UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Método KNN Aplicado a Análise Bibliométrica

Wanderson Aparecido da Silva Alves

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Wanderson Aparecido da Silva Alves

Método KNN Aplicado a Análise Bibliométrica

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Leandro Franco de Souza

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

L864a	Alves, Wanderson Aparecido da Silva Método KNN Aplicado a Análise Bibliométrica / Wanderson Aparecido da Silva Alves ; orientador Leandro Franco de Souza. – São Carlos, 2024. 48 p. : il. (algumas color.) ; 30 cm. Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024. 1. Inteligência Artificial. 2. Séries Temporais. 3. Modelos de Regressão. 4. Redes Neurais. I. de Souza, Leandro Franco, orientador. II. Título.

Wanderson Aparecido da Silva Alves

KNN Method Applied to Bibliometric Analysis

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Leandro Franco de Souza

Original version

São Carlos

2024

Este trabalho é dedicado a minha esposa Carla Pompilio, aos meus filhos Dr. Wanderson Pompílio e Caroline Pompílio, a minha nora Amanda Gehrke e a minha netinha: a pequena Laura. Ademais, dedico a todos os amantes de tecnologia e estudantes de Inteligência Artificial.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, pela Graça e pela sabedoria concedidas para enfrentar cada etapa desta jornada.

Aos meus familiares, que sempre acreditaram em mim e me apoiaram nos momentos mais desafiadores, oferecendo amor, paciência e compreensão durante minha dedicação aos estudos.

Aos professores deste MBA, que com sua dedicação, compartilharam conhecimento e foram fundamentais para minha evolução profissional e pessoal.

Ao professor e orientador Dr. Leandro Souza pelo brilhantismo e todo cuidado e zelo na construção desse trabalho.

E, por fim, a todos os amigos e profissionais que, direta ou indiretamente, contribuíram para a realização deste trabalho. A cada um de vocês, minha eterna gratidão.

RESUMO

Alves, W. A. S. **resumo título**. 2024. 48p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

O trabalho propõe avaliar a aplicabilidade de utilização de Inteligência Artificial (IA) no processo de pesquisas científicas, utilizando análise bibliométrica de temas acadêmicos aplicando tipos de mapeamento de coocorrência de palavras-chave, cocitação e acoplamento bibliográfico. Serão apresentadas as técnicas de pré-processamento e análise dos corpos textuais com aplicação de IA como alternativa aos softwares utilizados na pesquisa publicada no artigo Transformação digital na esfera pública: uma análise bibliométrica por (ALVES *et al.*, 2024). Tradicionalmente, softwares como Gephi, VOSviewer e IraMuTeq são utilizados para análise de redes de cocitação, coocorrência de palavras-chave e acoplamento bibliográfico, mas apresentam limitações em flexibilidade analítica e complexidade. Este estudo propõe a utilização de algoritmos para automação no pré-processamento dos dados, além do uso do algoritmo de aprendizado de máquina K-Nearest Neighbors (KNN) como alternativa para simplificar e ampliar o potencial analítico na análise bibliométrica, oferecendo escalabilidade e maior precisão ao processar grandes volumes de dados. A metodologia abrange conceitos de bibliometria e utiliza dados coletados na base Scopus sobre “transformação digital” e “governança” para realizar análises. São aplicadas técnicas de preparação de dados, como tokenização, vetorização e normalização, para aprimorar a eficiência do KNN na identificação de padrões e agrupamentos. O KNN, com sua simplicidade e eficácia, permite a classificação e agrupamento de publicações e autores com base em similaridades, eliminando a dependência de visualizações complexas. Por meio da definição de métricas de similaridade e ajuste do valor de "K", o estudo busca identificar clusters de temas e autores, facilitando a compreensão das tendências e lacunas em pesquisas sobre digitalização em governos e organizações.

Palavras-chave: Transformação Digital; Análise Bibliométrica; Inteligência Artificial (IA); K-Nearest Neighbors (KNN); Mapeamento Científico.

ABSTRACT

Alves, W. A. S. **KNN Method Applied to Bibliometric Analysis**. 2024. 48p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

This study aims to assess the feasibility of employing Artificial Intelligence (AI) in scientific research by conducting a bibliometric analysis of academic topics. The analysis will utilize techniques such as keyword co-occurrence mapping, co-citation analysis, and bibliographic coupling. We will introduce AI-powered methods for pre-processing and analyzing textual corpora as an alternative to tools used in the study "Digital Transformation in the Public Sphere: a bibliometric analysis" por (ALVES *et al.*, 2024). Traditionally, software like Gephi, VOSviewer, and IraMuTeq have been used for network analysis, but they have limitations in flexibility and complexity. This research proposes to automate data pre-processing using algorithms and to leverage the K-Nearest Neighbors (KNN) machine learning algorithm for simplifying and enhancing bibliometric analysis. KNN offers scalability and precision when handling large datasets. Our methodology incorporates bibliometric concepts and employs data from the Scopus database on "digital transformation" and "governance." Data preparation techniques such as tokenization, vectorization, and normalization will be applied to optimize KNN's performance in pattern recognition and clustering. KNN's simplicity and effectiveness allow for classifying and grouping publications and authors based on similarities, eliminating the need for complex visualizations. By defining similarity metrics and adjusting the "K" value, we aim to identify clusters of topics and authors, thereby facilitating the understanding of trends and knowledge gaps in government and organizational digitization research.

Keywords: Digital Transformation; Bibliometric Analysis; Artificial Inteligence; K-Nearest Neighbors (KNN); Scientific Mapping.

LISTA DE FIGURAS

Figura 2.1 – Evolução do tema de pesquisa 27
Figura 2.2 – Rede de coocorrência de palavras-chave dos autores 28
Figura 2.3 – Visualização da rede de coocorrência de palavras-chave dos autores após o refinamento da pesquisa. 30
Figura 2.4 – Visualização da rede de cocitação de referências citadas. 32
Figura 2.5 – Visualização da rede de acoplamento bibliográfico de documentos. . . . 33
Figura 2.6 – Dendrograma com as classes textuais identificadas. 34
Figura 3.1 – Artigos publicados por ano. 36
Figura 3.2 – Matriz de coocorrência como um grafo. 37
Figura 3.3 – Nuvem de Palavras no corpus textual. 38
Figura 3.4 – Palavras chaves mais frequentes. 38
Figura 3.5 – Tabela com palavras chaves mais frequentes. 39
Figura 3.6 – Bigramas mais frequentes. 40
Figura 3.7 – Trigramas mais frequentes. 41
Figura 3.8 – Identificação de Clusters. 42
Figura 3.9 – Palavras chaves com maiores TFIDF. 43
Figura 3.10–Rede KNN. 43

LISTA DE TABELAS

Tabela 1 – Principais arestas da rede de coocorrência de palavras-chave dos autores. 28
Tabela 2 – Resultados da pesquisa bibliográfica por peso das arestas. 29
Tabela 3 – Palavras-chave com as maiores centralidades de autovetor. 31
Tabela 4 – Artigos com a maior centralidade de autovetor da rede de cocitação de referências citadas. 31
Tabela 5 – Artigos com a maior centralidade de autovetor da rede de acoplamento bibliográfico de documentos. 33

LISTA DE ABREVIATURAS E SIGLAS

STLF	*Short Term Load Forecaster* ou Previsor de carga de curto-prazo
SVM	*Support Vector Machine*
SVM	*Support Vector Regression*
XGBoost	*eXtreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO 23

2 REVISÃO BIBLIOGRÁFICA 25
2.1 Bibliometria 25
2.2 Fonte de Dados 26
2.3 Refinamento da Pesquisa 27
2.4 Análise e discussão dos resultados de pesquisas anteriores 29
2.4.1 Rede de coocorrência de palavras-chave 29
2.4.2 Rede de cocitação de referências citadas 30
2.4.3 Rede de acoplamento bibliográfico de documentos 32
2.4.4 Análise dos dados textuais 32

3 METODOLOGIA E RESULTADOS 35
3.1 Preparo dos Dados com IA 35
3.1.1 Pré-processamento dos dados textuais 35
3.2 Aplicação do KNN 39
3.2.1 Avaliação dos dados textuais 40
3.3 Identificação de Padrões 44
3.4 Interpretação das Relações 44
3.5 Comparação com Análises Anteriores 44

4 CONCLUSÕES 45

Referências 47

1 INTRODUÇÃO

A transformação digital já é realidade como uma das principais ferramentas para aumento da eficiência nas organizações, inclusive nos governos de modo geral. A governança eletrônica surge como um pilar para apoiar as organizações e governos no direcionamento de caminhos e nas tomadas de decisões.

Nesse cenário, a análise bibliométrica desempenha um papel importante ao fornecer insights sobre o desenvolvimento e a evolução das pesquisas na área de transformação digital e inteligência artificial (IA) de modo geral. Tradicionalmente, ferramentas como Gephi, VOSviewer e IraMuTeq são amplamente utilizadas para realizar essas análises por meio da visualização de redes de coautoria, cocitação e de coocorrência de palavras-chave. Contudo, essas ferramentas podem apresentar desafios, como a complexidade na configuração e a limitação em termos de flexibilidade analítica.

Com o avanço da IA, surge uma oportunidade para substituir essas ferramentas tradicionais por métodos de aprendizado de máquina, como o algoritmo K-Nearest Neighbors (KNN). O KNN, que é conhecido por sua simplicidade e eficácia na classificação de dados com base em similaridade, permite uma abordagem inovadora para a análise bibliométrica, identificando padrões de agrupamento e relação sem depender exclusivamente de visualizações complexas. Além disso, a IA oferece a vantagem de escalabilidade, permitindo que grandes volumes de dados sejam processados e analisados com maior rapidez e precisão.

