HUNGARIAN UNIVERSITY OF AGRICULTURE AND LIFE SCIENCES

PROJECT MANAGEMENT PROCESS MODEL FOR DIGITAL ERA BASED ON BIG DATA ANALYSIS

The Thesis of Doctoral (PhD) dissertation

By:

Mehrzad AbdiKhalife

Gödöllő, Hungary

2024

Hungarian University of Agriculture and Life Sciences

Name of Doctoral School:
Doctoral School of Economic and Regional Sciences

Discipline:
Management and Business Administration

Head of Doctoral School:	Prof. Dr. Zoltán Bujdosó PhD, Full Professor Hungarian University of Agriculture and Life Sciences
Supervisor:	Prof. Dr. Anna Dunay PhD, Full Professor John von Neumann University

……………………………… ………………………………
Approval of Head of Doctoral School Approval of Supervisor

Table of contents

1. INTRODUCTION

1.1. Importance of the topic

Scientific exploration is fundamentally driven by the needs that arise from theoretical gaps and practical challenges, and the motivations to address these needs. This research was initiated not by chance, but by the desire to fill specific gaps and meet certain requirements in the field. The motivation for this journey was fueled by the aspiration to provide meaningful solutions to these challenges and a passion for deepening the understanding of the subject. The researcher's previous role as a project manager provided a firsthand experience of the complexities of project management, highlighting the paradox of choice in decision-making tools.

This experience led to a desire to simplify this process in their academic work. The role of a project manager is central to this, as each decision they make has a domino effect, impacting various aspects of a project's lifecycle. Despite the wide range of available standards, software, and methodologies, the ultimate goal is to enhance the accuracy of decision-making and streamline processes. This research aims to dissect, understand, and potentially simplify these decision-making processes that are so crucial to effective project management. Classifying knowledge helps in organizing the vast amount of information into understandable categories, enabling researchers to comprehend the structure of a domain, identify gaps, and spot overlaps. Monitoring the most discussed articles provides insights into current trends and areas of interest in the scientific community, guiding researchers towards significant areas. Scientometrics analysis, which measures and analyzes scientific literature, evaluates the impact of publications and researchers and provides insights into the growth and focus of various scientific fields. A related concept, bibliometrics, focuses on evaluating publication patterns and citation analyses, offering insights into the influence of scholarly works.

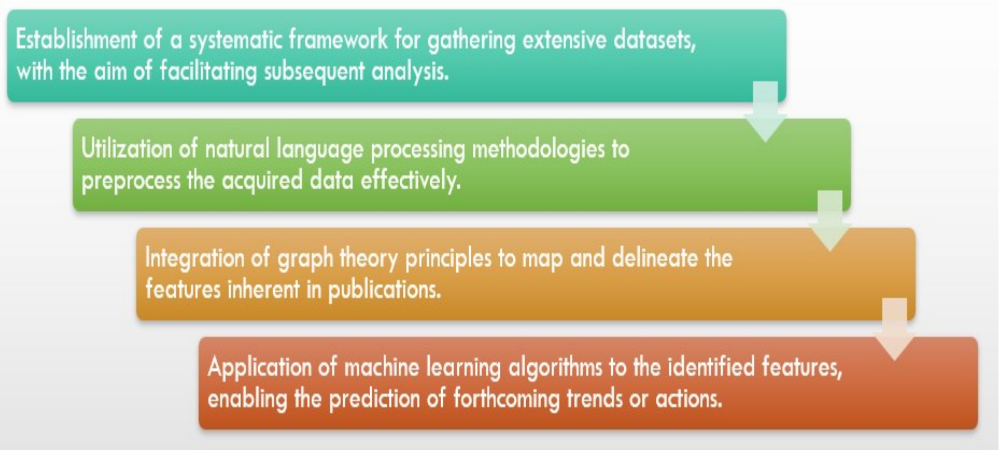
Network analysis of publications involves creating and analyzing networks based on citations, collaborations, and co-authorships. Visualizing these networks allows for understanding the relationships between researchers and institutions, highlighting the collaborative nature of science, and identifying influential nodes in the network. These tools, when used wisely, serve as invaluable aids for researchers to navigate the complex pathways of knowledge. Understanding the current landscape is crucial, but it only provides half the picture. Recognizing patterns can offer glimpses into future trends, despite the elusive nature of precise prediction. Therefore, the focus evolved to shed light on potential pathways for both theoretical and practical project managers.

1.2. Problem statement

In the fast-paced digital era, project management faces numerous complexities. Agile project management, renowned for its adaptability, has become a favored approach to navigate this dynamic environment. However, integrating big data analysis into agile project management is a complex and underexplored task. This research aims to develop a Project Management Process Model that effectively incorporates big data analysis. This model is designed to serve as a guide through the complex maze of evolving knowledge and methodologies, helping to make sense of the vast amount of information characteristic of the digital age.

The proposed model will include text mining and text analysis, facilitating the discovery of trends and extraction of specific details from large volumes of text. It will also incorporate the classification of knowledge, aiding in organizing the overwhelming amount of information into coherent categories. This process will be complemented by monitoring the most accessed or discussed articles, providing insights into current trends and areas of interest. The model will also integrate scientometrics analysis and network analysis of publications, offering insights into the growth and focus of different scientific fields and illuminating the collaborative nature of science. The research will investigate the impact of the proposed model on project outcomes and explore the factors influencing its successful implementation. The findings could contribute to project management by providing a comprehensive framework for effective decision-making, potentially enhancing project outcomes and efficiency in the digital era.

The framework of the research steps depicted in Fig. 1 offers a clearer perspective.



Establishment of a systematic framework for gathering extensive datasets, with the aim of facilitating subsequent analysis.

Utilization of natural language processing methodologies to preprocess the acquired data effectively.

Integration of graph theory principles to map and delineate the features inherent in publications.

Application of machine learning algorithms to the identified features, enabling the prediction of forthcoming trends or actions.

