Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

DOI: 10.37943/17ALVR8114

Andrii Biloshchytskyi
Doctor of Technical Sciences, Professor, Vice-Rector for Science and Innovation
a.b@astanait.edu.kz, orcid.org/0000-0001-9548-1959
Astana IT University, Kazakhstan
Professor Department of Information Technologies,
Kyiv National University of Construction and Architecture, Ukraine

Malika Shamgunova
Master's degree student, Department of Computer Science and Engineering
malika.shamgun@gmail.com, orcid.org/0009-0001-8079-3578
Astana IT University, Kazakhstan

Svitlana Biloshchytska
Doctor of Technical Sciences, Associate Professor
bsv@astanait.edu.kz, orcid.org/0000-0002-0856-5474
Professor of the Department of Computational and Data Science,
Astana IT University, Kazakhstan
Professor Department of Information Technologies,
Kyiv National University of Construction and Architecture, Ukraine

EXPLORATION OF THE THEMATIC CLUSTERING AND COLLABORATION OPPORTUNITIES IN KAZAKHSTANI RESEARCH

Abstract: In today's academic environment, the rapid growth of research publications calls for advanced methods to organize and understand the extensive collections of academic work. This study aims to systematically categorize a substantial number of research paper abstracts from Kazakhstani institutions, focusing on identifying key themes and potential interdisciplinary collaboration opportunities. The dataset includes 13,356 abstracts from the Scopus database, covering a wide range of academic fields. The methodology of this research goes beyond traditional hand-done analysis by using advanced text analysis tools to organize the text data efficiently. This initial phase is crucial for summarizing each abstract's core content. The next steps of the analysis use this organized data to find and group similar thematic areas, considering the complex and multi-dimensional nature of academic research topics. The results reveal a diverse array of research themes, highlighting the dynamic academic contributions from Kazakhstan. Significant areas such as environmental science, technological advancements, linguistics, and cultural studies are among the prominent clusters identified. These insights not only provide an overview of current research directions but also highlight the potential for cross-disciplinary partnerships. Moreover, the findings have important implications for decision-makers, scholars, and educational institutions by illuminating key research areas and collaborative possibilities. This thematic overview acts as a guide for shaping research policies, fostering academic connections, and efficiently distributing resources within the scholarly community. Ultimately, this study adds to the academic conversation by offering a way to navigate and utilize the wealth of information in scientific literature, promoting a more collaborative and integrated research environment.

Keywords: data preprocessing; natural language processing; thematic clustering; research abstracts.

DOI: 10.37943/17ALVR8114

Introduction (Literary review)

In the vast expanse of academic research, the trend towards specialization has led scholars to delve deeply into niche areas, culminating in a rich tapestry of publications that mirror their specific interests and expertise. This thematic specialization is not merely a reflection of academic diversity but serves as a critical compass for navigating the complex landscape of scientific collaboration. The alignment of research interests and objectives is paramount when forging partnerships, as it lays the foundation for synergistic endeavors that can push the frontiers of knowledge.

However, amidst this specialization, a significant challenge emerges: the difficulty in discerning overarching trends and potential collaborative opportunities within and across disciplines. This issue is particularly pronounced in the rapidly evolving domain of Information Technology (IT), where the pace of innovation and the breadth of research areas can obscure the thematic intersections that might foster groundbreaking collaborative research.

Recognizing this challenge, the present study seeks to get the thematic landscape of IT research within the context of Kazakhstan's academic output. By mapping the thematic contours of this landscape, the study aims to identify not only the dominant research themes but also the potential nodes of interdisciplinary collaboration. This endeavor is significant as it contributes to a more nuanced understanding of the IT research ecosystem, enabling stakeholders to strategically navigate and leverage collaborative opportunities.

To achieve this, the study employs a methodological approach that synthesizes quantitative and qualitative analyses of research abstracts from Kazakhstani institutions. This dual approach facilitates a comprehensive thematic exploration, allowing for the identification of both explicit and emergent research themes.

The hypothesis underpinning this investigation posits that a systematic analysis of research abstracts can reveal discernible thematic clusters, which, in turn, can highlight potential areas for collaborative research. This hypothesis is investigated through a meticulous examination of research abstracts, employing state-of-the-art text analysis techniques to distill and categorize thematic elements.

In sum, this study not only aims to chart the thematic territories of IT research in Kazakhstan but also to serve as a beacon for fostering collaboration within the academic community. By delineating the thematic and collaborative landscape, it contributes valuable insights to the ongoing discourse on research specialization and interdisciplinary partnerships.

Al-Obaydy, Hashim, Najm and Jalal propose an innovative approach for categorizing research articles into thematic groups, leveraging Term Frequency-Inverse Document Frequency (TF-IDF) and K-means clustering. The methodology is designed to address the challenges researchers face in navigating the vast corpus of scientific literature, aiming to cluster text documents into meaningful groups that represent similar scientific fields [1]. Shetty and Kallimani introduce an innovative approach leveraging K-Means clustering for extractive text summarization, focusing on preserving semantic richness while eliminating redundancy [2]. Biloshchytskyi, Kuchansky, Andrashko, Biloshchytska, Kuzka, and Terentyev's study proposes a nuanced method for evaluating the scientific research activities of academics, focusing on scalar and integral estimations based on publication citations. Unlike traditional bibliometric indicators like the h-index and g-index, their approach ensures no loss of citation information by incorporating all citations in their evaluation. They achieve this through a system of linear algebraic equations that account for citations among scientists, with the solutions providing scalar evaluations [3]. Alsmadi and Alhami's research provides a comprehensive exploration of clustering and classification methodologies applied to a substantial dataset of personal emails for various objectives, including folder and subject classifications. Through the development and implementation of algorithms tailored for this extensive text collection, the study show-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