Dessa forma, o presente estudo busca realizar a preparação dos corpos textuais e pré-processamento de forma automatizada, além de aplicar o KNN como ferramenta principal para a análise bibliométrica, explorando suas capacidades de identificar padrões, temas emergentes e relações entre os principais tópicos de um determinado assunto. Ao substituir as ferramentas tradicionais pela IA, esperamos simplificar o processo analítico e ampliar as possibilidades de interpretação dos dados, contribuindo para um entendimento mais profundo dos avanços e desafios na digitalização de organizações governamentais e institucionais.

O trabalho está dividido da seguinte forma: o Capítulo 2 apresenta a revisão bibliográfica e resultados de estudos anteriores; a metodologia e os resultados obtidos através da técnica KNN são apresentados no Capítulo 3 e no Capítulo 4 são apresentadas as principais conclusões que se pode obter através deste estudo.

2 REVISÃO BIBLIOGRÁFICA

O presente capítulo que trata da revisão bibliográfica e está subdividido da seguinte forma: os conceitos de bibliometria; a coleta e a análise dos dados; o refinamento da pesquisa; e a análise e discussão de resultados de pesquisas anteriores.

2.1 Bibliometria

De acordo com Pritchard (1969), a bibliometria é a aplicação de métodos matemáticos e estatísticos aos livros e demais comunicação escrita. A bibliometria analisa estatisticamente números de publicações e citações, assim como as relações entre publicações para sistematizar um campo de pesquisa (ALVES *et al.*, 2024; ELLEGAARD; WALLIN, 2015; KÜCHER; FELDBAUER-DURSTMÜLLER, 2019; ZUPIC; ČATER, 2015).

O mapeamento científico visualiza estatisticamente as conexões entre publicações, ajudando na interpretação do conteúdo dos estudos. Esse trabalho irá buscar representar os três tipos de mapeamentos aplicados na pesquisa publicada no artigo Transformação digital na esfera pública: uma análise bibliométrica (ALVES *et al.*, 2024), que são: coocorrência de palavras-chave, cocitação e acoplamento bibliográfico.

A Coocorrência de Palavras-Chave se trata de uma técnica que analisa as palavras presentes nos documentos para construir uma estrutura conceitual do campo estudado. Normalmente, as palavras que aparecem juntas em diversos documentos indicam uma relação conceitual entre os temas, formando um "mapa semântico"que reflete a estrutura cognitiva da área. Esse método pode ser aplicado em títulos, resumos, palavras-chave ou texto completo, analisando termos individuais ou compostos.

A Cocitação mede a frequência com que duas referências são citadas simultaneamente em outros trabalhos, sugerindo uma conexão conceitual entre elas. A cocitação identifica áreas de pesquisa densamente conectadas e ajuda a destacar frentes de pesquisa, pois as referências frequentemente citadas juntas são vistas como influentes. Esta técnica, contudo, reflete o estado do campo no momento da publicação e permite o monitoramento de mudanças nos paradigmas científicos.

Já o Acoplamento Bibliográfico é um método que mede a similaridade entre documentos pelo número de referências que compartilham. Documentos com muitas referências em comum tendem a abordar temas semelhantes. Diferente da cocitação, que pode variar ao longo do tempo, o acoplamento bibliográfico permanece estático, uma vez que as referências em um artigo são fixas. É mais eficaz em publicações contemporâneas e oferece insights sobre a conexão temática entre trabalhos próximos em termos temporais.

Portanto esses métodos se complementam ao identificar influências, conexões temáticas e agrupamentos de interesse em áreas de pesquisa, facilitando uma compreensão mais aprofundada do desenvolvimento e das frentes de estudo dentro de um domínio.

Existem inúmeros desafios e oportunidades quando precisamos trabalhar com grandes fontes de dados. As limitações das ferramentas tradicionais possibilitam novas oportunidades como utilização de algoritmos e outras abordagens de IA podem oferecer para simplificar e potencializar a análise bibliométrica.

2.2 Fonte de Dados

Os dados utilizados nessa pesquisa foram os mesmos utilizados por Alves *et al.* (2024), pois o foco dessa pesquisa é avaliar a aplicabilidade de utilização de IA no processo de pesquisas científicas, fazendo um paralelo com as técnicas e softwares utilizados no artigo Transformação digital na esfera pública: uma análise bibliométrica (ALVES *et al.*, 2024).

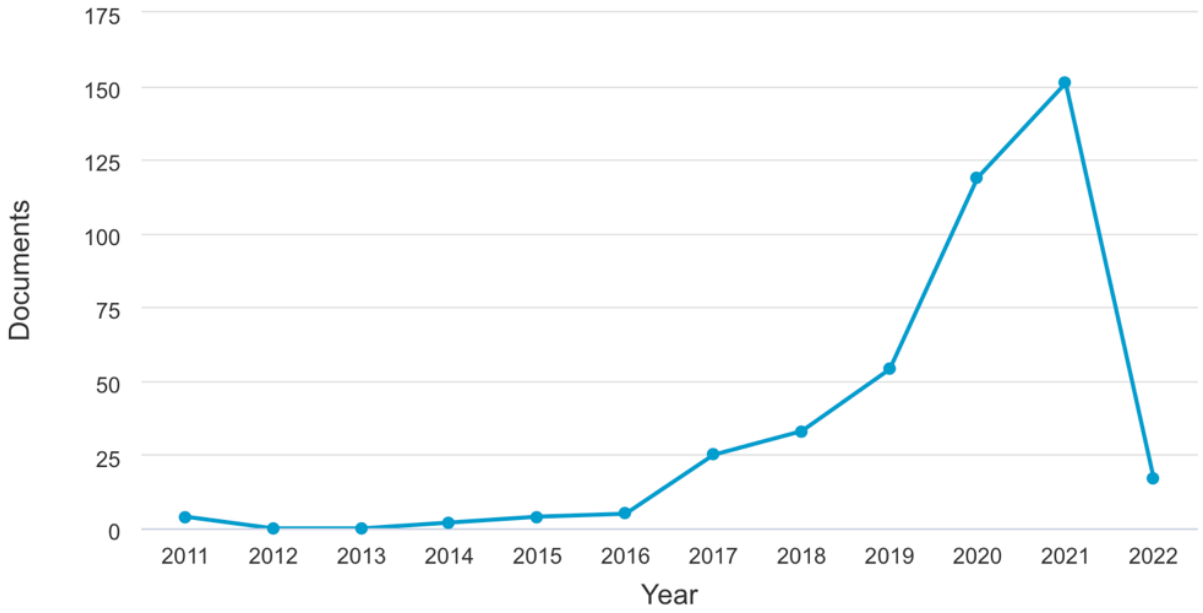
Alves *et al.* (2024) adotam uma abordagem bibliométrica seguindo para o mapeamento científico usando técnicas de análise de redes sobre metadados recuperados da pesquisa bibliográfica. Os métodos de análise de rede são aplicados ao estudo das relações entre um conjunto de atores (BORGATTI, 2002). Para a análise de redes foram utilizados os métodos publicados por Newman (2010), Eck e Waltman (2010), Eck e Waltman (2014).

Inicialmente foi realizada uma consulta na base de dados Scopus com a seguinte expressão de busca: “digital transformation” AND (“governance” OR “it governance” OR “data governance”). O resultado recuperou 420 referências, no período compreendido entre 2011 e 2022, delimitado aos tipos de documentos publicados em periódicos ou em conferências. A metodologia utilizada por Alves *et al.* (2024), seguiu as abordagens sugeridas por Moresi, Pinho e Costa (2022) exportando os metadados da consulta na base Scopus e utilizando as ferramentas Gephi e VOSviewer para análises, controle do vocabulário e normalização das referências bibliográficas; obtenção das redes de coocorrência de palavras-chave dos autores, do software Gephi (BASTIAN; HEYMANN; JACOMY, 2009) para o cálculo das métricas de análise de redes e refinamento da pesquisa a partir da análise das arestas da rede de coocorrência.

Por fim, Alves *et al.* (2024) realizam a análise dos dados textuais utilizando o software Iramuteq a partir dos artigos mais relevantes da pesquisa bibliográfica. A análise textual é uma metodologia flexível e pode auxiliar necessidades específicas de pesquisa, utilizando técnicas e abordagens para analisar textos (WHITE *et al.*, 2006).

Conforme apresentado na Figura 2.1, Alves *et al.* (2024) demonstram a evolução do tema pesquisado com o pico de documentos publicados e o período de tais publicações.

Figura 2.1 – Evolução do tema de pesquisa



Fonte: Pesquisa na base Scopus apresentada por Alves *et al.* (2024).