Figure 1. The framework of data collection, preprocessing, feature mapping, and predictive analysis (own research)

1.3. Research questions/ research objectives

1.3.1. Research Questions

1. Decoding and Anticipating Trends: How can I leverage the combined power of network analysis and machine learning to decode and anticipate trends in word analysis?

2. Visualizing Knowledge: What insights can be derived from visualizing connections, patterns, and clusters in the current state of knowledge?

3. Data Extraction and Analysis: How can web scraping techniques be used to extract articles from openly accessible websites, and what does this data reveal about global discourse and scholarly communication?

4. Temporal Analysis: What shifts in discourse, emerging topics, and fading trends can be identified by comparing articles published in different quarters?

5. Predictive Analysis: How can a temporal lens, when combined with predictive machine learning models, help me understand the present narrative landscape and make informed projections about its future evolution?

1.3.2. Research Objectives

1. Harnessing Tools: My first goal is to use network analysis and machine learning to decode and anticipate linguistic trends.

2. Knowledge Landscape: My second goal is to use network analysis to map and understand the current state of knowledge.

3. Data Sourcing: My third goal is to source and analyze data from the web to understand global discourse.

4. Temporal Segmentation: My fourth goal is to segment data by publication quarters to track discourse evolution.

5. Discourse Shifts: My fifth goal is to track shifts in discourse, detect emerging topics, and identify fading trends.

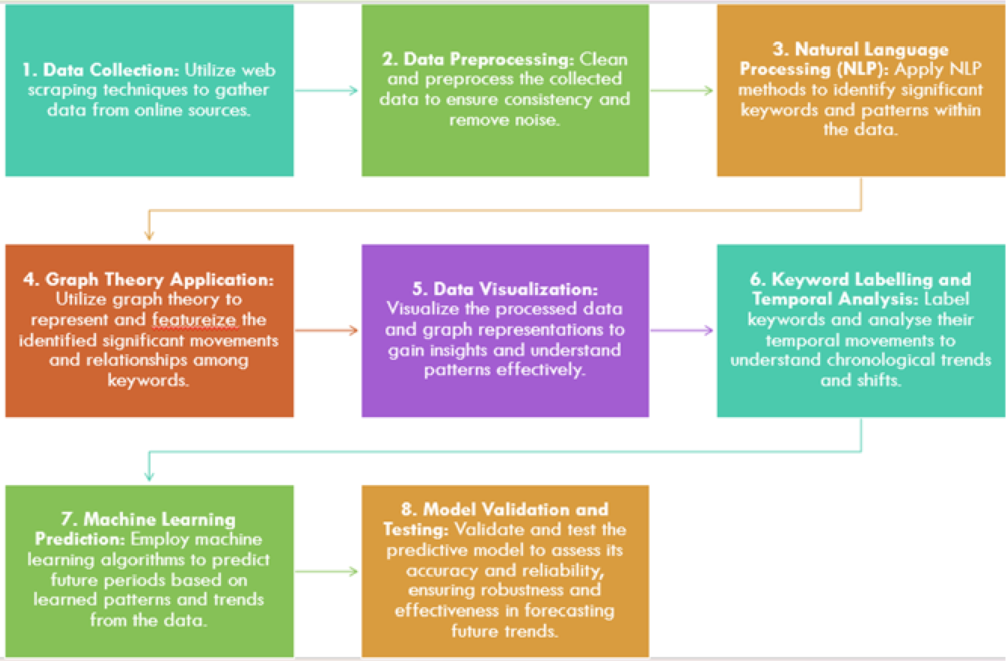
6. Predictive Analysis: My final goal is to use predictive machine learning models to understand the current narrative landscape and predict its future evolution.

1.4. Conceptual model / research model

The proposed model will use network analysis as a key tool to understand the current state of knowledge in the field, visualizing connections, patterns, and clusters to identify relationships, prominent themes, and emerging nodes of importance. Data for the analysis will be sourced from the World Wide Web using sophisticated web scraping techniques to extract articles, providing a comprehensive dataset reflecting global discourse in project management. The data will be segmented based on publication quarters to track temporal changes in discourse, enabling the detection of shifts, emerging topics, and fading trends. The model will also incorporate machine learning prediction models trained on the segmented data to predict future trends, providing timely insights into project management's future trends.

The model will integrate scientometrics analysis and bibliometrics to measure and analyze scientific literature, evaluate publication patterns, citation analyses, and relationships between authors, institutions, and countries. This will provide insights into the influence and reach of scholarly works, revealing collaboration and citation patterns that reflect the scientific community's structure and dynamics. Network analysis of publications will also be incorporated, constructing and analyzing networks based on citations, collaborations, and co-authorships to illuminate the collaborative nature of science and identify influential nodes in the network. For this purpose, 12793 publications were identified in the core collection of the Web of Science (CCWoS), and bibliometrics analysis was used to discover information among publications. The application of big data analysis in project management was also represented, with 92 publications selected from 129 search results using the keywords 'project management' and 'big data'.

The framework of the research steps, as illustrated in Fig. 2, provides a clearer understanding of the overall process.



1. Data Collection: Utilize web scraping techniques to gather data from online sources.

2. Data Preprocessing: Clean and preprocess the collected data to ensure consistency and remove noise.

3. Natural Language Processing (NLP): Apply NLP methods to identify significant keywords and patterns within the data.

4. Graph Theory Application: Utilize graph theory to represent and featureize the identified significant movements and relationships among keywords.

5. Data Visualization: Visualize the processed data and graph representations to gain insights and understand patterns effectively.

6. Keyword Labelling and Temporal Analysis: Label keywords and analyse their temporal movements to understand chronological trends and shifts.

7. Machine Learning Prediction: Employ machine learning algorithms to predict future periods based on learned patterns and trends from the data.

8. Model Validation and Testing: Validate and test the predictive model to assess its accuracy and reliability, ensuring robustness and effectiveness in forecasting future trends.