cases the effectiveness of classification based on NGram, particularly in the context of bilingual (English and Arabic) text data [4]. Rejito, Atthariq, and Abdullah's study explores the application of K-Means clustering in analyzing the tweet content of Tokopedia, a leading Indonesian e-commerce platform. Through text mining of 885 tweets, they identified 48 distinct clusters, which were further categorized into 5 major groups, revealing that tweets related to quizzes and prizes garnered the most engagement, while lifestyle content attracted the least [5]. Denny and Spirling's paper rigorously investigates the substantial influence of preprocessing decisions on the outcomes of unsupervised learning models in text analysis, specifically within political science research [6]. Hickman, Thapa, Tay, Cao, and Srinivasan's review provides a comprehensive examination of text preprocessing techniques in organizational research, highlighting the critical role these methodologies play in enhancing text mining processes. The study systematically assesses the implications of various preprocessing decisions on the reliability, validity, and statistical power of text analysis, offering empirically grounded recommendations for both open and closed vocabulary text mining [7]. Alhawarat and Hegazi's research revisits the integration of k-means clustering and topic modeling techniques for clustering Arabic documents, presenting a comparative study that emphasizes the enhancement of clustering quality through normalization and the combined method. Their study demonstrates that normalizing the vector space model significantly improves the quality and accuracy of k-means clustering [8]. Oti, Olusola, Eze, and Enogwe provide a thorough examination of the evolution and variations of k-means clustering algorithms, highlighting their pivotal role in partitioning data into distinct clusters based on Euclidean distance and the minimum distance rule [9]. Vijayarani, Ilamathi, and Nithya's comprehensive review delves into the critical preprocessing techniques in text mining, focusing on stop words elimination, stemming, and TF-IDF algorithms. Their exploration underscores the pivotal role of these preprocessing steps in enhancing the efficacy of text mining processes by improving the quality of the text data input [10]. Aubaidan, Mohd and Albared through a series of evaluations using overall purity and F-measure across various datasets of crime documents, the findings consistently indicate the superior performance of k-means++ over the traditional k-means, especially when applying cosine similarity measures [11]. Tabassum and Patil's survey provides an in-depth analysis of text pre-processing and feature extraction techniques critical to enhancing Natural Language Processing (NLP) applications. The study underscores the importance of meticulous pre-processing to remove irrelevant data and highlights the effectiveness of various techniques such as tokenization, stopword removal, and lemmatization in refining text data for computational processing. Furthermore, the survey emphasizes feature extraction methods like TF-IDF and Bag-of-Words, demonstrating their pivotal role in transforming processed text into numerical formats suitable for machine learning algorithms. The comparative analysis of these techniques reveals their significant impact on the performance of NLP tasks, suggesting a careful selection based on specific use case requirements [12]. Kadhim, Cheah, and Ahamed's research investigates the impact of preprocessing and dimension reduction techniques on the performance of text document clustering using the k-means algorithm. Their method, which integrates TF-IDF for term weighting and Singular Value Decomposition (SVD) for dimensionality reduction, demonstrates significant improvements in clustering English text documents. The experimental evaluation, conducted on BBC news and BBC sport datasets, shows that the proposed approach not only effectively reduces the dimensionality of the datasets but also maintains high clustering accuracy, achieving 95% for BBC news and approximately 94.67% for BBC sport, even when dimensions are reduced dramatically. This study underscores the importance of effective preprocessing and dimension reduction in enhancing the performance and accuracy of clustering algorithms in text mining applications [13]. Al-Anazi, AlMahmoud, and Al-Turaiki's study evaluates clustering techniques, including

DOI: 10.37943/17ALVR8114

k-means, k-means fast, and k-medoids, using different similarity measures for grouping capstone project documents from King Saud University. The research finds that k-means and k-medoids combined with cosine similarity yield the best clustering performance. The study also observes an improvement in clustering quality with an increase in the number of clusters. This work contributes to the understanding of clustering capstone projects, revealing categories such as e-health, Arabic and Islamic applications, and location-based services and others offering valuable insights for both students and academic administrations [14]. Bafna, Pramod, and Vaidya's paper evaluates the efficiency of document clustering using the TF-IDF approach alongside fuzzy K-means and hierarchical algorithms across various datasets, including News 20, Reuters, and emails. The study demonstrates that hierarchical agglomerative clustering (HAC) outperforms fuzzy K-means in terms of lower entropy and higher F-measure values for most datasets, indicating more coherent and distinct clusters [15]. Arora, Deepali, and Varshney's study conducts a comparative analysis of the K-Means and K-Medoids clustering algorithms on a large dataset, illustrating the efficiency and robustness of K-Medoids over K-Means [16]. Zhou, Xu, Liu, Chang and Xiao propose a method for measuring text similarity called Word Vector Distance Decentralization (WVDD), which is especially tailored for the complexities of semantic relationships in Chinese language texts. Experimental results, validated using the F-measure, highlight the method's effectiveness in handling semantic cognition, offering significant improvements over traditional models like Doc2Vec and Bag-of-Words [17]. Singh and Shashi's paper explores various text vectorization methods for clustering news articles to enhance multi-document summarization based on trending topics. By comparing the effectiveness of TF-IDF, Word2Vec, and Doc2Vec in clustering using the k-means algorithm, the study contributes valuable insights into natural language processing techniques for organizing large volumes of textual data [18]. The study by Naeem and Wumaier delves into optimizing the K-means clustering algorithm for English text, with a focus on identifying the optimal value of K. It offers a comprehensive exploration of various methods for determining K's true value, such as the Elbow Method, Gap Statistic Method, and Silhouette Method, addressing one of K-means' known limitations [19]. Kim and Gil introduce a research paper classification system leveraging TF-IDF and LDA schemes for efficient clustering based on abstract analysis. It adeptly combines well-established text mining techniques with K-means clustering to categorize a vast array of academic literature [20].

Analysis and comparison of the existing text processing techniques

The following examination delves into the mathematical foundation of stemming, lemmatization, Bag of Words (BoW), and Term Frequency-Inverse Document Frequency (TF-IDF).

Stemming is generally rule-based, and while there isn't a mathematical equation that defines the process, it operates on the principle of removing prefixes and suffixes according to patterns and heuristics. The Porter Stemming algorithm, for instance, employs a series of about 60 rules applied sequentially. Each rule is of the form:

$$(condition)\ S1 \rightarrow S2 \tag{1}$$

where $S1$ is a suffix to be replaced by $S2$ if a condition (usually related to the measure of the stem, or m is satisfied. The measure m is calculated as:

$$m = \frac{Count\ of\ VC\ sequences}{2} \tag{2}$$

Here, V stands for vowels and C for consonants in the word represented as a sequence of V and C. The more complex the word, the higher the measure, and certain rules only apply for words above a certain measure threshold.

Lemmatization, unlike stemming, requires an understanding of the word's part of speech and context. Lemmatization might involve complex algorithms that take into account the

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

word's morphological analysis. The algorithms often refer to extensive lexicons and involve database lookups, which, from a computational perspective, can be represented by functions that map inflected forms to their lemmas:

$$Lemma(word, pos) = f(word, pos) \tag{3}$$

where the function *f* takes a word and its part of speech *pos* and returns the corresponding lemma after looking up a lexicon or applying morphological rules.

Analysis and comparison of the existing feature extraction techniques

The Bag of Words (BoW) model treats documents as vectors in a multidimensional space where each dimension corresponds to a unique word from the corpus vocabulary. The value in each dimension is the count of the word's occurrences in the document. For a corpus C containing N documents $\{d_1, d_2, \ldots, d_N\}$ and M unique terms $\{t_1, t_2, \ldots, t_M\}$, the mathematical representation of the BoW model is a $M \times N$ matrix $\boldsymbol{B}$ where each entry b_{ij} is the frequency of term i in document j:

$$\mathbf{B} = \begin{bmatrix} b_{11} & \cdots & b_{1N} \\ \vdots & \ddots & \vdots \\ b_{M1} & \cdots & b_{MN} \end{bmatrix} \tag{4}$$

where $b_{ij} = \text{frequency of term } t_i \text{ in document } d_j$.

TF-IDF adds an additional layer by adjusting the raw term frequencies according to their importance across all documents. The term frequency (TF) of a word in a document is normalized by the total number of words in the document, while the inverse document frequency (IDF) is calculated using the logarithmically scaled fraction of the total number of documents by the number of documents containing the word:

$$TF_{ji} = \frac{f_{ji}}{\sum_k f_{jk}} \tag{5}$$

$$IDF_j = \log\left(\frac{D}{|\{d: i \in d\}|}\right) \tag{6}$$

$$TFIDF_{ji} = TF_{ji} \times IDF_i \tag{7}$$

where:

f_{ij} is the raw count of term i in document j,

$\sum_k f_{jk}$ is the total number of terms in document j,

D is the total number of documents, and

$|\{d: i \in d\}|$ is the number of documents that contain term i.