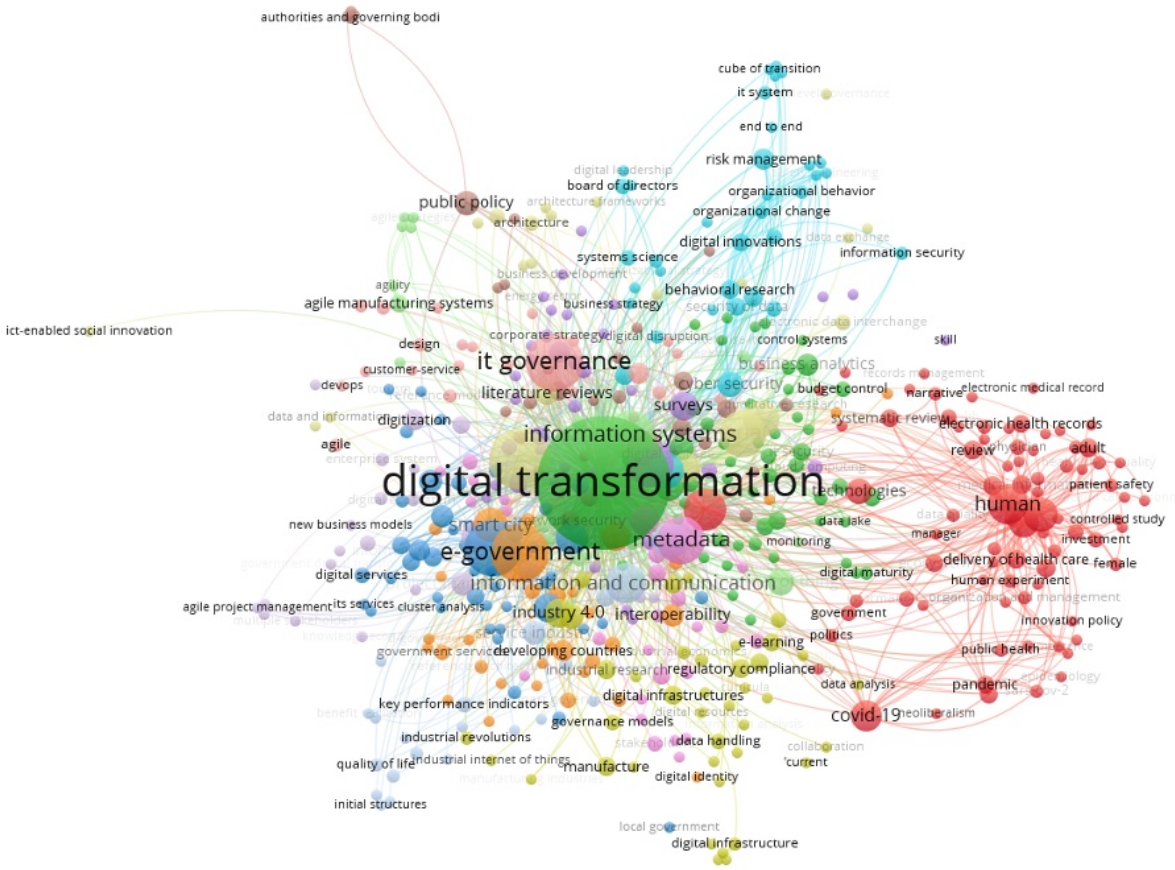
2.3 Refinamento da Pesquisa

Para o refinamento da pesquisa, Alves *et al.* (2024) adotaram o referencial de Moresi *et al.* (2021). Primeiramente, os metadados dos documentos obtidos na pesquisa bibliográfica foram importados para o VOSviewer (ECK; WALTMAN, 2022). A partir disso, foi gerada uma rede de coocorrência de palavras-chave dos autores, considerando um mínimo de duas ocorrências para cada palavra, resultando em um grafo com 601 nós, 18 comunidades e 8.430 arestas. O VOSviewer possibilita o controle do vocabulário por meio de um arquivo TXT, denominado tesauro que precisou ser trabalhado manualmente. Após incluir o tesauro, a rede de coocorrência de palavras-chave dos autores, com ao menos duas ocorrências, apresentou um grafo com 493 nós, 14 comunidades e 6.344 arestas, conforme ilustrado na Figura 2.2.

Alves *et al.* (2024) salvam a rede no formato GML e importam para o Gephi. Após o cálculo das métricas da rede de coocorrência das palavras-chave dos autores, selecionam o Laboratório de Dados e a opção de arestas. A Tabela 1 exibe as 14 arestas com os maiores pesos. Como se trata de uma rede não direcionada, os nós de origem e destino não possuem orientação específica, representando unicamente as coocorrências entre as respectivas palavras-chave.

Na sequência, Alves *et al.* (2024) elaboraram uma nova expressão de busca, combinando as palavras-chave de cada aresta utilizando o operador lógico AND. Esse processo foi realizado de forma iterativa, em que a nova expressão era consultada na base bibliográfica, e a quantidade de documentos recuperados era registrada. Cada par de palavras-chave

Figura 2.2 – Rede de coocorrência de palavras-chave dos autores



Fonte: imagem Gerada por Alves *et al.* (2024) utilizando o software VOSviewer.

Tabela 1 – Principais arestas da rede de coocorrência de palavras-chave dos autores.

Origem	**Destino**	**Peso**
350 - digital transformation	667 - it governance	28.0
350 - digital transformation	395 - e-governance	24.0
350 - digital transformation	398 - e-government	21.0
350 - digital transformation	904 - public governance	16.0
237 - data analytics	350 - digital transformation	13.0
398 - e-government	904 - public governance	10.0
213 - covid-19	350 - digital transformation	8.0
395 - e-governance	398 - e-government	8.0
395 - e-governance	1006 - smart city	8.0

Fonte: gerado por Alves *et al.* (2024) utilizando o software Gephi.

adicional foi integrado à busca por meio do operador lógico OR, conforme detalhado na Tabela 2, sendo consultado novamente na base de dados. O processo foi concluído ao atingir a saturação, com o total de 2.015 documentos recuperados.

Após essa etapa, Alves *et al.* (2024) demonstraram os resultados da pesquisa apresentando as redes de coocorrência de palavras-chave dos autores, de cocitação de

Tabela 2 – Resultados da pesquisa bibliográfica por peso das arestas.

Índice	**Expressão de busca**	**Quantidade de documentos**
#1	(“digital transformation"AND "it governance”)	46
#2	#1 OR (“digital transformation"AND "e-governance”)	75
#3	#2 OR (“digital transformation"AND "e- government”)	247
#4	#3 OR (“digital transformation"AND "public governance”)	252
#5	#4 OR (“data analytics"AND "digital transformation”)	538
#6	#5 OR (“e-government"AND "public governance”)	582
#7	#6 OR (“covid-19"AND "digital transformation”)	1223
#8	#7 OR (“e-governance"AND "e-government”)	2167
#9	#8 OR (“e-governance"AND "smart city”)	2245

Fonte: gerado por Alves *et al.* (2024) realizando nova pesquisa na base Scopus.

referências citadas e de acoplamento bibliográfico de documentos.

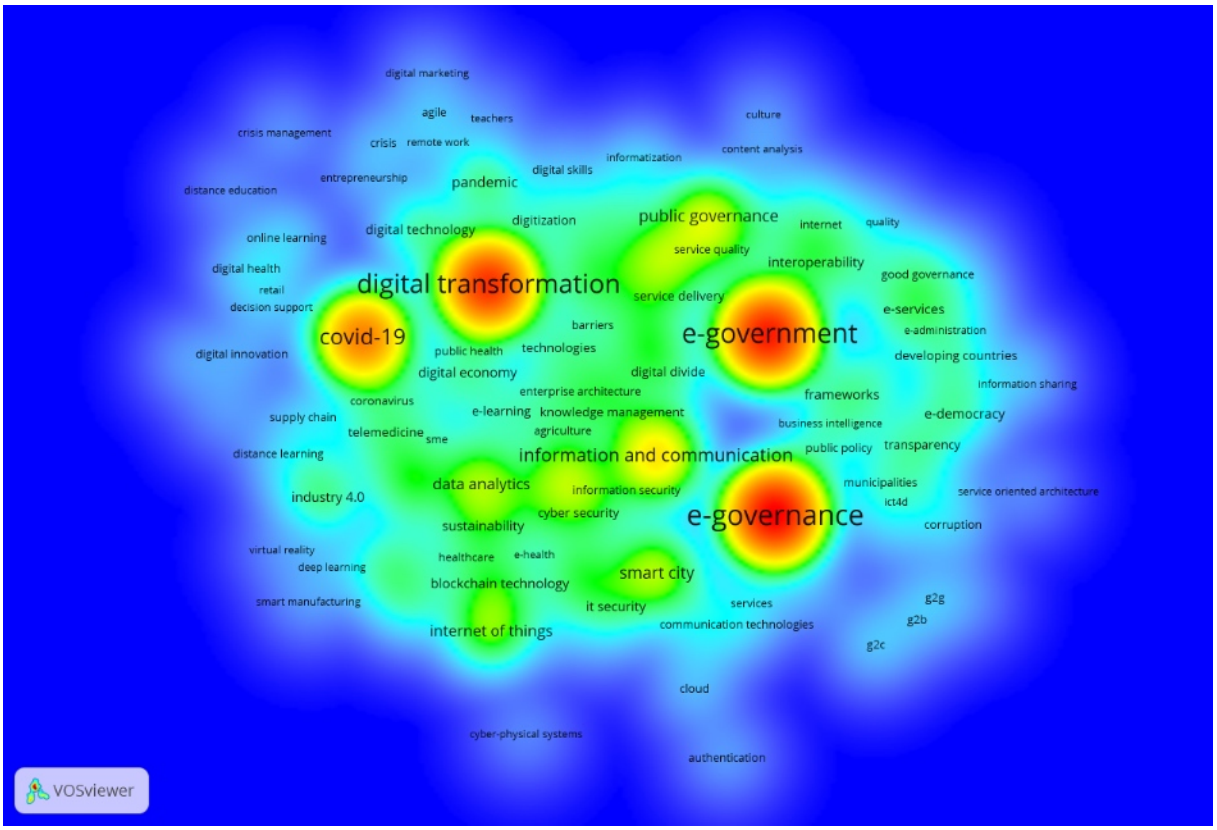
2.4 Análise e discussão dos resultados de pesquisas anteriores

2.4.1 Rede de coocorrência de palavras-chave

Alves *et al.* (2024) processaram no VOSviewer os metadados da pesquisa bibliográfica refinada (ECK; WALTMAN, 2014), utilizando a opção de análise de coocorrência das palavras-chave dos autores presentes em cada publicação. Inicialmente, sem o controle do vocabulário e considerando um mínimo de cinco ocorrências para cada par de palavras-chave, a rede gerada contou com 234 nós, 13 comunidades e 2.492 arestas. Em seguida, foi criado um tesauro para o controle do vocabulário, resultando em uma nova rede com 173 nós, 12 comunidades e 1.578 arestas. A Figura 2.3 apresenta um mapa de densidade da rede de coocorrência de palavras-chave, destacando termos como digital transformation, e-governance, e-government, covid-19, it governance, public governance, smart city e big data. Alves *et al.* (2024) afirmam que esses resultados refletem a formulação da expressão de busca, além de observarem o surgimento de temas periféricos que podem representar oportunidades de pesquisa, como digital twins, decision support, e-democracy, telemedicine, enterprise architecture, e-business e remote work. Alves *et al.* (2024) citam ainda que a análise visual da rede não permite identificar com precisão as palavras-chave emergentes, sendo necessário o cálculo de métricas de análise de redes para aprofundar essa identificação.