Figure 2. Conceptual model of the data analysis and prediction framework employed in this thesis (own research)

1.5. Thomas Kuhn science philosophy

Thomas Kuhn's philosophy of science argues that scientific advancement is not characterized by a linear and continuous progression but rather by periodic "paradigm shifts." These shifts represent fundamental changes in the basic concepts and experimental practices of a scientific discipline, often triggered by an accumulation of anomalies that the current paradigm cannot account for. During periods of "normal science," the research conducted by the scientific community is based on an established paradigm—a framework that includes the beliefs, techniques, and standards shared by the community. This period is marked by cumulative progress within the confines of the existing paradigm. In contrast, "revolutionary science" emerges when the current paradigm is no longer sufficient to explain observed phenomena, leading to a scientific revolution and the establishment of a new paradigm.

Kuhn's seminal work, "The Structure of Scientific Revolutions," emphasizes that the notion of scientific truth at any given time is shaped not only by objective criteria but also by the prevailing consensus within the scientific community. This introduces a social dimension to the concept of scientific truth, suggesting that it is, to some extent, constructed by the community's subjective agreement rather than by empirical evidence alone.

Moreover, Kuhn's influence extends beyond the philosophy of science. The terms he introduced, such as "paradigm shift" and "normal science," have permeated academic and public discourse, illustrating the broad impact of his ideas. His philosophy underscores the importance of historical and social contexts in understanding the evolution of scientific knowledge, thereby challenging the simplistic view of science as a straightforward march towards truth.

2. MATERIALS AND METHODS

2.1. Data source

Collecting and filtering relevant data is a critical step in the research process. After identifying potential sources of information, it is important to filter the data based on specific keywords related to the research question. This helps to ensure that the data collected is relevant and useful for the study.

Data collection tools and steps are summarized in Fig. 3.

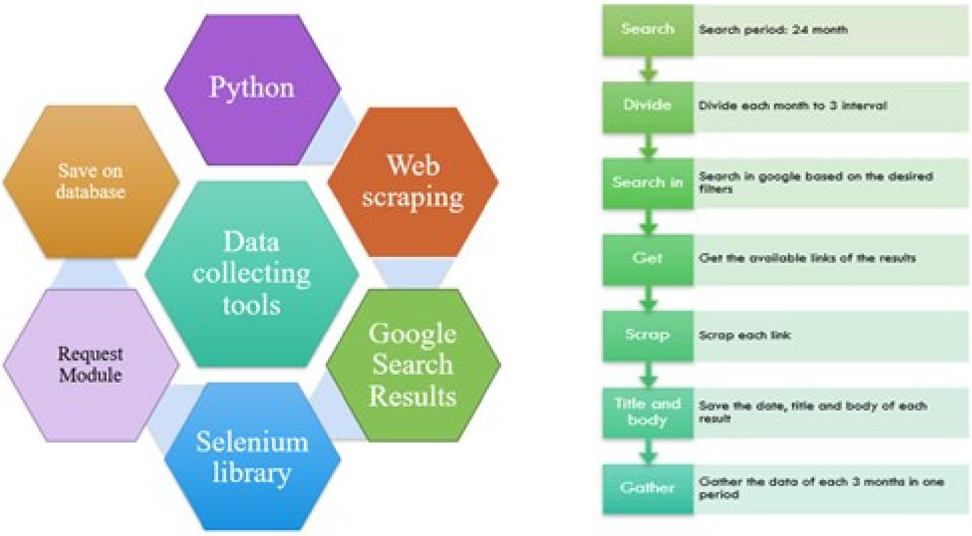

Python
Web scraping
Save on database
Data collecting tools
Google Search Results
Request Module
Selenium library
Search
Search period: 24 month
Divide
Divide each month to 3 interval
Search in
Search in google based on the desired filters
Get
Get the available links of the results
Scrap
Scrap each link
Title and body
Save the date, title and body of each result
Gather
Gather the data of each 3 months in one period

Figure 3. Data collection tools and steps (own research)

2.1.1. Academic source

Academic sources can be defined as a body of work produced by experts in a particular field of study. These experts are often scholars, researchers, or academics who have conducted studies and experiments to test hypotheses related to their area of expertise. The work is then published in peer-reviewed journals or academic books, which are rigorously reviewed by other experts in the same field. The process ensures that the research meets high standards of quality and accuracy, making it a valuable resource for other researchers.

These sources provide several advantages that make them valuable for research purposes. Mainly, academic sources are reliable because they are based on rigorous research methods and adhere to scholarly standards. This ensures that the information presented in these sources is accurate and trustworthy.

Google Scholar, Web of Science, and Scopus are three of the most widely used academic search engines in the world. Each of these platforms has its unique features and benefits that make them popular among researchers, academics, and students. It is the largest scientific database, covering a broad range of disciplines and sources. It provides easy access to scholarly literature, including articles, books, theses, and conference papers.

2.1.2. Non-academic source

Traditionally, research has focused on using academic sources such as journals, books, and articles. However, non-academic sources can also provide valuable information that can enrich a study. Non-academic sources can provide perspectives that are not found in academic literature, and they can help researchers to explore topics from a variety of angles. Non-academic sources refer to sources such as newspapers, magazines, blogs, and social media platforms that are not peer-reviewed or written by scholars. These sources include books, magazines, newspapers, government reports, personal interviews, historical documents, blogs, podcasts, videos, and social media posts. While these sources may not have undergone the same rigorous peer-review process as academic sources, they can still provide valuable insights into a research topic. These sources also provide several advantages that make them valuable for research purposes.

Another consideration is data privacy regulations. The General Data Protection Regulation (GDPR) governs data protection and privacy for citizens of the European Union (EU). The regulation outlines strict requirements for collecting, storing, and processing personal data, including names, email addresses, and other identifying information. Organizations must obtain consent from individuals before collecting their personal data and ensure that the data is protected from unauthorized access and use.