The TF-IDF score for term i in document j is given by $t_{ij} = tf_{ij} \times idf_i$, where tf_{ij} is the term frequency of t_i in d_j, and idf_{ij} is the inverse document frequency of t_i across all documents. The TF-IDF matrix $\boldsymbol{T}$ is then:

$$\mathbf{T} = \begin{bmatrix} t_{11} & \cdots & t_{1N} \\ \vdots & \ddots & \vdots \\ t_{M1} & \cdots & t_{MN} \end{bmatrix} \tag{8}$$

The result is a weighted model where each document is represented by a vector $\mathbf{d}_i$ in a V-dimensional space, and each vector component d_{ij} corresponds to the TF-IDF score of the word i in document j.

DOI: 10.37943/17ALVR8114

Implications for Research Abstract Processing

In the processing of research abstracts, these mathematical models facilitate the extraction of features that reflect the content's semantics. The choice between stemming and lemmatization impacts the resulting word forms fed into subsequent mathematical models for text representation. Stemming may produce a more compact feature space, potentially advantageous in high-dimensional datasets where the computational burden is a concern. Lemmatization, while more computationally intensive, yields a richer and more accurate feature set, beneficial for algorithms that rely on semantic precision. When abstracts are vectorized using BoW or TF-IDF, the underlying mathematical structure determines how the document space is modeled. BoW's simplicity can be beneficial for algorithms that can handle high-dimensional, sparse data. In contrast, TF-IDF's nuanced representation, emphasizing terms that provide the most informational gain about the document's content, is especially useful in modelling and clustering tasks where the relative importance of terms is critical. The mathematical frameworks behind stemming, lemmatization, BoW, and TF-IDF offer distinct perspectives on text processing, each with trade-offs in complexity, computational cost, and semantic fidelity. For research abstracts, the choice hinges on the desired balance between computational efficiency and the depth of linguistic processing required for the task. In clustering, classification, and information retrieval within academic corpora, leveraging the detailed representation provided by lemmatization and TF-IDF can significantly enhance performance and insights gained.

Fig. 1a represents the Bag of Words matrix after stemming. Each row corresponds to an abstract, and each column represents a stemmed term. The values indicate the presence or frequency of each term in the corresponding abstract. This matrix is instrumental in analyzing the text data, highlighting the distribution of common word roots across the dataset.

In Fig. 1b, the Bag of Words matrix after lemmatization is depicted. Similar to Figure 1a, it shows the frequency of terms, but these terms have been reduced to their lexicographical base form, or lemma. This approach retains more of the word's original meaning and grammatical correctness.

Stemmed BoW Matrix

	advanc	advantag	and	are	art	carbon	certain	chang	climat	comput	...	problem	process	quantum	reduc	significantli	state	sustain	system	the
Abstract 1	1	0	0	0	1	0	0	0	0	0	...	0	1	0	0	1	1	0	0	2
Abstract 2	0	1	1	0	0	0	1	0	0	2	...	1	0	1	0	0	0	0	0	0
Abstract 3	0	0	0	1	0	1	0	1	1	0	...	0	0	0	1	0	0	1	1	1

3 rows × 36 columns

Figure 1a. Stemmed Bag of Words Matrix representation

Lemmatized BoW Matrix

	advanced	advantage	and	are	art	carbon	certain	change	climate	computational	...	problem	processing	quantum	reducing	significantly	state
Abstract 1	1	0	0	0	1	0	0	0	0	0	...	0	1	0	0	1	1
Abstract 2	0	1	1	0	0	0	1	0	0	1	...	1	0	1	0	0	0
Abstract 3	0	0	0	1	0	1	0	1	1	0	...	0	0	0	1	0	0

Figure 1b**.** Lemmatized Bag of Words Matrix representation

Fig. 1c provides a visual representation of the Term Frequency-Inverse Document Frequency (TF-IDF) matrix using stemmed words. Unlike the simple frequency counts in the BoW matrix, the TF-IDF matrix reflects the importance of each term within an abstract relative to its

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

commonality across all abstracts. This serves to prioritize unique terms which might be more relevant for analysis.

The lemmatized TF-IDF matrix is shown in Fig. 1d. It builds on the concept of the TF-IDF matrix in Fig. 1c but uses lemmatized terms. This can be particularly useful for maintaining the integrity of the dataset's language structure while still benefiting from the TF-IDF model's ability to emphasize terms that provide the most informational value about each document's content.

	advanc	advantag	and	are	art	carbon	certain	chang	climat	comput	...	problem	process	quantum	reduc	significantli
Abstract 1	0.264992	0.000000	0.000000	0.000000	0.264992	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.264992	0.000000	0.000000	0.264992
Abstract 2	0.000000	0.267959	0.267959	0.000000	0.000000	0.000000	0.267959	0.000000	0.000000	0.535917	...	0.267959	0.000000	0.267959	0.000000	0.000000
Abstract 3	0.000000	0.000000	0.000000	0.276458	0.000000	0.276458	0.000000	0.276458	0.276458	0.000000	...	0.000000	0.000000	0.000000	0.276458	0.000000

3 rows × 36 columns

Figure 1c. Stemmed TF-IDF Matrix representation

Lemmatized TF-IDF Matrix

	advanced	advantage	and	are	art	carbon	certain	change	climate	computational	...	problem	processing	quantum	reducing
Abstract 1	0.264992	0.000000	0.000000	0.000000	0.264992	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.264992	0.000000	0.000000
Abstract 2	0.000000	0.289554	0.289554	0.000000	0.000000	0.000000	0.289554	0.000000	0.000000	0.289554	...	0.289554	0.000000	0.289554	0.000000
Abstract 3	0.000000	0.000000	0.000000	0.276458	0.000000	0.276458	0.000000	0.276458	0.276458	0.000000	...	0.000000	0.000000	0.000000	0.276458

3 rows × 37 columns

Figure 1d. Lemmatized TF-IDF Matrix representation

The differences between the Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF)

In BoW, b_{ij} is simply the count of occurrences of term t_i in document d_j. In TF-IDF, tf_{ij} is similar to b_{ij} but is adjusted by the idf component to reflect the term's importance. BoW does not account for the term's commonality across documents. TF-IDF introduces idf_{ij}, decreasing the weight for terms that occur in many documents, thereby highlighting terms unique to specific documents. BoW matrices tend to be sparse but can emphasize common terms. TF-IDF matrices are also sparse but adjust the emphasis towards terms that are significant in a document-specific context. BoW matrices capture the surface-level textual features without semantic weighting. TF-IDF matrices encode an additional layer of semantic importance by weighing the terms, making them more suitable for tasks requiring nuanced understanding, like document similarity and topic modeling. The transition from BoW to TF-IDF representation is akin to moving from a simple count-based model to a more refined model that considers both the term frequency and the information it carries within the corpus context. Mathematically, this is encapsulated in the transformation from $\mathbf{B}$ to $\mathbf{T}$, where the latter matrix adjusts each term's contribution based on its distribution across the corpus, offering a more discriminative and informative representation for subsequent analysis.

In Fig. 2a, we observe the heatmap showcasing the term frequency distribution within the stemmed Bag of Words model. This representation enables us to discern how stemming consolidates similar words by trimming them down to their root forms, thereby simplifying the textual data while preserving its essential content.