Alves *et al.* (2024) aprofundam a análise dos dados calculando as métricas de análise de redes utilizando o software Gephi (BASTIAN; HEYMANN; JACOMY, 2009). Citam que entre as métricas avaliadas, destaca-se o grau médio, que representa o número médio de conexões por nó na rede (NEWMAN, 2010); a modularidade, que mede a intensidade da divisão da rede em comunidades (NEWMAN *et al.*, 2009; BLONDEL *et al.*, 2008); e a

Figura 2.3 – Visualização da rede de coocorrência de palavras-chave dos autores após o refinamento da pesquisa.



Fonte: Imagem Gerada por Alves *et al.* (2024) utilizando o software VOSviewer.

centralidade de autovetor, que calcula a influência de cada nó com base nos autovalores únicos da matriz de adjacências (NEWMAN *et al.*, 2009; RUHNAU, 2000). O Gephi disponibiliza um recurso de laboratório de dados, permitindo a extração detalhada dessas métricas. A Tabela 3 apresenta as palavras-chave com centralidade de autovetor superior a 0,4000, indicando as mais influentes na rede (ALVES *et al.*, 2024).

2.4.2 Rede de cocitação de referências citadas

Alves *et al.* (2024) criam a rede de cocitação de documentos utilizando o software VOSviewer (ECK; WALTMAN, 2022). Importam os metadados com a aplicação de um arquivo de tesauro para normalizar as referências bibliográficas, eliminando duplicidades e aprimorando a precisão dos resultados. Citam que sem o uso do tesauro, e considerando um mínimo de cinco cocitações por documento, a rede gerada continha 63 nós, 6 comunidades e 365 arestas. Afirmam que com a inclusão do tesauro de referências, a rede ajustada apresentou 61 nós, 6 comunidades e 363 arestas, conforme ilustrado na Figura 2.4.

Exportam o grafo para o Gephi (BASTIAN; HEYMANN; JACOMY, 2009), onde calculam as métricas de análise de redes, como grau médio, modularidade e centralidade de autovetor. Com base nessas métricas ordenam os documentos em ordem decrescente

Tabela 3 – Palavras-chave com as maiores centralidades de autovetor.

Palavra-chave	**Ano Médio**	**Grau**	**Centralidade de autovetor**
e-government	2014,925	134	1,000
e-governance	2015,139	132	0,999
digital transformation	2020,475	121	0,942
covid-19	2020,844	90	0,762
information and communication technology	2015,865	78	0,712
smart city	2018,693	58	0,570
big data	2019,047	53	0,559
public governance	2017,026	56	0,554
artificial intelligence	2020,447	50	0,533
data analytics	2019,875	50	0,523
internet of things	2019,661	47	0,503
innovations	2018,737	41	0,459
cloud computing	2017,514	40	0,440
machine learning	2019,929	40	0,412

Fonte: gerado por Alves *et al.* (2024) utilizando o software Gephi.

Tabela 4 – Artigos com a maior centralidade de autovetor da rede de cocitação de referências citadas.

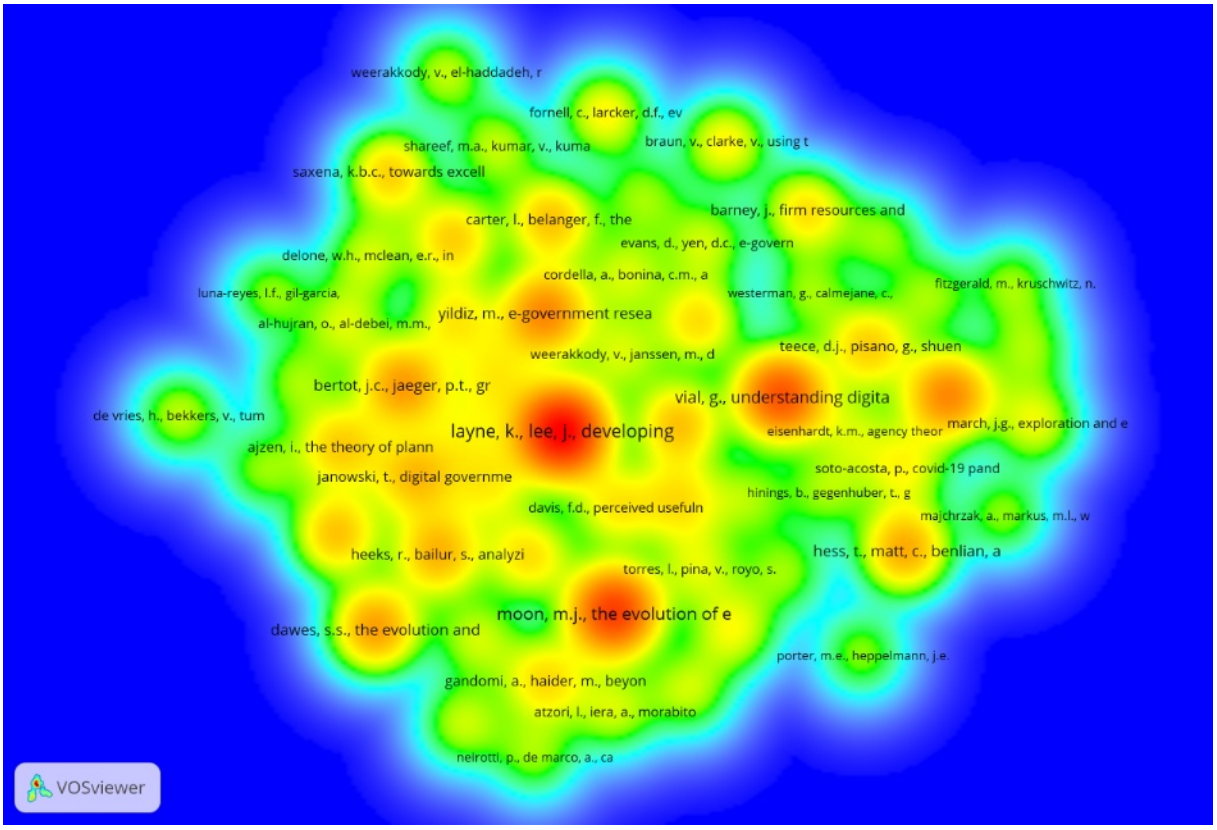
Documento	**Tema**	**Grau**	**Centralidade de autovetor**
Layne; Lee (2001)	e-government	32	1,000
West (2004)	digital government	27	0,926
Yildiz (2007)	e-government	25	0,820
Moon (2002)	e-government	23	0,763
Bertot et al. (2010)	e-government	24	0,741
Andersen; Henriksen (2006)	e-government	20	0,686
Heeks; Bailur (2007)	e-government	19	0,676
Coursey; Norris (2008)	e-government	18	0,669
Ebrahim; Irani (2005)	government data processing	20	0,657
Weerakkody et al. (2011)	e-government	19	0,647

Fonte: gerado por Alves *et al.* (2024) utilizando o software Gephi.

pela centralidade de autovetor, o que permitiu identificar as referências mais influentes. A Tabela 4 apresenta os 10 documentos com as maiores centralidades de autovetor.

Segundo Alves *et al.* (2024), a análise da rede de cocitação indicou que os artigos mais influentes abordavam o tema de governo eletrônico e afirmam que o resultado é consistente com os resultados da rede de coocorrência de palavras-chave. Por fim, citam que a palavra-chave mais influente nessa rede foi e-government, com grau 134 em um grafo de 173 nós.

Figura 2.4 – Visualização da rede de cocitação de referências citadas.



Fonte: Imagem Gerada por Alves *et al.* (2024) utilizando o software VOSviewer.