2.2. Application of graph theory

Creating networks based on co-occurrence or frequency of keyword occurrence can help us identify important concepts and relationships in the data. For example, if there are frequent co-occurrences between the keyword's "risk" and "project management," it may suggest that risk management is an important consideration in project management. Similarly, if certain keywords have high node sizes or edge weights, it may indicate that these concepts are particularly relevant or significant in the dataset.

Once the networks have been created, I can visualize and analyze the data to gain insights into project management practices and strategies. For example, I can identify clusters of keywords that are closely related and examine their relationships to other clusters. I can also measure the centrality of different nodes or groups of nodes, which can provide information about which concepts or themes are most central to the dataset. By analyzing the networks, I can gain a deeper understanding of the relationships between different concepts and how I may impact project management outcomes.

In addition to creating networks based on keyword co-occurrence or frequency of occurrence, there are other approaches that I can use to create networks in project management. For example, I can use semantic analysis to identify relationships between keywords based on their meanings, rather than their co-occurrence or frequency. I can also use network clustering algorithms to identify subgroups of nodes that are highly connected within the network.

Network analysis is a powerful tool for understanding the structure and dynamics of complex systems. Once the networks are created and drawn, I can use various network analysis algorithms to gain insights into the underlying patterns and relationships in the data.

2.3. Using NLP in graph theory

Once the network is created, researchers can apply various network analysis algorithms to gain insights into the structure and dynamics of the text. For example, they might use clustering algorithms to group words or entities together based on their similarity. Alternatively, they might use centrality algorithms to identify the most important words or entities within the text.

One way that NLP can benefit graph theory and network science is by providing more granular data for analysis. For example, researchers might use sentiment analysis to assign positive or negative values to each node in the network. This can help identify which entities or concepts are associated with positive or negative sentiment and how this sentiment changes over time.

Another benefit of using NLP in graph theory and network science is that it can enable more sophisticated analysis of text data. For example, researchers might use topic modeling to identify the key topics within a text and map out the relationships between them. They might then create a network where each node represents a topic, and the edges represent the relationships between them. This approach can provide valuable insights into the underlying structure of the text and how different topics relate to each other.

NLP can also be used to identify key phrases or keywords within a text that are relevant to a particular research question. By extracting these phrases or keywords and mapping out the connections between them, researchers can gain insights into the most important concepts or themes within the text.

One potential application of using NLP in graph theory and network science is in the analysis of social media data. By creating networks based on social media interactions, researchers can gain insights into the social dynamics that drive behavior on these platforms. For example, they might analyze the connections between users and the topics they discuss to identify areas of agreement or disagreement. Research model is summarized by Fig. 4.

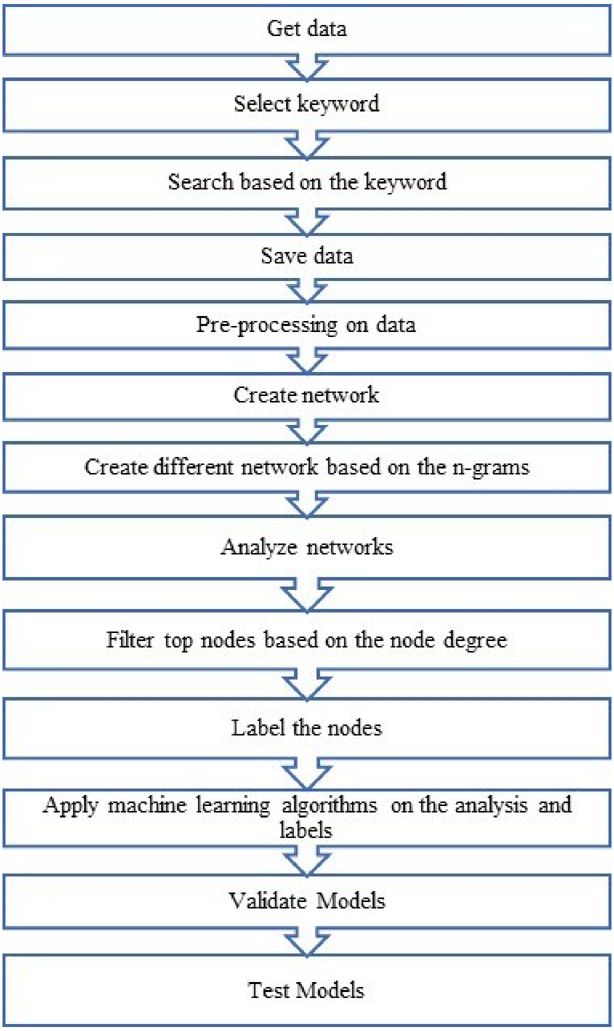

Get data
Select keyword
Search based on the keyword
Save data
Pre-processing on data
Create network
Create different network based on the n-grams
Analyze networks
Filter top nodes based on the node degree
Label the nodes
Apply machine learning algorithms on the analysis and labels
Validate Models
Test Models

Figure 4. The machine learning workflow: From collection to analysis (own research)