Moving to Fig. 2b, the heatmap displays the term frequency distribution for the lemmatized Bag of Words model. Lemmatization, in contrast to stemming, reduces words to their lexico-

DOI: 10.37943/17ALVR8114

graphically correct base forms, taking into consideration the morphological analysis of words. This approach maintains a more linguistically accurate representation of the text, which can be crucial for certain NLP tasks where the precise form of words carries significant meaning.

Fig. 2c delves into the stemmed Term Frequency-Inverse Document Frequency (TF-IDF) model, where we extend our analysis beyond mere term frequencies. TF-IDF introduces an additional layer of information by considering the overall importance of terms within the corpus. This weighted approach helps in emphasizing terms that are uniquely significant to certain documents, thereby offering a more nuanced understanding of the textual data.

Lastly, Fig. 2d presents the heatmap for the lemmatized TF-IDF model. Similar to the stemmed TF-IDF model, this representation leverages both term frequencies and their document-wide significance but does so using lemmatized terms. This combination allows us to appreciate the subtleties in how lemmatization, coupled with TF-IDF weighting, captures the essence of the text, potentially revealing different patterns or themes compared to its stemmed counterpart.

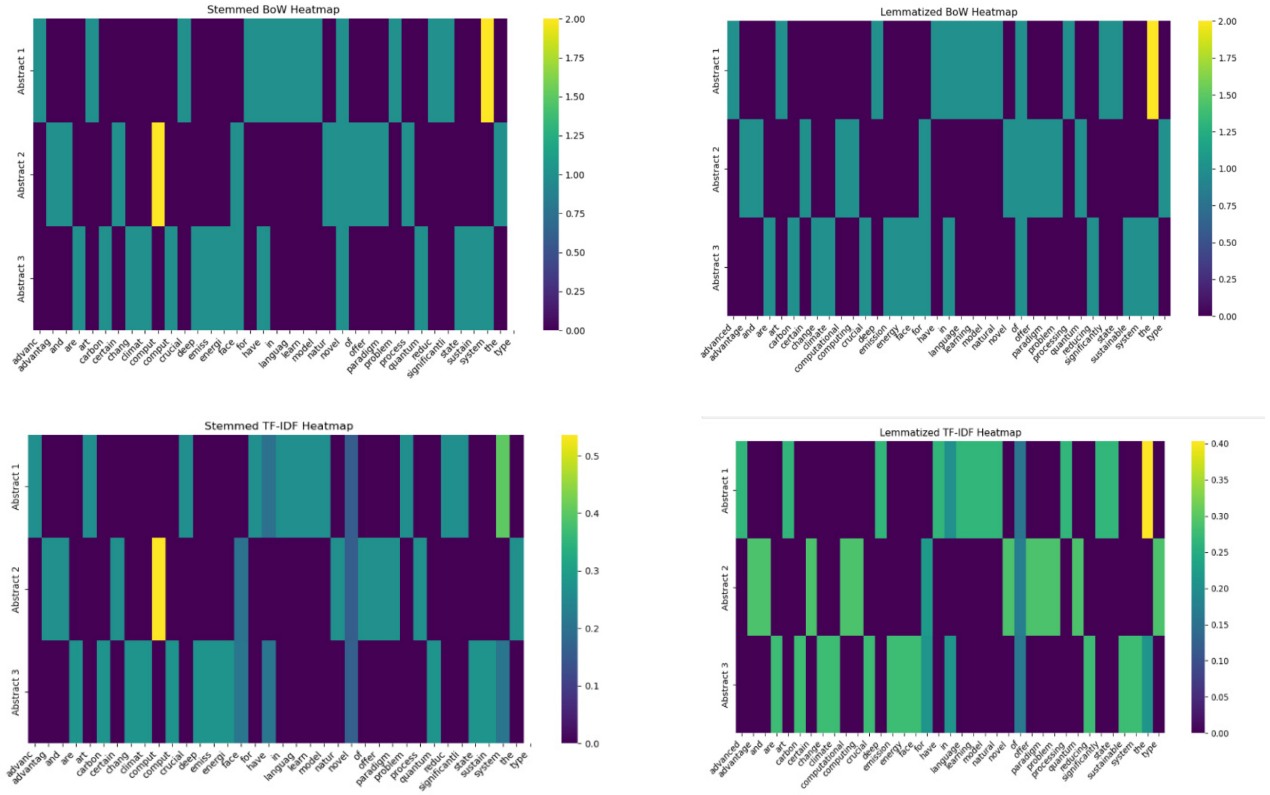

Figure 2 (a, b, c, d). Heatmaps of stemmed bag of words, lemmatized bag of words, stemmed term frequency-inversed document frequency, lemmatized term frequency-inversed document frequency

Together, these figures provide a comprehensive visual comparison between stemming and lemmatization, as well as between the Bag of Words and TF-IDF methodologies, illustrating the diverse ways in which textual data can be quantified and analyzed.

The processing of stop words

The removal of stop words plays a crucial role in processing and analyzing text data, especially in research abstract corpora. Stop words are common words that carry minimal individual meaning and occur frequently across texts, such as “the,” “is,” “at”. Their removal is a standard preprocessing step in many Natural Language Processing (NLP) and text mining tasks.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

In the vector space model, a document d can be represented as a vector $\vec{d} = \{w_1, w_2, \ldots, w_n\}$, where each w_i corresponds to a term's weight in the document. When stop words are removed, the dimensionality of $\vec{d}$ is effectively reduced, and the weights are recalculated based only on the remaining terms. This results in a more concise and meaningful representation of each document, focusing on terms that carry more semantic significance. The removal of stop words is a foundational preprocessing step in text analysis, especially for research abstracts, where the emphasis is on extracting meaningful insights from concise texts. By eliminating these common words, researchers can achieve a more focused analysis, reduce computational complexity, and enhance the performance of various NLP algorithms.

In Fig. 3a, the heatmap illustrates the frequency distribution of terms in the stemmed

Bag of Words model with stop words removed. Similarly, Fig. 3b presents the lemmatized Bag of Words model, allowing for a comparison between stemming and lemmatization in terms of term frequency. Fig. 3c and 3d extend this analysis to the Term Frequency-Inverse Document Frequency (TF-IDF) representation, highlighting the impact of term weighting and normalization in the stemmed and lemmatized text data, respectively.