2.4.3 Rede de acoplamento bibliográfico de documentos

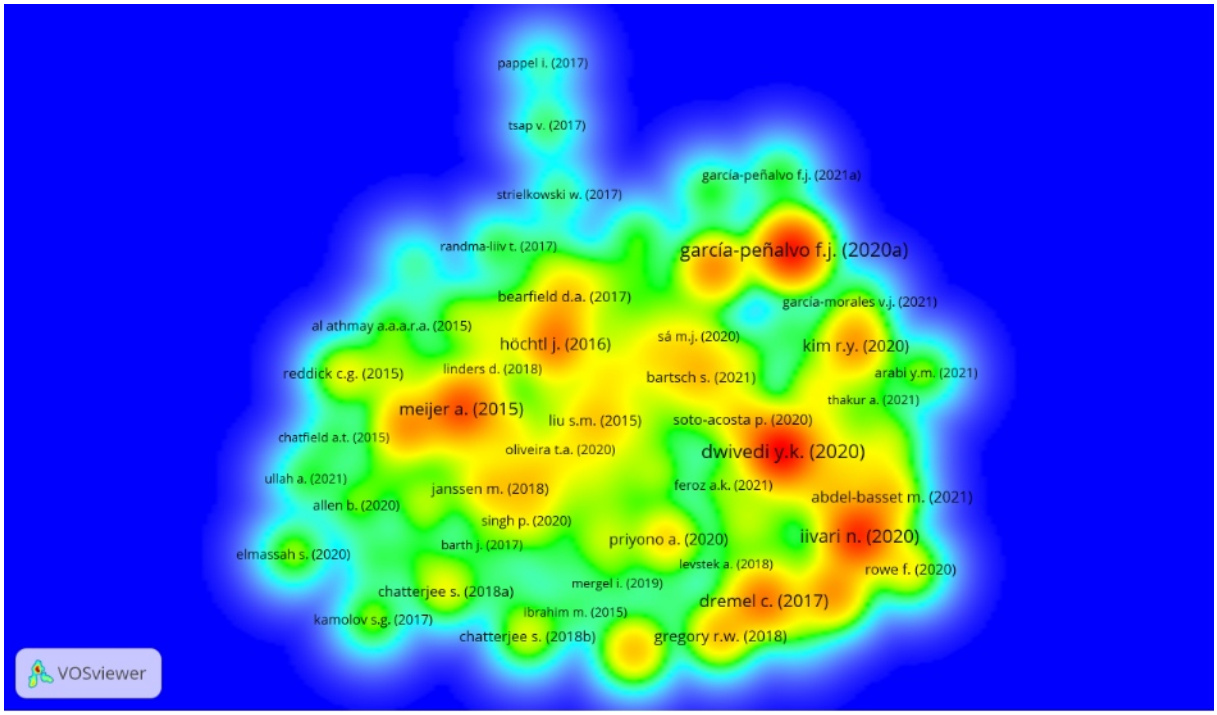
Dando continuidade ao refinamento da pesquisa, Alves *et al.* (2024) recuperam 1.590 documentos na plataforma Scopus, restringindo-se ao período de 2015 a 2022. Criam a rede de acoplamento bibliográfico de documentos com o software VOSviewer (ECK; WALTMAN, 2022). Realizaram a importação dos metadados com a opção de análise de acoplamento bibliográfico, estabelecendo um mínimo de 15 documentos em comum. Afirmam que a rede gerada conta com 92 nós, 13 comunidades e 247 arestas, conforme ilustrado na Figura 2.6.

Após essa etapa, Alves *et al.* (2024) exportam o grafo para o Gephi (BASTIAN; HEYMANN; JACOMY, 2009), e calculam as métricas de análise de redes, incluindo grau médio, modularidade e centralidade de autovetor. A Tabela 5 destaca os 10 documentos com as maiores centralidades de autovetor, indicando aqueles com maior influência na rede (ALVES *et al.*, 2024).

2.4.4 Análise dos dados textuais

Alves *et al.* (2024) construíram um corpus com os 40 artigos mais relevantes identificados para a pesquisa, e realizada uma análise textual utilizando o software Iramuteq.

Figura 2.5 – Visualização da rede de acoplamento bibliográfico de documentos.



Fonte: Imagem Gerada utilizando o software VOSviewer.

Tabela 5 – Artigos com a maior centralidade de autovetor da rede de acoplamento bibliográfico de documentos.

Documento	**Tema**	**Grau**	**Centralidade de autovetor**
Pereira et al. (2018)	smart governance	19	1,000
Janssen; Helbig (2018)	e-government	15	0,821
Chatfield; Reddick (2016)	smart city	15	0,812
Linders et al. (2018)	e-government	14	0,804
Allen et al. (2020)	smart city	11	0,681
Reddick et al. (2015)	e-government	9	0,639
Lee-Geiller; Lee (2019)	e-government	14	0,633
Chatfield; Reddick (2015)	e-government	10	0,630
Nielsen (2016)	e-government	12	0,623
McNutt et al. (2016)	e-government	11	0,584

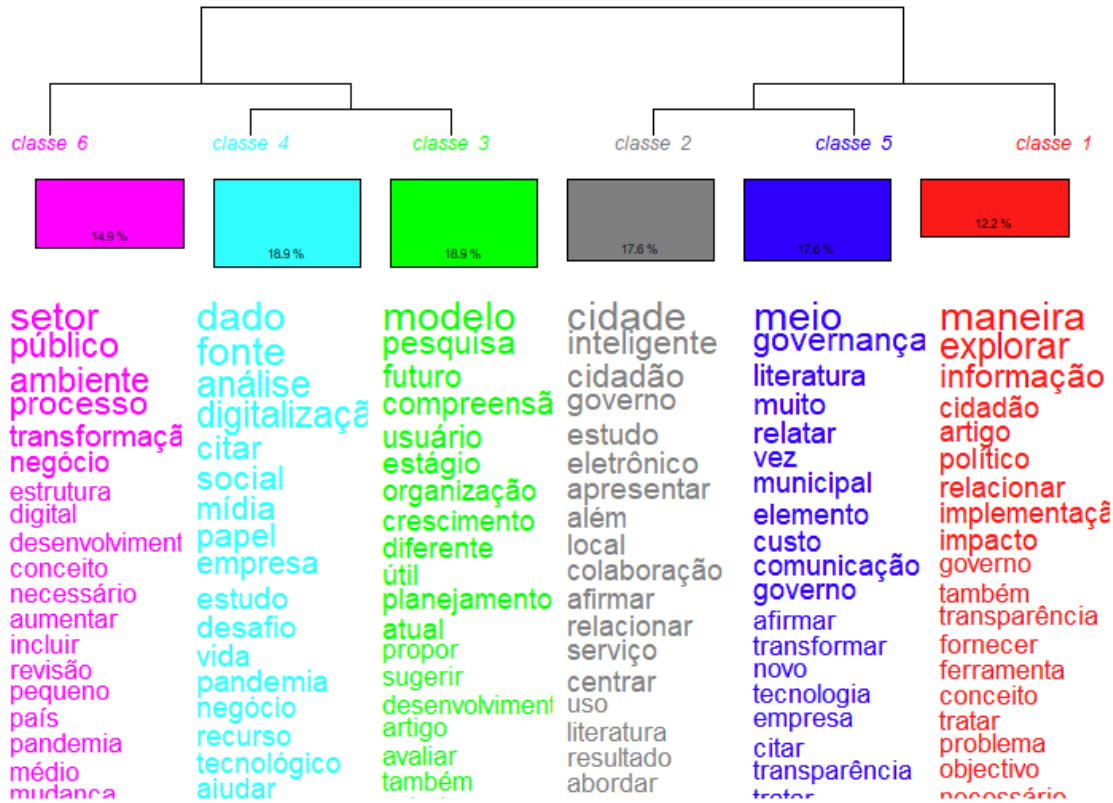
fonte: gerado por Alves *et al.* (2024) utilizando o software Gephi.

Citam que após o carregamento do corpus, aplicou-se uma análise estatística clássica, que revelou os seguintes resultados: 40 textos, 2.984 ocorrências, 741 formas, das quais 442 eram únicas (hápax), correspondendo a 14,81% das ocorrências e 59,65% das formas. A média de ocorrências por texto foi de 75%.

Para aprofundar a análise textual, Alves *et al.* (2024) realizaram análises de especificidades, a Análise Fatorial de Correspondência (AFC) e a classificação pelo método

de Reinert. A Figura 2.6 apresenta o dendrograma das classes textuais identificadas (CAMARGO; JUSTO, 2013). As classes obtidas foram interpretadas como categorias para análise dos dados textuais pelos autores. Nesse processo, os materiais gerados pelo Iramuteq, aliados à leitura e interpretação dos resumos pelo pesquisador, contribuíram para a construção da compreensão dos dados, servindo como base para a formulação de inferências.

Figura 2.6 – Dendrograma com as classes textuais identificadas.



Fonte: Imagem Gerada utilizando o software Iramuteq.

Para a auxiliar na análise de conteúdo, Alves *et al.* (2024) afirmam que o software Iramuteq identificou seis classes a partir do corpus, as quais foram agrupadas em dois sub-corpora distintos. Segundo os autores, as classes 6, 4 e 3 estão mais relacionadas em um grupo, enquanto no segundo grupo as classes 2, 5 e 1 estão relacionadas. A partir desses agrupamentos os autores fazem análises das proximidades das classes com atribuição de nomes e construção de narrativas.

3 METODOLOGIA E RESULTADOS

Esse capítulo apresenta a metodologia e os resultados de análise bibliométrica utilizando o método K-NN. Ele está subdividido em duas seções, o preparo dos dados para usar com Inteligência Artificial e a aplicação do KNN.