3. RESULTS AND THEIR DISCUSSION

3.1. Data collection

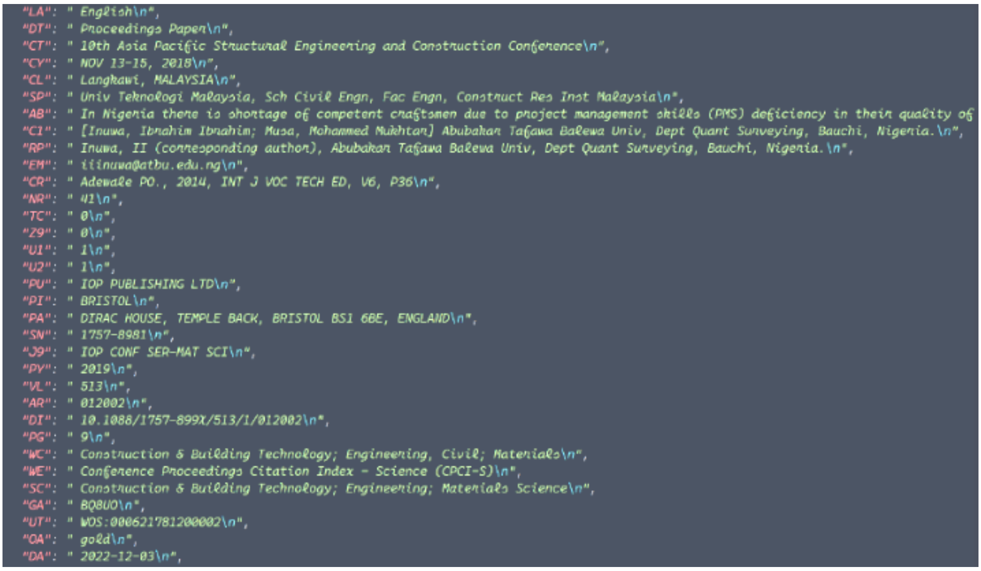
Academic data collection involves the gathering of information related to academic disciplines, such as scientific experiments, survey responses, academic publications, and educational records. These data sources serve as the foundation for generating new knowledge, testing hypotheses, and validating theories. Researchers often rely on academic data to make significant contributions to their respective fields.

Non-academic data, on the other hand, extends beyond the traditional boundaries of academia. It encompasses data from sources such as social media, online reviews, customer feedback, and real-world observations. Non-academic data collection is particularly valuable for interdisciplinary research and for understanding the practical implications of academic findings in real-world scenarios. For academic data, tried to get the scientific articles from journals. Each paper consists of various metadata, like title, body, keywords, journal name, and etc. These data are the inputs for the next steps. All academic data stored in a single Json file, which each element in this file is the information about specific record.

As illustrated in the accompanying image, the dataset under consideration presents information pertaining to individual records. It is noteworthy that each record within this dataset offers a consistent set of data elements. While it is important to acknowledge that not every record within the dataset contains every available data element, a substantial portion of these records share commonalities, consistently encompassing the same set of information. To elucidate, some records may lack temporal data, yet the majority of records consistently feature both a title and a body section.

Within the context of academic data analysis, the title and body components of each record are subjected to meticulous filtering and are subsequently utilized as primary inputs for in-depth scrutiny and further analytical exploration. This choice of utilizing the title and body components stems from their intrinsic significance and recurring relevance within the dataset. These two components, namely the title and body, emerge as indispensable reservoirs of valuable insights, capable of furnishing users with a wealth of pertinent information. This intrinsic importance underscores the rationale for their selection as principal data sources for subsequent analysis and research endeavors.

The rationale behind this selection is grounded in the recognition that titles typically encapsulate succinct yet informative descriptors, serving as a brief but comprehensive representation of the record's content. On the other hand, the body section augments this by providing a more detailed exposition of the record's content, elucidating its nuances and intricacies. These dual components, when employed in tandem, synergistically enhance the analytical capabilities of researchers, facilitating the extraction of rich and meaningful insights from the dataset. Consequently, the combination of title and body data serves as a potent foundation for rigorous academic inquiry, offering researchers a robust starting point for their investigative endeavors and the potential to unlock a plethora of valuable information. A preview of academic data is illustrated by Fig. 5.



```
"LA": " English\n",
"DT": " Proceedings Paper\n",
"CT": " 10th Asia Pacific Structural Engineering and Construction Conference\n",
"CY": " NOV 13-15, 2018\n",
"CL": " Langkawi, MALAYSIA\n",
"SP": " Univ Teknologi Malaysia, Sch Civil Engn, Fac Engn, Construct Res Inst Malaysia\n",
"AB": " In Nigeria there is shortage of competent craftsmen due to project management skills (PMS) deficiency in their quality of
"C1": " [Inuwa, Ibrahim Ibrahim; Musa, Mohammed Mukhtar] Abubakar Tafawa Balewa Univ, Dept Quant Surveying, Bauchi, Nigeria.\n",
"RP": " Inuwa, II (corresponding author), Abubakar Tafawa Balewa Univ, Dept Quant Surveying, Bauchi, Nigeria.\n",
"EM": " iiinuwa@atbu.edu.ng\n",
"CR": " Adewale PO., 2014, INT J VOC TECH ED, V6, P36\n",
"NR": " 41\n",
"TC": " 0\n",
"Z9": " 0\n",
"U1": " 1\n",
"U2": " 1\n",
"PU": " IOP PUBLISHING LTD\n",
"PI": " BRISTOL\n",
"PA": " DIRAC HOUSE, TEMPLE BACK, BRISTOL BS1 6BE, ENGLAND\n",
"SN": " 1757-8981\n",
"J9": " IOP CONF SER-MAT SCI\n",
"PY": " 2019\n",
"VL": " 513\n",
"AR": " 012002\n",
"DI": " 10.1088/1757-899X/513/1/012002\n",
"PG": " 9\n",
"WC": " Construction & Building Technology; Engineering, Civil; Materials\n",
"WE": " Conference Proceedings Citation Index - Science (CPCI-S)\n",
"SC": " Construction & Building Technology; Engineering; Materials Science\n",
"GA": " BQ8UO\n",
"UT": " WOS:000621781200002\n",
"OA": " gold\n",
"DA": " 2022-12-03\n",
```

Figure 5. A Preview of academic data (own research)