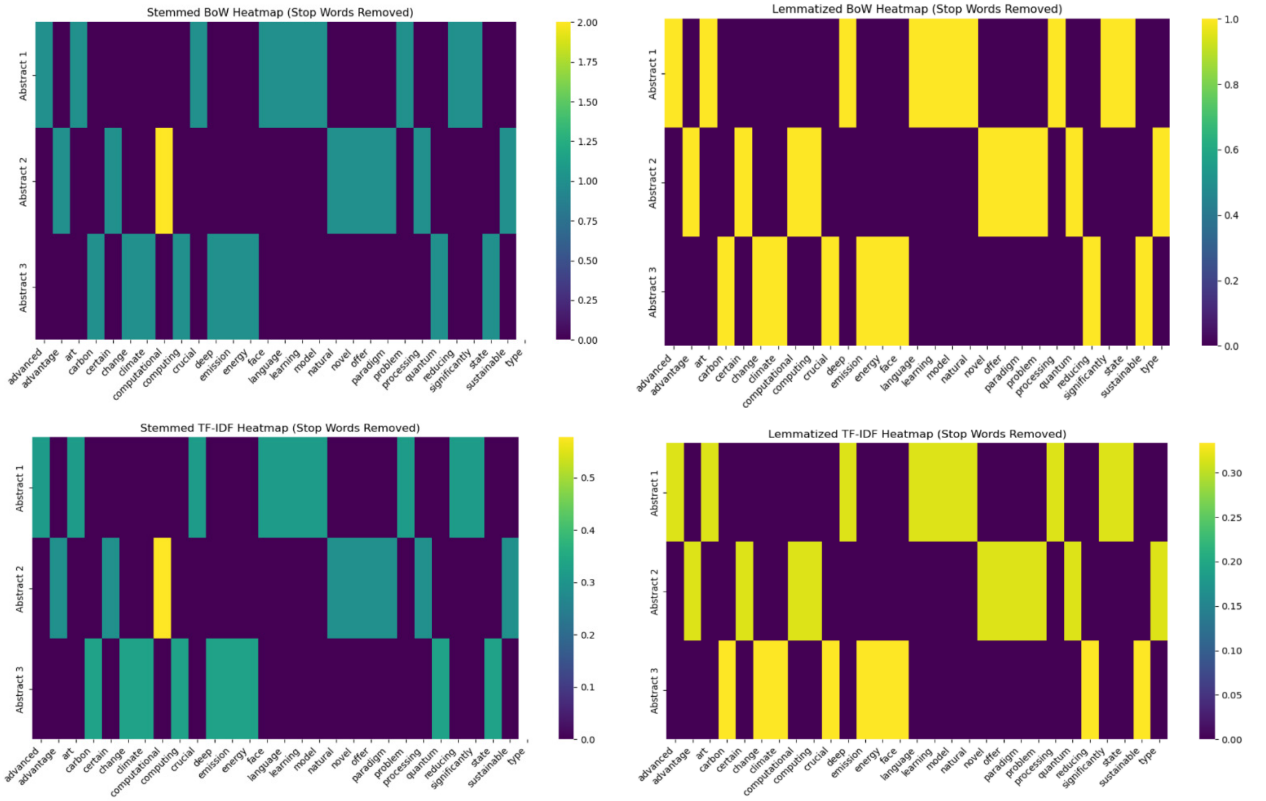

Figure 3 (a,b,c,d). Heatmap representations with stop words removed

Methods and Materials

Data Collection Method

The dataset was methodically collected from the Scopus database, a prominent repository of peer-reviewed literature abstracts and citations. The data collection focused on English-language articles to maintain linguistic clarity in the analysis. Only open-access articles were included to ensure data availability and support the reproducibility of the findings. The study was confined to journal articles to benefit from their structured format and peer-reviewed quality. The selection criteria emphasized articles with Kazakhstani author affiliations to highlight the region's academic contributions. The dataset includes only final versions of ar-

DOI: 10.37943/17ALVR8114

ticles from journals to ensure the use of complete and reliable academic materials. A detailed data extraction template was used to gather bibliometric details, such as full author names for authorship analysis, paper titles for thematic insight, citation counts for impact assessment, DOIs for permanent electronic access, URLs for full-text access, abstracts for content synthesis, and both author and index keywords for conceptual and indexing purposes, along with Scopus Identification Numbers (EIDs) for unique document identification. The second dataset is also collected from Scopus but the limit was set to subject area Computer Science papers. The sample size of the first dataset is 13356 records. The sample size of the second dataset is 2062 records.

Data preprocessing steps and noise removal

Data preprocessing involved essential noise removal and text refinement to ensure the dataset's quality and analysis readiness. Initially, entries lacking substantial content, specifically those marked "[No abstract available]," were excluded. Further refinement entailed the removal of non-ASCII characters from the abstracts, employing a regular expression to retain only standard English letters, numerals, and certain punctuation, thereby discarding any irrelevant symbols or characters.

The textual content underwent tokenization to break down the text into individual, analyzable units. Subsequent steps included the elimination of stop words–common but insignificant words like "the," "is," and "in"–using a predefined list from the Natural Language Toolkit (NLTK). Additionally, a regular expression was applied to remove unnecessary punctuation, further purifying the text.

Lemmatization was then applied, standardizing inflected word forms to their base lemmas, thus reducing the dataset's complexity and enhancing its uniformity. These collective preprocessing efforts were instrumental in refining the dataset for more effective clustering and thematic analysis.

Feature Extraction

Following the preprocessing steps, including lemmatization, the next phase in our analysis involved feature extraction using the Term Frequency-Inverse Document Frequency (TF-IDF) technique on the corpus of abstracts. This method quantitatively evaluates the importance of a word within a document in the context of a collection of documents, or corpus. The TF-IDF value is calculated using two components: Term Frequency (TF) and Inverse Document Frequency (IDF).

The Term Frequency (TF) for a word is defined as the ratio of the number of times the word appears in an abstract to the total number of words in that abstract, expressed as

$$TF(t,a) = \frac{f_{t,a}}{\sum_{t' \epsilon a} f_{t',a}} \tag{9}$$

where $f_{t,a}$ is the frequency of term t in abstract a and $\sum_{t' \epsilon a} f_{t',a}$ is the sum of all term frequencies in abstract a.

The Inverse Document Frequency (IDF) is calculated using

$$IDF(t,A) = \log \frac{N}{|\{a \epsilon A : t \epsilon a\}|} \tag{10}$$

where N represents the total number of abstracts in the corpus A, and $|\{a \epsilon A : t \epsilon a\}|$ is the count of abstracts that contain term t. The logarithmic component in the IDF equation ensures that terms appearing in numerous abstracts do not disproportionately influence the analysis.

The comprehensive TF-IDF score for a term t within an abstract a across the corpus A is thus

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

$$TFIDF(t, a, A) = TF(t, a) \times IDF(t, A) \quad (11)$$

Terms with high TF-IDF scores, indicating their prevalence in a few abstracts, are deemed to carry more analytical significance. Conversely, terms with low scores, found commonly across the corpus, are considered less critical.

By converting the processed abstracts into a matrix of TF-IDF features, we constructed a numerical representation of the textual content, capturing the relative weight of terms both within individual abstracts and across the corpus as a whole. This numerical framework is vital for the next stage of clustering analysis, as it enables a quantitative comparison of the textual content of the abstracts, grouping them into clusters based on thematic similarities.

Clustering research abstracts and identifying number of clusters

To effectively group the corpus of research abstracts A, we employed the k-means++ clustering algorithm, an iterative portioning method that optimizes the clustering process. The k-means++ algorithm enhances the k-means approach by selecting initial cluster centers (centroids) in a manner that reduces variance within clusters. Our implementation iterated over a range of 2 to 35 potential clusters

For the Elbow Method, the optimal number of clusters k is determined by observing the point where the Within-Cluster Sum of Squares (WCSS) begins to diminish less rapidly as k increases. WCSS for a given k is defined as

$$WCSS(k) = \sum_{i=1}^{k} \sum_{a \epsilon A_i} ||a - \mu_i||^2 \quad (12)$$

where A_i denotes the set of abstracts in cluster *i, a* represents an individual abstract, and μ_i is the centroid of cluster A_i.