3.1 Preparo dos Dados com IA

3.1.1 Pré-processamento dos dados textuais

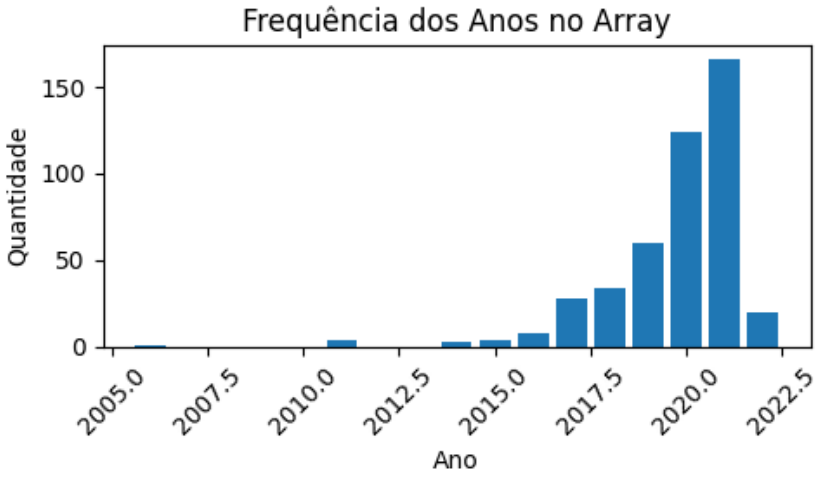
Seguindo a estratégia de avaliação da aplicabilidade de utilização de Inteligência Artificial (IA) no processo de pesquisas científicas, serão apresentadas as técnicas de pré-processamento e análise dos corpos textuais com aplicação de IA como alternativa aos softwares utilizados por Alves *et al.* (2024).

Assim, foi realizado a tokenização e limpeza dos dados, ou seja, houve a separação do texto em unidades menores (tokens) e a remoção de elementos irrelevantes para essa análise, como stopwords e símbolos. Outro aspecto importante dessas atividades de preparação dos dados é a facilitação do processamento dos dados.

Foi realizada a técnica de vetorização denominada Embedding para converter palavras-chave e termos em valores numéricos. A vetorização é essencial para que o KNN possa comparar as publicações e identificar similaridades. Por fim, foi realizada a Normalização e Escalonamento. Como o KNN é sensível a escalas, é recomendável normalizar ou escalonar os dados para garantir que os valores estejam dentro de uma mesma faixa, o que melhora a precisão das classificações e dos agrupamentos.

Os dados foram carregados em um notebook Python desenvolvido para realizar o pré-processamento e automação. Inicialmente foi plotado a ocorrência de arquivos por ano, conforme apresentado na Figura 3.1.

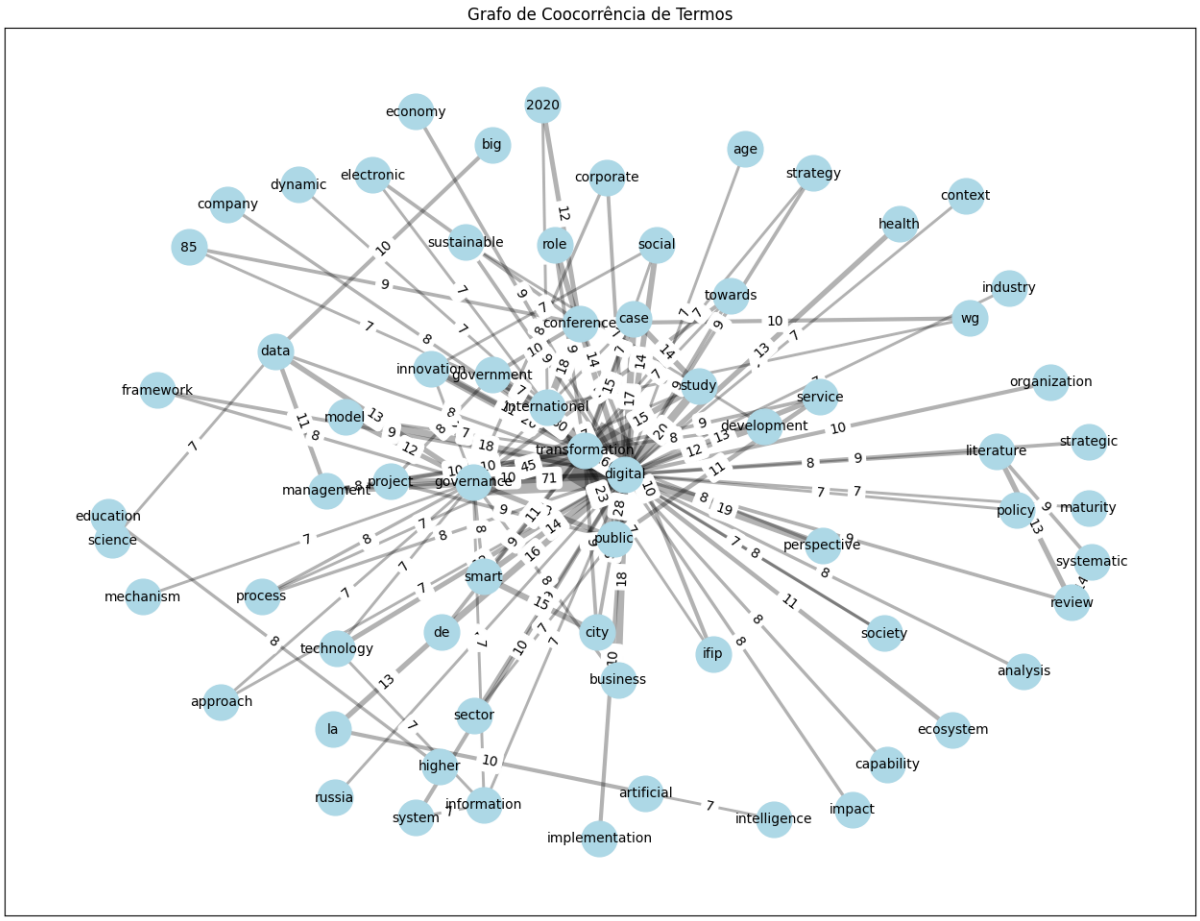
Figura 3.1 – Artigos publicados por ano.



Fonte: Imagem Gerada pelo autor em Python.

Na sequência houve o processamento dos corpus textuais com criação de dataframes para tokenização e criação de estruturas para geração de redes de coocorrência, conforme apresentado na Figura 3.2.

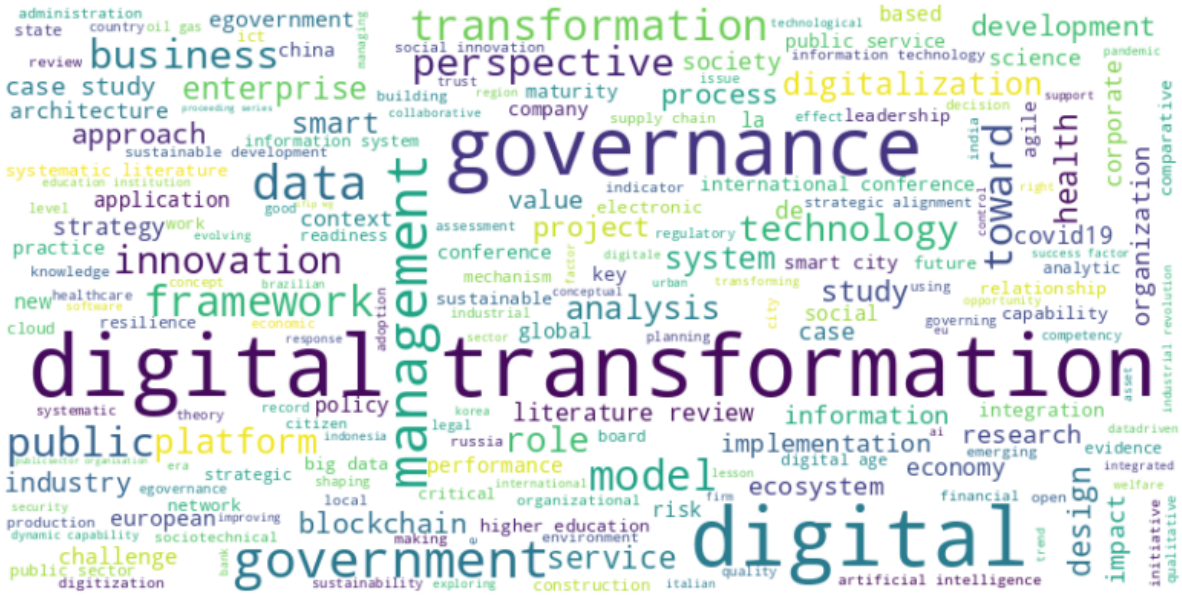
Figura 3.2 – Matriz de coocorrência como um grafo.



Fonte: Imagem Gerada pelo autor em Python.

Prosseguindo com o processamento dos dados e início das análises foi gerado uma nuvem de palavras no corpus textual, conforme Figura 3.3.

