? competing visions. **Information Processing & Management**, v. 35, n. 3, p. 227-243, 1999. DOI https://doi.org/10.1016/S0306-4573(98)00059-4. Disponível em: https://doi.org/10.1016/S0306-4573(98)00059-4. Acesso em: 24 abr. 2025.

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. **Plataforma Sucupira**. Brasília: CAPES, 2024 Disponível em http://www.periodicos.capes.gov.br. Acesso em: 18 maio 2025.

CRESWELL, J. W. **Investigação qualitativa e projeto de pesquisa:** escolhendo entre cinco abordagens. 3. ed. Porto Alegre: Penso, 2014.

DÖRRE, J.; GERSTL, P.; SEIFFERT, R.. Text mining: finding nuggets in mountains of textual data. *In:* ACM SIGKDD, 50., San Diego, 1999. **Proceedings**[…]. San Diego: ACM, 1999. p. 308-401. DOI https://doi.org/10.1145/312129.312299. Disponível em: https://dl.acm.org/doi/10.1145/312129.312299. Acesso em: maio 2025.

FACELI, K. **Inteligência artificial:** uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2022.

GREENSTEIN, D. Digital libraries and their challenges. **Library Trends**, v. 49, n. 2, p. 290-303, Fall 2000.

HEVNER, A.; CHATTERJEE, S. Design science research in information systems: theory and practice. **Design Research in Information Systems**, v. 2, p. 9-22, mar. 2010. DOI: https://doi.org/10.1007/978-1-4419-5653-8_2. Disponível em: https://doi.org/10.18225/ci.inf.v42i2.1388. Acesso em: 24 abr. 2025.

IOANNIDIS, Y. **Digital libraries**: future directions for a european research programme. 3. ed. Roma**:** DELOS, 2001.

KITCHENHAM, B. A.; CHARTERS, S. **Guidelines for perform-ing systematic literature reviews in software engineering**. UK: Tech. rep.: Keele University, 2007. 65p.

KODRATOFF, Y. Knowledge discovery in texts: a definition, and applications. *In:* RÁS, Z. W.; SKOWRON, A. (ed.) **Foundations of inteligente systems**: International Symposium on Methodologies for Intelligent Systems 1999. Berlin: Springer, 1999. p. 16-29. (Lecture Notes in Computer Science, v. 1609) DOI https://doi.org/10.1007/BFb0095087. Disponível em: https://doi.org/10.1007/BFb0095087. Acesso em: 24 abr. 2025.

LESK, M. **Practical digital libraries:** books, bytes, and bucks. California: Morgan Kaufmann, 1997.

LEY, M. The DBLP computer science bibliography: evolution, research issues, perspectives. *In*: LAENDER, A.H.F.; OLIVEIRA, A.L. **String processing and information retrieval**: International symposium on string processing and information retrieval 2002. Berlin: Springer, 2002. p. 1-10. (Lecture Notes in Computer Science, v. 2476). DOI https://doi.org/10.1007/3-540-45735-6_1. Disponível em: https://doi.org/10.1007/3-540-45735-6_1. Acesso em: 24 abr. 2025.

NORVIG, P.; RUSSELL, S. **Inteligência artificial:** uma abordagem moderna. 4. ed. Rio de Janeiro: LTC, 2022.

PETERSEN, K.; FELDT, R.; MUJTABA, S.; MATTSSON, M.. Systematic mapping studies in software engineering. *In:* INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING (EASE), 12., Italy, 2008. [**Conference Proceedings**]. Italy: BCS Learning & Development, 2008. p. 68-77. DOI: 10.14236/ewic/EASE2008.8. Disponível em: https://www.scienceopen.com/hosted-document?doi=10.14236/ewic/EASE2008.8. Acesso em: 24 abr. 2025.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L.. Guidelines for conducting systematic mapping studies in software engineering: an update. **Information and software technology**,v. 64, p. 1-18, 2015. DOI https://doi.org/10.1016/j.infsof.2015.03.007. Disponível em: https://doi.org/10.1016/j.infsof.2015.03.007. Acesso em: 24 abr. 2025.

PORTER, M. F. An algorithm for suffix stripping. **Program**: electronic library and information systems, v. 14, n. 3, p. 130-137, 1980. DOI https://doi.org/10.1108/eb046814. Disponível em: https://doi.org/10.1108/eb046814. Acesso em: 24 abr. 2025.

SAFFADY, W. Digital library concepts and technologies for the management of library collections: an analysis of methods and costs. **Library Technology Reports,** v. 31, n. 3, p. 221-380, 1995. Disponível em: http://link.gale.com/apps/doc/A17443511/AONE?u=anon~db7d6b9b&sid=googleScholar&xid=679c185f. Acesso em: 24 abr. 2025.

SCHOLAR GOOGLE. c2025. Disponível em: http://scholar.google.com. Acesso em: 23 abr. 2025.

SNYDER, H. Literature review as a research methodology: an overview and guidelines. **Journal of Business Research**, v. 104, p. 333-339, nov. 2019. DOI https://doi.org/10.1016/j.jbusres.2019.07.039. Disponível em: https://doi.org/10.1016/j.jbusres.2019.07.039. Acesso em: 24 abr. 2025.

van AKEN, J. E.; ROMME, G. Reinventing the future: adding design science to the repertoire of organization and management studies. **Organization Management Journal**, v. 6, n. 1, p.

5-12, 2009. DOI https://doi.org/10.1057/omj.2009.1. Disponível em: https://doi.org/10.1057/omj.2009.1. Acesso em: 24 abr. 2025.

WITTEN, I. H.; MOFFAT, A.; BELL, T. C. **Managing gigabytes**: compressing and indexing documents and images. 2. ed. Boston: Morgan Kaufmann, 1999, 484 p.

WU, L. *et al*. Development of benchmark datasets for text mining and sentiment analysis to accelerate regulatory literature review. **Regulatory Toxicology and Pharmacology,** v. 137, p. 105-287, 2023. DOI https://doi.org/10.1016/j.yrtph.2022.105287. Disponível em: https://doi.org/10.1016/j.yrtph.2022.105287. Acesso em: 24 abr. 2025.

Artigo submetido em: 08 nov. 2024
Artigo aceito em: 11 abr. 2025