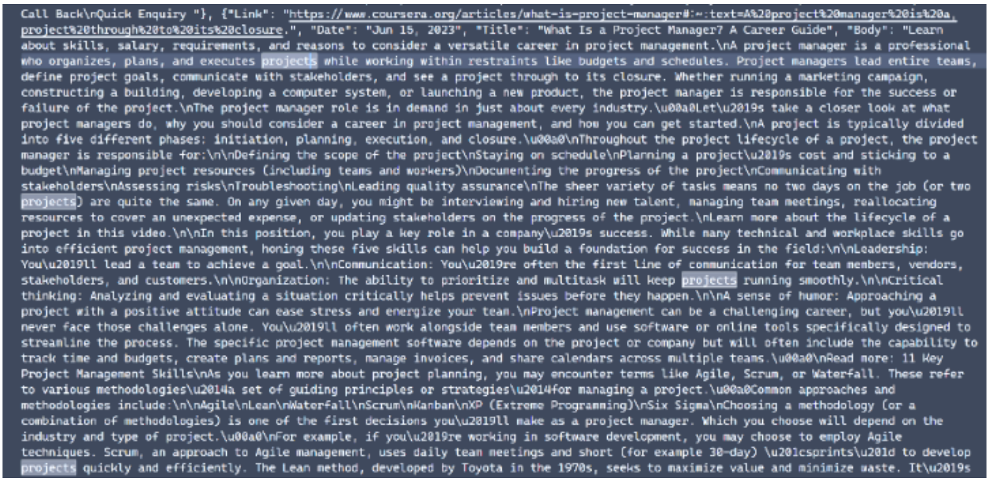
In the realm of data analysis, academic datasets exhibit a discernibly more structured and regular format in comparison to their non-academic counterparts. This characteristic owes its existence to a confluence of factors, primarily stemming from the meticulous curation and filtering mechanisms imposed by academic journals and their associated databases. This attribute of structuredness, as discussed in subsequent sections, bestows upon academic data a distinct advantage in terms of ease of processing, setting it apart from the relatively less structured landscape of non-academic data. While it is essential to acknowledge that the disparity between these two categories of data may not be seismic in nature, it undoubtedly contributes significantly to the enhanced manageability and utility of academic datasets within the domain of data analysis.

The structured nature of academic data, as previously alluded to, finds its origins in the rigorous editorial and review processes intrinsic to academic journals. These processes serve as a crucible for ensuring the quality, consistency, and organization of the information contained within academic publications. Additionally, academic databases, which often serve as repositories for scholarly works, further reinforce this structured paradigm by imposing standardized formatting and indexing practices. Consequently, academic data is imbued with a regularity and predictability that greatly simplifies its manipulation and analysis.

In contrast, non-academic data sources typically lack the stringent curation and filtering mechanisms characteristic of academic data. This divergence can lead to a higher degree of heterogeneity, unpredictability, and noise within non-academic datasets. The absence of a standardized framework for organizing and presenting information can present formidable challenges when attempting to process and extract meaningful insights from such data.

It is important to underscore that while the disparity in structure between academic and non-academic data may not always result in a profound divergence in terms of analytical outcomes, it undeniably expedites the workflow associated with academic data analysis. This increased ease of use can translate into efficiency gains, facilitating more expedient research and knowledge generation processes within academic contexts.

In Fig. 6, a record of non-academic data has been shown. As obvious from the picture each result consists of four elements (i.e., Link, Date, Title, and Body). Also, it can be mentioned these data is not as clean as academic data.



```
Call Back\nQuick Enquiry "}, {"Link": "https://www.coursera.org/articles/what-is-project-manager#:~:text=A%20project%20manager%20is%20a,
project%20through%20to%20its%20closure.", "Date": "Jun 15, 2023", "Title": "What Is a Project Manager? A Career Guide", "Body": "Learn
about skills, salary, requirements, and reasons to consider a versatile career in project management.\nA project manager is a professional
who organizes, plans, and executes projects while working within restraints like budgets and schedules. Project managers lead entire teams,
define project goals, communicate with stakeholders, and see a project through to its closure. Whether running a marketing campaign,
constructing a building, developing a computer system, or launching a new product, the project manager is responsible for the success or
failure of the project.\nThe project manager role is in demand in just about every industry.\u00a0Let\u2019s take a closer look at what
project managers do, why you should consider a career in project management, and how you can get started.\nA project is typically divided
into five different phases: initiation, planning, execution, and closure.\u00a0\nThroughout the project lifecycle of a project, the project
manager is responsible for:\n\nDefining the scope of the project\nStaying on schedule\nPlanning a project\u2019s cost and sticking to a
budget\nManaging project resources (including teams and workers)\nDocumenting the progress of the project\nCommunicating with
stakeholders\nAssessing risks\nTroubleshooting\nLeading quality assurance\nThe sheer variety of tasks means no two days on the job (or two
projects) are quite the same. On any given day, you might be interviewing and hiring new talent, managing team meetings, reallocating
resources to cover an unexpected expense, or updating stakeholders on the progress of the project.\nLearn more about the lifecycle of a
project in this video.\n\nIn this position, you play a key role in a company\u2019s success. While many technical and workplace skills go
into efficient project management, honing these five skills can help you build a foundation for success in the field:\n\nLeadership:
You\u2019ll lead a team to achieve a goal.\n\nCommunication: You\u2019re often the first line of communication for team members, vendors,
stakeholders, and customers.\n\nOrganization: The ability to prioritize and multitask will keep projects running smoothly.\n\nCritical
thinking: Analyzing and evaluating a situation critically helps prevent issues before they happen.\n\nA sense of humor: Approaching a
project with a positive attitude can ease stress and energize your team.\nProject management can be a challenging career, but you\u2019ll
never face those challenges alone. You\u2019ll often work alongside team members and use software or online tools specifically designed to
streamline the process. The specific project management software depends on the project or company but will often include the capability to
track time and budgets, create plans and reports, manage invoices, and share calendars across multiple teams.\u00a0\nRead more: 11 Key
Project Management Skills\nAs you learn more about project planning, you may encounter terms like Agile, Scrum, or Waterfall. These refer
to various methodologies\u2014a set of guiding principles or strategies\u2014for managing a project.\u00a0Common approaches and
methodologies include:\n\nAgile\nLean\nWaterfall\nScrum\nKanban\nXP (Extreme Programming)\nSix Sigma\nChoosing a methodology (or a
combination of methodologies) is one of the first decisions you\u2019ll make as a project manager. Which you choose will depend on the
industry and type of project.\u00a0\nFor example, if you\u2019re working in software development, you may choose to employ Agile
techniques. Scrum, an approach to Agile management, uses daily team meetings and short (for example 30-day) \u201csprints\u201d to develop
projects quickly and efficiently. The Lean method, developed by Toyota in the 1970s, seeks to maximize value and minimize waste. It\u2019s
```

Figure 6. A Preview of non-academic data (own research)

The methodological approach for this non-academic data gathering operation involved a series of meticulously designed steps, each contributing to the overall objective of accumulating a meaningful dataset on project management.