For the Silhouette Score, the cohesion and separation of clusters are evaluated for each abstract a. The silhouette score for a is calculated as

$$s(a) = \frac{b(a) - \bar{d}(a, A_i)}{\max\{\bar{d}(a, A_i), b(a)\}} \quad (13)$$

where $\bar{d}(a, A_i)$ is the mean distance from abstract a to all other abstracts within its own cluster A_i, indicating intra-cluster cohesion, and, $b(a)$ represents the lowest mean distance from a to the abstracts in any other cluster, indicating the nearest cluster to which a does not belong, thus measuring inter-cluster separation.

Results

To determine the number of clusters the Elbow method showed in Fig. 4a that the number of clusters should be set to 20 and the Silhouette method showed in Fig. 4b that the number of clusters should be set to 32.

DOI: 10.37943/17ALVR8114

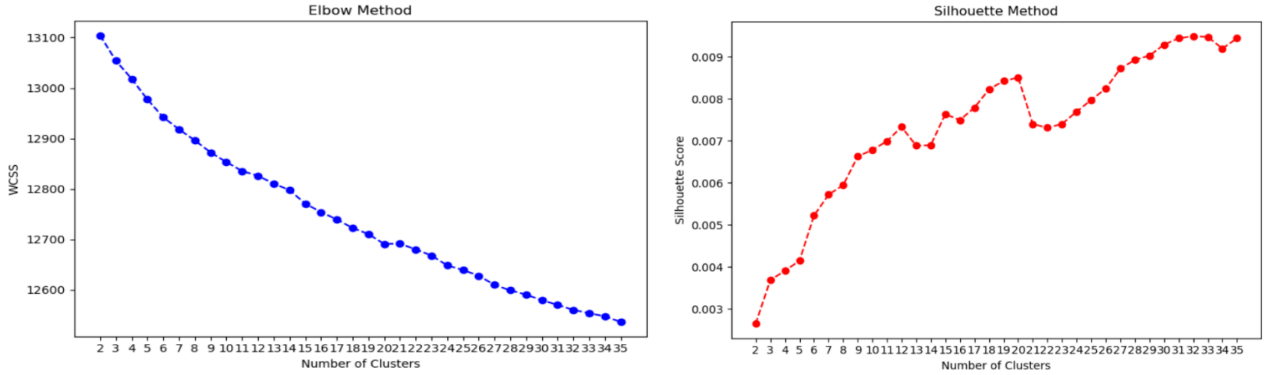

Figure 4 (a, b). Elbow Method and Silhouette Method for determination of optimal k

The Fig. 5 shows top 10 terms per cluster with overall 32 clusters. Cluster 13 encompasses 20.6% of the data points, indicating a significant concentration in this category. The prominent terms within this cluster include 'study,' 'result,' 'method,' 'kazakhstan,' 'data,' 'model,' 'effect,' 'analysis,' 'research,' and 'article.' Given this distribution, a further iteration of k-means clustering may be warranted for more granular categorization, or alternatively, a refinement of the tf-idf matrix could be achieved by removing stop words that are commonly used in research abstract terminology to enhance the distinctiveness of the cluster characterizations.

```
Top terms per cluster:
Cluster 0: scalar, gravity, field, solution, cosmological, inflation, model, gravitational, theory, equation
Cluster 1: development, kazakhstan, economic, country, social, economy, tourism, state, republic, article
Cluster 2: soil, fertilizer, crop, plant, water, content, kazakhstan, yield, study, area
Cluster 3: energy, neutron, mev, state, nucleus, reaction, cluster, decay, ion, experimental
Cluster 4: education, educational, professional, pedagogical, student, development, school, university, training, competence
Cluster 5: teacher, research, education, school, student, professional, future, training, teaching, pedagogical
Cluster 6: coating, ceramic, strength, phase, concrete, property, resistance, irradiation, material, increase
Cluster 7: problem, equation, boundary, solution, differential, integral, value, condition, solvability, existence
Cluster 8: system, model, algorithm, network, proposed, method, based, data, control, power
Cluster 9: energy, financial, economic, price, renewable, country, bank, oil, market, kazakhstan
Cluster 10: plasma, magnetic, field, star, surface, electron, line, particle, temperature, dense
Cluster 11: hole, black, mirror, particle, accretion, spacetime, collapse, horizon, solution, singularity
Cluster 12: dark, model, cosmological, universe, parameter, energy, matter, gravity, hubble, redshift
Cluster 13: study, result, method, kazakhstan, data, model, effect, analysis, research, article
Cluster 14: section, jet, zeus, cross, gev, hera, q2, luminosity, detector, photon
Cluster 15: patient, treatment, group, clinical, disease, study, therapy, year, cancer, month
Cluster 16: specie, plant, genus, population, kazakhstan, habitat, diversity, family, lake, region
Cluster 17: milk, meat, product, camel, flour, content, protein, acid, food, fat
Cluster 18: ore, mining, rock, deposit, mineral, uranium, copper, leaching, gold, mine
Cluster 19: operator, inequality, space, hardy, weighted, function, type, paper, differential, boundedness
Cluster 20: film, thin, surface, nm, substrate, oxide, ion, electron, deposition, spectroscopy
Cluster 21: health, risk, country, covid, disease, 19, child, woman, age, hiv
Cluster 22: equation, solution, problem, numerical, wave, nonlinear, system, method, model, differential
Cluster 23: coal, waste, fuel, combustion, gas, process, material, composition, oil, production
Cluster 24: student, learning, university, research, education, school, study, teaching, educational, technology
Cluster 25: water, river, basin, groundwater, irrigation, resource, lake, reservoir, area, kazakhstan
Cluster 26: gene, genetic, wheat, strain, genome, population, sequence, kazakhstan, resistance, dna
Cluster 27: cell, cancer, protein, vaccine, mouse, tumor, expression, virus, extract, immune
Cluster 28: heat, fiber, temperature, sensor, solar, thermal, grating, bragg, energy, flow
Cluster 29: carbon, oil, catalyst, acid, surface, composite, polymer, ion, temperature, solution
Cluster 30: language, kazakh, russian, linguistic, cultural, text, translation, word, english, article
Cluster 31: algebra, free, lie, prove, nilpotent, symmetric, automorphism, dimensional, derivation, identity
```

Figure 5. Top terms per cluster

Computer Science area of research abstracts clusters

To determine the number of clusters the Elbow method showed in Fig. 6a that the number of clusters should be set to 29. To determine the number of clusters the Silhouette method showed in Fig. 6b that the number of clusters should be set to 9.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

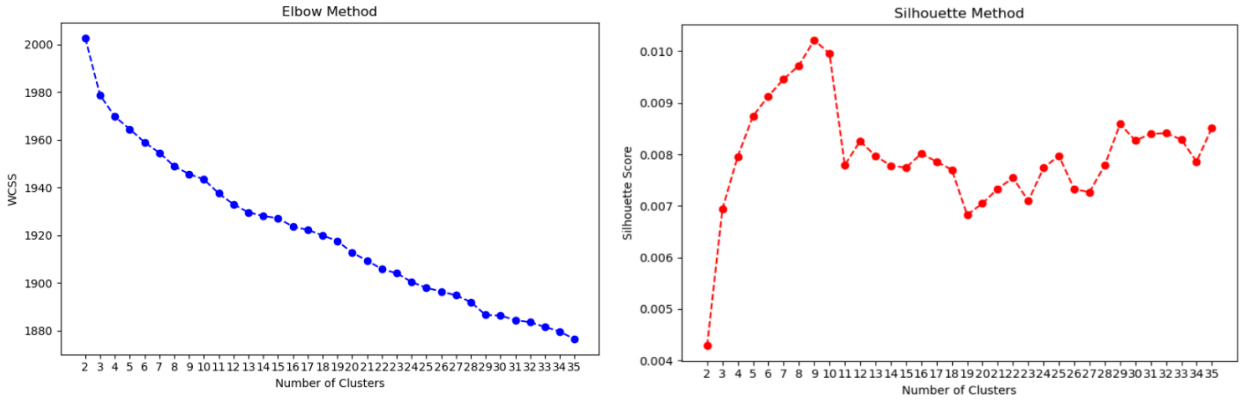

Figure 6 (a, b). Elbow Method for determination of optimal k (Subject area: Computer Science)