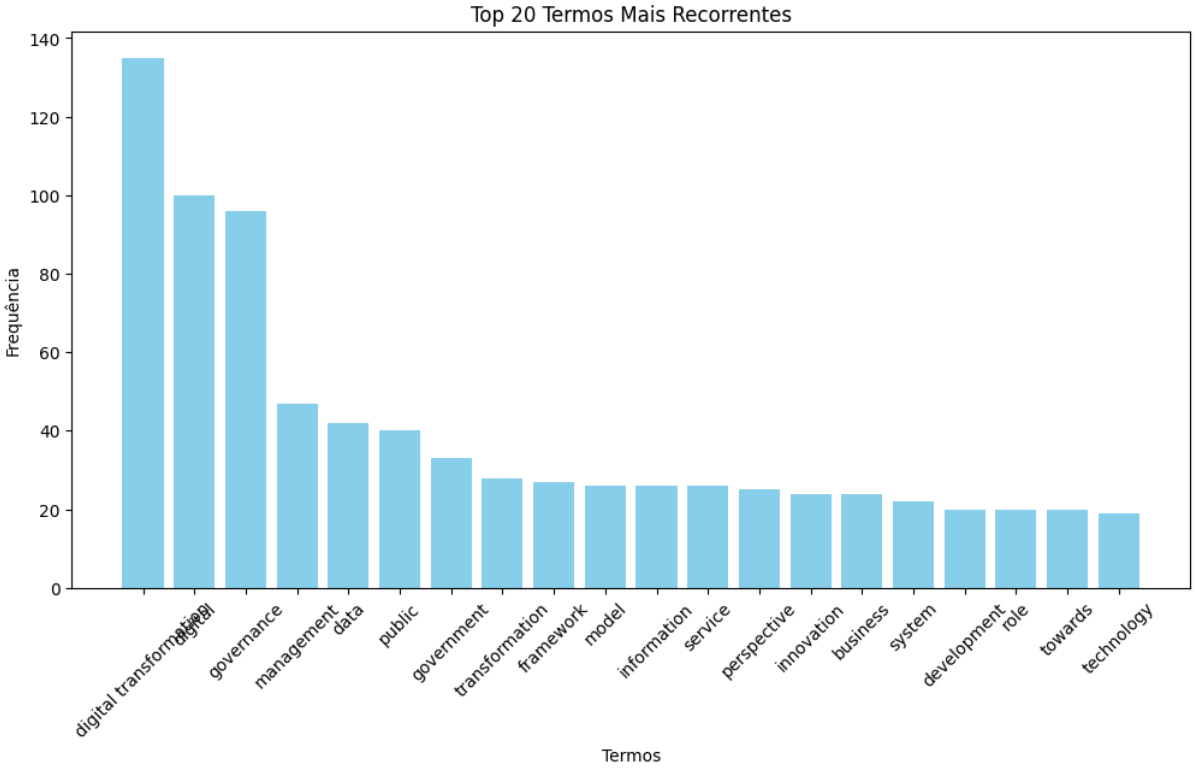
Figura 3.3 – Nuvem de Palavras no corpus textual.



Fonte: Imagem Gerada pelo autor em Python.

Prosseguindo a análise de avaliação dos temas tratados nos corpus textuais, houve a geração do gráfico das palavras chaves mais frequentes, conforme apresentado nas Figuras 3.4 e 3.5.

Figura 3.4 – Palavras chaves mais frequentes.



Fonte: Imagem Gerada pelo autor em Python.

Figura 3.5 – Tabela com palavras chaves mais frequentes.

Out[ ]:

	Termo 1	Termo 2	Coocorrência
0	digital	transformation	166
1	digital	governance	71
2	governance	transformation	45
3	digital	government	30
4	digital	public	28
...	...	...	...
82	governance	process	8
83	business	governance	8
84	company	digital	8
85	government	project	8
86	data	digital	8

87 rows × 3 columns

Fonte: Imagem Gerada pelo autor em Python.

3.2 Aplicação do KNN

O KNN permite que as publicações ou autores sejam agrupados por similaridade, identificando relações e padrões importantes sem depender de visualizações de redes complexas. Esta técnica é particularmente útil por sua simplicidade e precisão em classificar dados baseados em proximidade. Assim, a escolha do algoritmo KNN se mostrou como uma alternativa inovadora aos softwares Gephi, VOSviewer e IraMuTeq. Se fez necessário a definição da Métrica de Similaridade que basicamente está na escolha de uma métrica de distância (como distância euclidiana ou cosseno) que melhor represente a proximidade entre publicações, autores ou palavras-chave. Essa métrica é fundamental para o funcionamento do KNN. Outro fator é a escolha do Valor de K: pois se trata do critério para definição do número de vizinhos mais próximos, “K”, equilibrando entre precisão e robustez nos agrupamentos. Um valor de K muito baixo pode levar a ruído nos dados, enquanto um valor muito alto pode diluir relações importantes. Por fim, buscou-se a formação de clusters para aplicação das técnicas de Classificação e Agrupamento. Assim, é possível identificar grupos de artigos, autores ou temas semelhantes. Esse agrupamento ajuda a mapear os principais focos das pesquisas e possíveis lacunas.

3.2.1 Avaliação dos dados textuais

Dando prosseguimento para preparação dos dados e geração do KNN, foram desenvolvidas análises N-gramas. A Figura 3.6 apresenta os Bigramas mais frequentes, enquanto a Figura 3.7 apresenta os Trigramas.

Figura 3.6 – Bigramas mais frequentes.

[17]

	word	tfidf_sum
33	digit transform	80.013940
25	digit govern	18.649013
42	govern digit	12.735056
87	transform govern	8.516999
9	case studi	7.770124
73	smart citi	6.637325
69	public servic	6.504847
89	transform public	6.470297
53	inform system	6.323796
57	literatur review	6.222763
47	higher educ	5.689268
6	big data	5.273068
54	inform technolog	5.215253
80	sustain develop	5.161227
21	digit age	5.091204

Fonte: Imagem Gerada pelo autor em Python.

Figura 3.7 – Trigramas mais frequentes.

	word	tfidf_sum
6	digit transform govern	10.000000
10	govern digit transform	9.286314
18	systemat literatur review	8.000000
8	digit transform public	6.954110
11	higher educ institut	5.000000
9	digit transform strategi	4.000000
14	manag digit transform	3.400034
7	digit transform project	3.159415
5	digit transform conceptu	3.000000
1	adopt digit transform	3.000000
12	implement digit transform	2.721007

Fonte: Imagem Gerada pelo autor em Python.

Foram construídas clusters para identificando estruturas na rede, conforme apresentado na Figura 3.8.

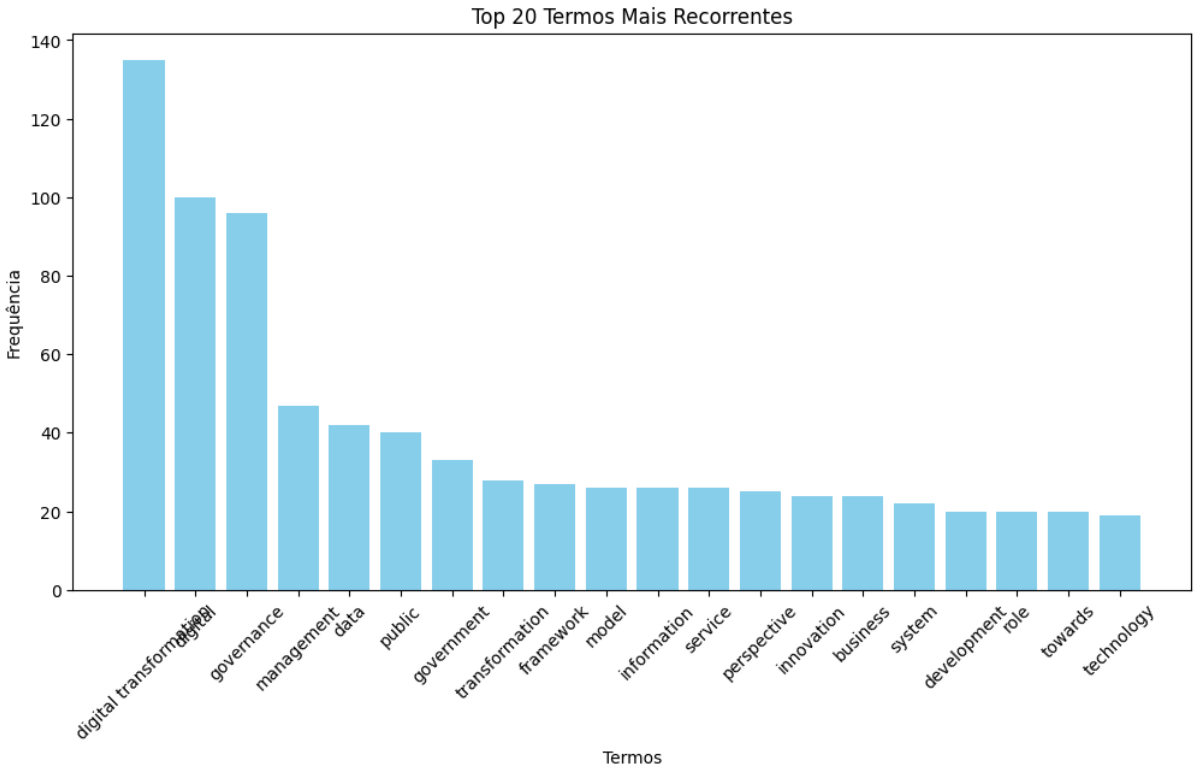
Figura 3.8 – Identificação de Clusters.

	Title	cluster
0	Development of an assessment model for industr...	0
1	Web-enabled supply chain management: Key antec...	1
2	It consumerization and the transformation of i...	2
3	E-government in Canada: Transformation for the...	3
4	Digital transformation in latecomer industries...	4
...	...	...
447	Standing Conference of Eastern, Central, and S...	0
448	Lecture Notes in Informatics (LNI), Proceeding...	48
449	From territories to tourist areas: Ending some...	0
450	PERIKLIS - electronic democracy in the 21st ce...	0
451	Enterprise business technology governance: Val...	13

Fonte: Imagem Gerada pelo autor em Python.

Na sequência, foram selecionadas as palavras chaves com maiores TFIDF (Term Frequency-Inverse Document Frequency) dos clusters para analisar o resultado obtido, conforme Figura 3.9.