The initial step revolved around the careful selection of a keyword that would serve as the foundation of the search process. 'Project management' was chosen due to its broad relevance and significance in various professional domains. This keyword was expected to yield diverse sources of information, encompassing the latest trends, practices, and discussions within the field.

The Google search engine was identified as the optimal tool for data collection. Google's unparalleled indexing capabilities and user-friendly interface made it the ideal choice for sourcing web-based data. Furthermore, its advanced search features allowed us to fine-tune results, which proved pivotal in achieving the research's temporal objectives.

To ensure the relevance and timeliness of the acquired data, I implemented a specific date range for the search results. This date range was set to span from January 1, 2023, to July 31, 2023, reflecting the desire to capture the most recent developments and discussions in project management. By limiting the scope to this seven-month period, I sought to balance the need for current data with the requirement for a substantial dataset for analysis.

Each webpage discovered during the search process was recognized as a repository of valuable information. To harness this data effectively, I systematically extracted key elements from each webpage, including the title, body text, and publication date. These elements were considered fundamental for constructing a dataset that would support comprehensive analysis and insights.

Despite the methodical approach, several challenges emerged during the data gathering process, each requiring careful consideration and resolution:

An initial challenge lay in the absence of publication dates on certain webpages. Without this temporal information, it was difficult to assess the relevance of the content. To address this issue, I relied on alternative strategies, such as searching for publication date metadata within the webpage's source code or cross-referencing information with other credible sources.

This non-academic data gathering venture resulted in the acquisition of a substantial dataset encompassing a diverse array of web-based resources related to project management. The data, comprising titles, body text, and, where available, publication dates, holds the potential to serve as a valuable resource for future analysis, research, and insights into the evolving landscape of project management practices. Furthermore, the challenges encountered during this endeavor have provided valuable lessons and insights into the complexities of web-based data collection, which can inform future non-academic research efforts in the digital era.

Historical datasets play a pivotal role in informing research, decision-making, and understanding trends over time. This academic text delves into the complexities encountered in the pursuit of constructing a historical dataset for non-academic purposes. The objective was to create 10 to 12 distinct time periods, each containing at least 1000 records, focusing on the field of project management. This endeavor required data collection from web sources and involved a multi-faceted approach, beset with challenges and innovative solutions.

One of the initial and persistent challenges was the presence of webpage restrictions designed to deter data scraping. Some websites employed mechanisms such as CAPTCHA challenges and robots.txt files, making it difficult to extract the desired data elements, including titles and body text.

These restrictions necessitated the exploration of workarounds and alternative data sources to compensate for the missing data.

The scale of search results generated by the Google search engine posed a significant challenge. Typically, the search results included around 30 pages, with each page containing approximately 10 webpages. This abundance of data required a structured approach to ensure comprehensive coverage. Consequently, the decision was made to divide the search results into three stages, with each month further divided into three date ranges. While this strategy aimed to maximize data collection, it introduced its own set of challenges.

Despite meticulous planning and dividing the search results into date ranges, an unexpected challenge emerged. Some months did not yield the desired volume of data necessary to meet the 1000-record threshold. This disparity raised questions about the completeness and representativeness of the dataset for each month.

To address webpage restrictions, the research team employed a multi-pronged approach. First, CAPTCHA-solving solutions were explored to bypass challenges encountered during data scraping. Second, alternative data sources were identified to compensate for inaccessible webpages. These included publicly available datasets and archived content from websites with more permissive data policies. The integration of these supplementary sources helped mitigate the impact of webpage restrictions.

To cope with the overwhelming volume of search results, a structured three-stage data collection process was adopted. Each month was divided into three distinct date ranges, allowing for more granular searches. This approach enabled the research team to obtain a more comprehensive dataset while managing the scale of information. Despite this, the challenge of ensuring an equitable distribution of data across all months persisted.

In response to the issue of insufficient data for some months, a strategic decision was made to aggregate records from each consecutive three-month period into a single period. This consolidation yielded ten distinct periods, each spanning three months. By grouping data in this manner, the research team aimed to achieve the desired minimum of 1000 records per period. While this approach helped address the challenge of data scarcity in individual months, it required careful consideration to ensure meaningful analysis could still be conducted.

As mentioned, the non-academic data collected from web. Therefore, I used python to automate this task, especially, used requests and selenium. The Requests library, a versatile Python library for handling HTTP requests, played a fundamental role in the initial data collection phase. It was employed to initiate HTTP GET requests to the Google search engine, where the query term 'project management' was used to fetch search results. Requests' efficiency in sending HTTP requests and processing server responses enabled the rapid retrieval of links to webpages. Furthermore, Requests was pivotal in extracting the publication dates associated with each webpage. This operation involved the parsing of the HTML content of the search results, locating and extracting date information embedded within the source code. Requests' simplicity and speed made it an ideal tool for this initial data retrieval task.

Selenium, renowned for its web automation capabilities, assumed a crucial role in the subsequent phases of data collection. Selenium was employed for its unique ability to interact with dynamic web content, execute user-like interactions with webpages, and extract structured data. In this project, Selenium was used to directly access and scrape webpage contents, focusing on the extraction of titles and body text. Its strengths lay in its adaptability to a variety of webpage layouts and its capacity to navigate webpages with complex structures and interactive elements. Selenium's dynamic interaction capabilities, including finding HTML elements, simulating user actions, and capturing data, proved essential for the comprehensive collection of textual data from webpages.