The Fig. 7 shows top 10 terms per cluster with overall 9 clusters.

```
Top terms per cluster:
Cluster 0: equation, problem, solution, boundary, differential, condition, operator, inverse, method, value
Cluster 1: signal, noise, channel, interference, radio, system, frequency, performance, communication, network
Cluster 2: energy, temperature, heat, water, result, power, model, material, study, method
Cluster 3: network, neural, model, image, learning, detection, data, accuracy, machine, method
Cluster 4: algorithm, system, control, robot, proposed, function, design, model, method, time
Cluster 5: student, education, teacher, research, educational, technology, study, school, university, learning
Cluster 6: language, kazakh, speech, text, word, corpus, translation, sentence, model, recognition
Cluster 7: information, system, model, development, data, process, management, method, decision, scientific
Cluster 8: product, grain, plant, crop, production, acid, flour, food, study, protein
```

Figure 7. Silhouette Method for determination of optimal k (Subject area: Computer Science)

Upon application of the K-Means clustering algorithm with an optimal cluster count of 9, as indicated by silhouette scores, the research abstracts were categorized into distinct clusters. Additionally, Fig. 8 represents a pie chart was constructed to illustrate the proportional distribution of data points across the nine clusters, offering a visual representation of the cluster sizes relative to the entire dataset.

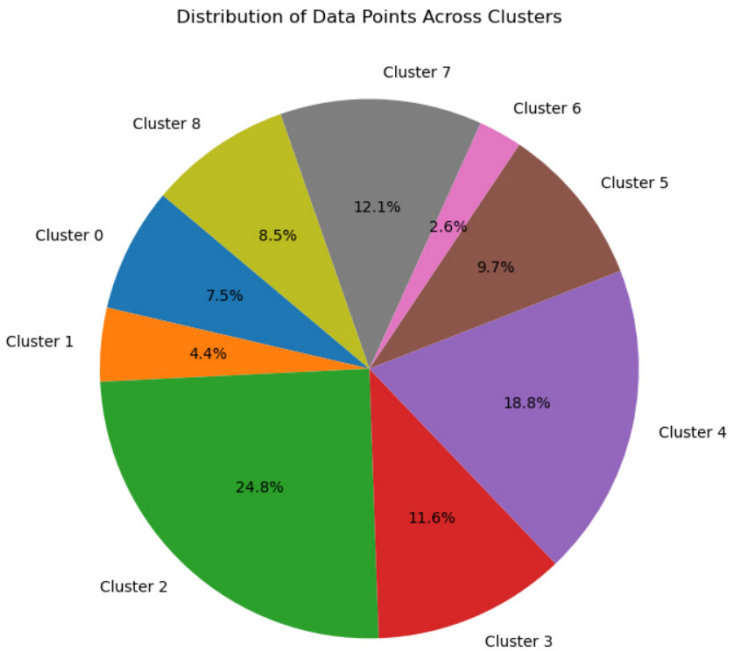

Figure 8. The distribution of data points across clusters (Subject area: Computer Science)

DOI: 10.37943/17ALVR8114

Discussion

The information technology system, leveraging the Scopus API, incorporates thematic tags and the relationships between scientists. These tags are vital for users of the platform, providing a methodical approach to navigating the extensive body of scholarly work. By grouping research papers into related categories, these tags make it easier for researchers to locate studies in their particular fields of interest (Figure 9, Figure 10).

Fig. 9 illustrates the database structure, which not only makes the search process more efficient but also aids in revealing potential links among various papers and areas of research that may not be immediately obvious.

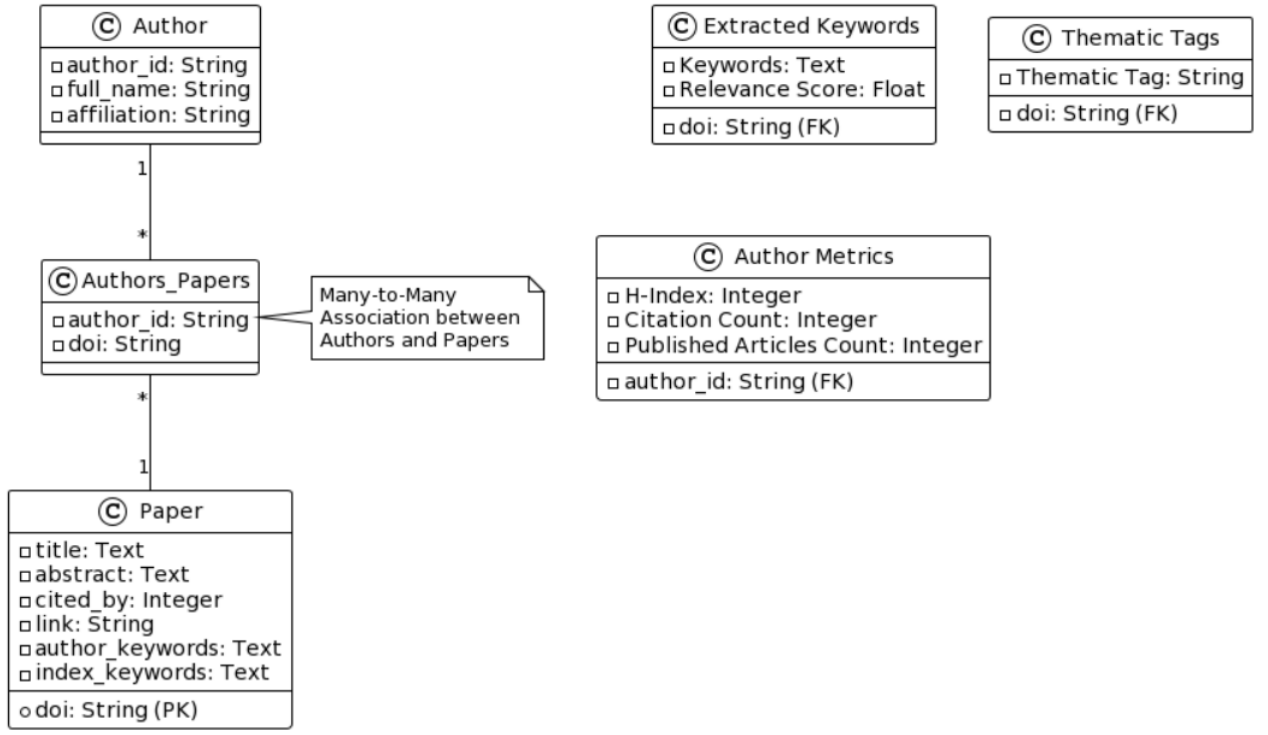

Figure 9. Database schema

Fig. 10 presents the microservices framework, underscoring how the thematic tags, created through the K-means clustering algorithm on this research platform, establish a groundwork for broader, future academic inquiries. By sorting a large number of scholarly articles into defined thematic clusters, these tags not only facilitate easier access for researchers but also organize the data in a way that's conducive to comprehensive longitudinal and overarching research studies.