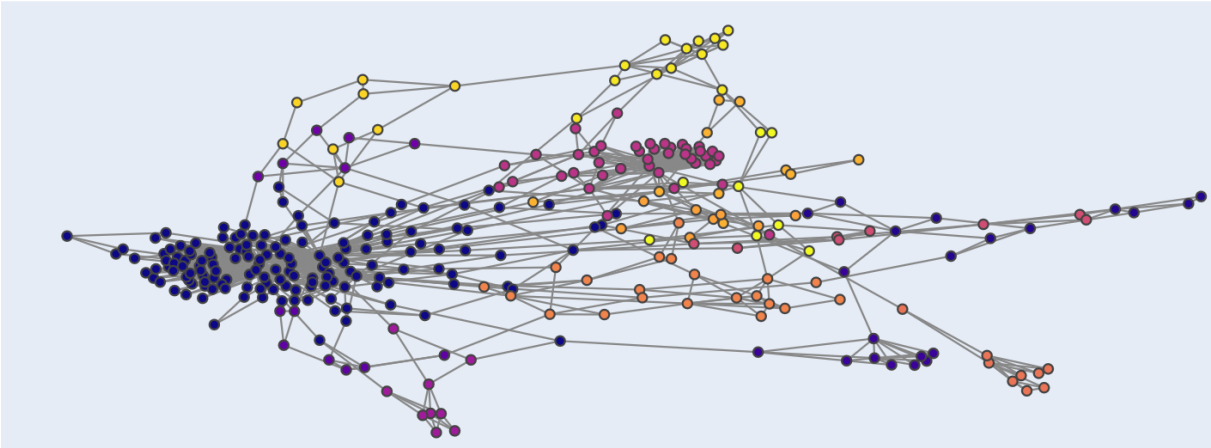
Figura 3.9 – Palavras chaves com maiores TFIDF.



Fonte: Imagem Gerada pelo autor em Python.

Mantendo apenas os documentos dos clusters selecionados, é possível ressaltar os tópicos/temas mais relevantes da base de dados, conforme a estrutura da rede K-NN, conforme Figura 3.10.

Figura 3.10 – Rede KNN.



Fonte: Imagem Gerada pelo autor em Python.

Essas análises permitem observar que é possível obter resultados equivalentes aos obtidos por Alves *et al.* (2024) com a aplicabilidade de utilização de Inteligência Artificial

(IA) no processo de pesquisas científicas.

3.3 Identificação de Padrões

A análise revelou vários padrões significativos nos dados. A Rede de Coocorrência de Palavras-Chave, após o refinamento da pequisa, destacou-se a presença de termos centrais como digital transformation, e-government, e-governance, smart city e big data. A Rede de Cocitação destacou que os artigos mais influentes estão concentrados no tema e-government, com destaque para Layne e Lee (2001) e West (2004). Já a Rede de Acoplamento Bibliográfico, identificou agrupamentos em torno de temas emergentes como smart governance e artificial intelligence, indicando conexões contemporâneas entre os documentos.

3.4 Interpretação das Relações

A proximidade entre e-government e digital transformation indica que essas áreas são co-dependentes e frequentemente exploradas em conjunto, refletindo esforços para modernizar a administração pública. Os artigos mais cocitados desempenham papel central na definição de paradigmas teóricos, com Layne e Lee (2001) sendo consistentemente referenciados em múltiplos estudos. A análise temporal mostrou um aumento significativo nas publicações após 2020, associado ao impacto da pandemia e à aceleração da digitalização. Os clusters formados pelo KNN indicaram que as principais tendências incluem smart cities, governança eletrônica e big data, enquanto áreas periféricas como digital twins começam a ganhar atenção.

3.5 Comparação com Análises Anteriores

As ferramentas tradicionais identificaram padrões semelhantes, mas com maior dependência de visualizações complexas e configurações manuais. A introdução do KNN simplificou a análise ao automatizar a formação de clusters e eliminar a necessidade de tesauros extensos. O uso do KNN proporcionou maior precisão na identificação de padrões sem comprometer a profundidade da análise. Os resultados do KNN foram equivalentes ou superiores, especialmente na detecção de termos emergentes e agrupamentos não evidentes nas redes visuais.

As análises anteriores destacaram tendências como e-governance e IT governance. O KNN, além de corroborar essas descobertas, identificou novas áreas como telemedicine e e-democracy, ampliando o escopo de insights.

4 CONCLUSÕES

Este estudo evidenciou que a aplicação de inteligência artificial (IA), com destaque para o algoritmo K-Nearest Neighbors (KNN), oferece uma abordagem eficiente, escalável e inovadora para a análise bibliométrica. Ao substituir o uso manual de ferramentas tradicionais como Gephi, VOSviewer e Iramuteq, a solução baseada em KNN permitiu a identificação de padrões, agrupamentos e tendências na literatura científica de forma simplificada e com maior precisão.

Como principais contribuições, é possível destacar a eficiência no processamento, tanto na automação do pré-processamento dos dados como na tokenização e vetorização, simplificando a análise de grandes volumes de dados. Além disso, o KNN se mostrou robusto para análises complexas, oferecendo insights que, em muitos casos, superaram os métodos tradicionais.

Ademais, esse estudo permitiu-se a flexibilidade analítica com a escolha de métricas de similaridades e ajustes de parâmetros permitindo a personalização da pesquisa. Essas observações confirmam que a aplicação de IA (KNN) pode não apenas reproduzir os resultados obtidos com ferramentas tradicionais, mas também trazer maior escalabilidade, eficiência e novos insights ao processo de análise bibliométrica, ampliando as possibilidades analíticas.

Como perspectivas e sugestão para estudos futuros, recomendamos a investigação de outras métricas de distância do KNN e a implementação de outros algoritmos como o DBSCAN e K-Means, bem como o uso de redes neurais e modelos de aprendizado profundo para enriquecer a análise com exploração de relações mais complexas e possibilidade de previsão de tendências futuras. O aprofundamento de implementações de análises com dados qualitativos, como análise de sentimento ou avaliação de impacto social para ampliar o escopo da análise bibliométrica.

Por fim, recomendamos a construção de ferramentas baseadas em Python e IA que possam ser disponibilizadas como plataformas acessíveis para pesquisadores, eliminando a necessidade de softwares comerciais, permitindo a análises em outras áreas do conhecimento, promovendo uma visão mais ampla sobre os avanços científicos e tecnológicos.

REFERÊNCIAS

ALVES, W. A. d. S. *et al.* Digital transformation in the public sphere: a bibliometric analysis. **Ciência da Informação**, v. 52, n. 2, p. 264–294, 2024.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi: an open source software for exploring and manipulating networks. *In*: **Proceedings of the international AAAI conference on web and social media**. [*S.l.: s.n.*], 2009. v. 3, n. 1, p. 361–362.

BLONDEL, V. D. *et al.* Fast unfolding of communities in large networks. **Journal of statistical mechanics: theory and experiment**, IOP Publishing, v. 2008, n. 10, p. P10008, 2008.

BORGATTI, E. Ucinet 6 for windows/borgatti everett and freeman. **Harvard: Analytic Technologies**, 2002.

CAMARGO, B. V.; JUSTO, A. M. Iramuteq: um software gratuito para análise de dados textuais. **Temas em psicologia**, Sociedade Brasileira de Psicologia, v. 21, n. 2, p. 513–518, 2013.

ECK, N. J. V.; WALTMAN, L. Visualizing bibliometric networks. *In*: **Measuring scholarly impact: Methods and practice**. [*S.l.: s.n.*]: Springer, 2014. p. 285–320.

ECK, N. J. van; WALTMAN, L. Crossref as a source of open bibliographic metadata. MetaArXiv, 2022.

ECK, N. V.; WALTMAN, L. Software survey: Vosviewer, a computer program for bibliometric mapping. **scientometrics**, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV ..., v. 84, n. 2, p. 523–538, 2010.

ELLEGAARD, O.; WALLIN, J. A. The bibliometric analysis of scholarly production: How great is the impact? **Scientometrics**, Springer, v. 105, p. 1809–1831, 2015.

KÜCHER, A.; FELDBAUER-DURSTMÜLLER, B. Organizational failure and decline–a bibliometric study of the scientific frontend. **Journal of Business Research**, Elsevier, v. 98, p. 503–516, 2019.

LAYNE, K.; LEE, J. Developing fully functional e-government: A four stage model. **Government information quarterly**, Elsevier, v. 18, n. 2, p. 122–136, 2001.

MORESI, E. A. D.; PINHO, I.; COSTA, A. P. How to operate literature review through qualitative and quantitative analysis integration? *In*: SPRINGER. **World Conference on Qualitative Research**. [*S.l.: s.n.*], 2022. p. 194–210.

MORESI, E. A. D. *et al.* Learning assessment: Mapping research and educational policy agendas. *In*: SPRINGER. **World Conference on Qualitative Research**. [*S.l.: s.n.*], 2021. p. 31–44.

NEWMAN, M. E. **Networks: an introduction**. [*S.l.: s.n.*]: Oxford university press, 2010.

NEWMAN, Y. *et al.* Enhancing phosphorus phytoremedation potential of two warm-season perennial grasses with nitrogen fertilization. **Agronomy journal**, Wiley Online Library, v. 101, n. 6, p. 1345–1351, 2009.

PRITCHARD, A. Statistical bibliography or bibliometrics. **Journal of documentation**, v. 25, p. 348, 1969.

RUHNAU, B. Eigenvector-centrality—a node-centrality? **Social networks**, Elsevier, v. 22, n. 4, p. 357–365, 2000.

WEST, D. M. E-government and the transformation of service delivery and citizen attitudes. **Public administration review**, Wiley Online Library, v. 64, n. 1, p. 15–27, 2004.

WHITE, M. D. *et al.* Content analysis: A flexible methodology. **Library trends**, Johns Hopkins University Press, v. 55, n. 1, p. 22–45, 2006.

ZUPIC, I.; ČATER, T. Bibliometric methods in management and organization. **Organizational research methods**, Sage Publications Sage CA: Los Angeles, CA, v. 18, n. 3, p. 429–472, 2015.