This not only streamlines the search process but also helps in uncovering connections between different papers and research topics that might not be immediately apparent.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

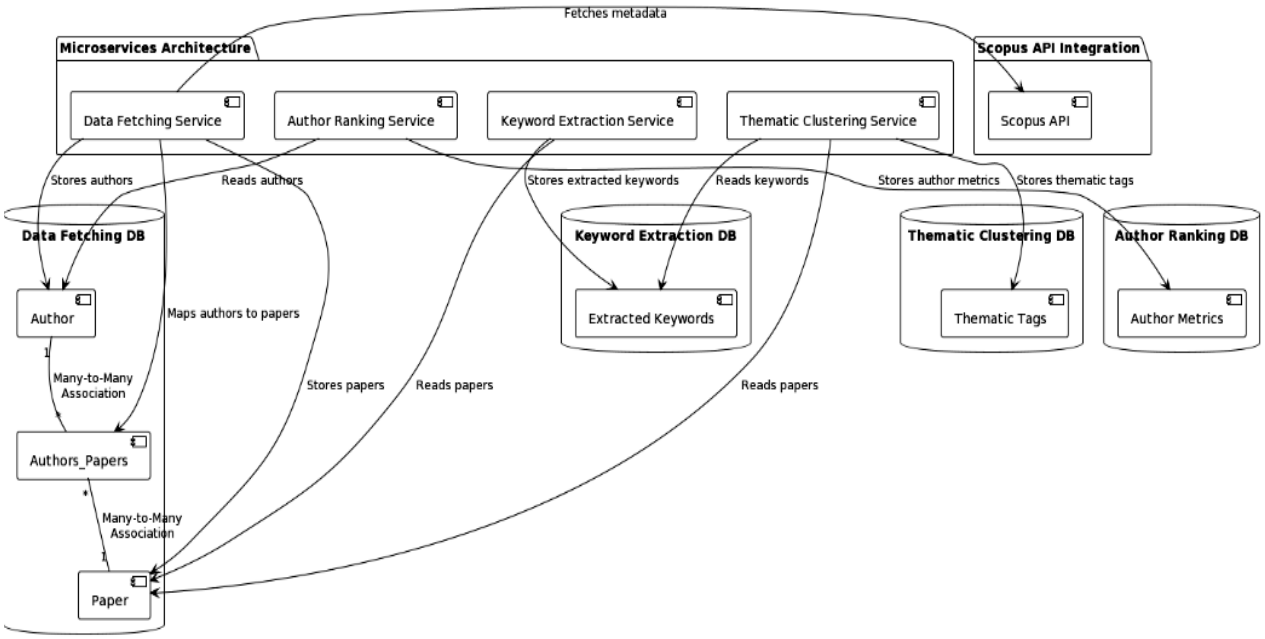

Figure 10. Microservices architecture

Conclusion

This research advances the intersection of information technology and academic study by developing an IT system that employs text mining and clustering techniques to organize a sizable compilation of research abstracts from Kazakhstani institutions. Utilizing TF-IDF vectorization coupled with K-Means clustering, the system identifies predominant research themes, demonstrating the extensive scope of scholarly work in the area. The process effectively segments literature into distinct thematic clusters, simplifying literature discovery and promoting scholarly collaboration. Focusing specifically on the Computer Science field, the system generates in-depth insights for targeted scholarly exploration. A key feature of this IT system is its recommendation engine, which, within each thematic cluster, suggests potential collaborations among authors based on the thematic alignment of their work. This not only enhances the accessibility of relevant studies but also fosters a networked academic community, potentially accelerating the advancement of collective knowledge within and across disciplines.

Acknowledgement

This paper was written in the framework of the state order to implement the science program according to the budget program 217 “Development of Science”, IRN No. AP19678627 with the topic: “Development of the information technology for the formation of multi-university scientific and educational communities based on the scientometric analysis theory”.

References

[1] Al-Obaydy, W.I., Hashim, H.A., Najm, Y.A., & Jalal, A.A. (2022). Document classification using term frequency-inverse document frequency and K-means clustering. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(3), 1517-1524.

[2] Shetty, K., & Kallimani, J.S. (2017, December). Automatic extractive text summarization using K-means clustering. *2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT),* 881-890.

DOI: 10.37943/17ALVR8114

[3] Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., Kuzka, O., & Terentyev, O. (2017). Evaluation methods of the results of scientific research activity of scientists based on the analysis of publication citations. *Vostochno-Evropejskij zhurnal peredovyh tehnologij*, 3 (2), 4-10.
[4] Alsmadi, I., & Alhami, I. (2015). Clustering and classification of email contents. *Journal of King Saud University-Computer and Information Sciences*, 27(1), 46-57.
[5] Rejito, J., Atthariq, A., & Abdullah, A. S. (2021). Application of text mining employing k-means algorithms for clustering tweets of Tokopedia. *Journal of Physics: Conference Series*, 1722 (1), 012019.
[6] Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168-189.
[7] Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, *25*(1), 114-146.
[8] Alhawarat, M., & Hegazi, M. (2018). Revisiting k-means and topic modeling, a comparison study to cluster arabic documents. *IEEE Access*, *6*, 42740-42749.
[9] Oti, E. U., Olusola, M.O., Eze, F.C., & Enogwe, S.U. (2021). Comprehensive review of K-Means clustering algorithms. *International Journal of Advances in Scientific Research and Engineering*, 7(8), 64.
[10] Vijayarani, S., Ilamathi, M.J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.
[11][11] Aubaidan B., Mohd M., Albared M. (2014). Comparative study of k-means and k-means++ clustering algorithms on crime domain. *Journal of Computer Science,* 10 (7), 1197-1206.
[12] Tabassum, A., & Patil, R.R. (2020). A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, 7(06), 4864-4867.
[13] Kadhim, A.I., Cheah, Y.N., & Ahamed, N.H. (2014, December). Text document preprocessing and dimension reduction techniques for text document clustering. *2014 4th international conference on artificial intelligence with applications in engineering and technology*, 69-73.
[14] Al-Anazi, S., AlMahmoud, H., & Al-Turaiki, I. (2016). Finding similar documents using different clustering techniques. *Procedia Computer Science*, 82, 28-34.
[15] Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF-IDF approach. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 61-66.
[16] Arora, P., Deepali Dr., Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78, 507-512.
[17] Zhou, S., Xu, X., Liu, Y., Chang, R., & Xiao, Y. (2019). Text similarity measurement of semantic cognition based on word vector distance decentralization with clustering analysis. *IEEE Access*, 7, 107247-107258.
[18] Singh, A.K., & Shashi, M. (2019). Vectorization of text documents for identifying unifiable news articles. *International Journal of Advanced Computer Science and Applications*, 10(7).
[19] Naeem, S., & Wumaier, A. (2018). Study and implementing K-mean clustering algorithm on English text and techniques to find the optimal value of K. *International Journal of Computer Applications*, 182(31), 7-14.
[20] Kim, S.W., & Gil, J.M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*, 9, 1-21.