During the initial phase, Requests was employed to initiate HTTP GET requests to the Google search engine, where the search query 'project management' was executed. The search results, comprised of a diverse array of webpages, were retrieved as HTML content. Concurrently, Requests was used to extract the publication dates associated with each webpage. This process involved parsing the HTML content of the search results, employing techniques such as Regular Expressions or parsing libraries, to locate and extract date information embedded within the HTML source code. This step was critical for establishing temporal context within the dataset, ensuring that the collected data was accurately time-stamped.

Following the acquisition of links and dates, Selenium took center stage. Selenium was employed to dynamically interact with webpages and extract the desired content, including titles and body text. Selenium's flexible nature allowed it to navigate through webpages, locate HTML elements, and simulate user actions, enabling the retrieval of textual data. Its adaptability to diverse webpage structures and its ability to handle JavaScript-driven interactions made Selenium an indispensable tool for the comprehensive collection of webpage data. The process involved initiating Selenium WebDriver sessions, opening webpages, navigating through webpage structures, and extracting data elements based on CSS selectors or XPath expressions.

3.2. NLP implementation

Subsequent to the acquisition of both academic and non-academic datasets, my research embarked on a crucial phase known as data preprocessing. This phase is fundamentally essential in the domain of data science, particularly in the context of Natural Language Processing (NLP). Its primary objective is to transform raw and often unstructured data into a structured and refined format that is amenable to rigorous analysis and modeling.

Data preprocessing is a multi-faceted process that encompasses several intricate steps, each aimed at addressing specific challenges inherent in the data. One of the foundational steps in this process, which warrants a deeper exploration, is tokenization.

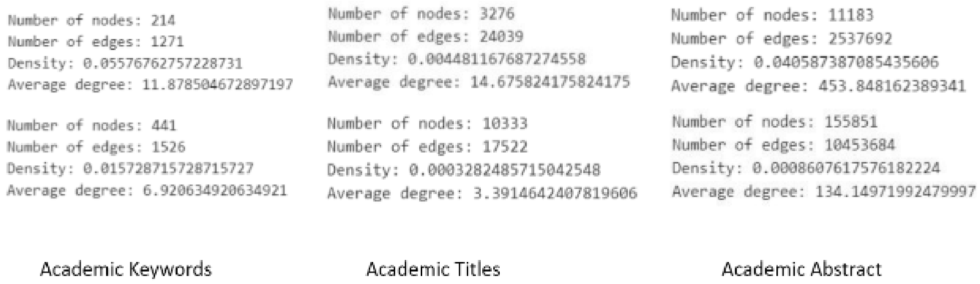
The selection of an appropriate tokenization strategy is not a trivial matter and can have a significant impact on the subsequent analysis. Different delimiters can be chosen based on the characteristics of the data and the specific research questions being addressed. Here, I delve into two essential techniques that play a pivotal role in this phase—stemming and lemmatization. One such critical step that follows stemming or lemmatization is the removal of stop words. This procedure plays a pivotal role in fine-tuning textual data for subsequent analysis, as it seeks to eliminate words that serve as linguistic "fillers" and bear minimal intrinsic meaning. These linguistic fillers predominantly encompass conjunctions such as "because," "and," and "since," as well as prepositions like "under," "above," "in," and "at." While these words constitute a substantial portion of human language, they often lack substantive relevance when developing NLP models.

Subsequently, following the data cleaning process, a meticulously organized dataset emerges, setting the stage for the crucial task of word frequency calculation. Various techniques are available for achieving this objective, with two prominent methodologies frequently employed within the realm of natural language processing: TF-IDF Vectorization and Count Vectorization. In this study for generating the co-occurrence matrix I used the Count Vectorizer in the scikit-learn python module. Unigrams, bigrams, and trigrams are different approaches to representing and analyzing text data, each with its own advantages and use cases. The choice of which one is better depends on specific task and the characteristics of dataset. In this study, I used all of them to get the best results.

Initially, the count vectorizer matrix undergoes a transposition operation. Subsequently, a further transformation is performed by multiplying the transposed matrix by itself. In practical terms, this multiplication operation calculates the co-occurrence relationships between every n-gram and all other n-grams present in the corpus.

3.3. Network creation

Subsequent to the rigorous process of n-gram generation, a critical phase of my research unfolded as I embarked on the construction of network models built around these n-grams using python. To create the network representation, I can use either an adjacency matrix or an edge list (Fig. 7). My methodological approach involved the creation of a dedicated Graph for each distinct n-gram present within my comprehensive dataset.



```
Number of nodes: 214
Number of edges: 1271
Density: 0.05576762757228731
Average degree: 11.878504672897197

Number of nodes: 441
Number of edges: 1526
Density: 0.015728715728715727
Average degree: 6.920634920634921
```

Academic Keywords

```
Number of nodes: 3276
Number of edges: 24039
Density: 0.004481167687274558
Average degree: 14.675824175824175

Number of nodes: 10333
Number of edges: 17522
Density: 0.0003282485715042548
Average degree: 3.3914642407819606
```

Academic Titles

```
Number of nodes: 11183
Number of edges: 2537692
Density: 0.040587387085435606
Average degree: 453.848162389341

Number of nodes: 155851
Number of edges: 10453684
Density: 0.0008607617576182224
Average degree: 134.14971992479997
```

Academic Abstract

Figure 7. Details of the generated networks for academic data (own research)

The visualization of complex networks, characterized by a substantial volume of data and a multitude of interconnected nodes, often poses a significant challenge in terms of clarity and comprehensibility. Then, I filtered the top nodes based on degree to visualize.

3.4. Network visualization

In the following, for instance, unigram and bigram of one of the datasets presented (Fig. 8 and Fig. 9.).

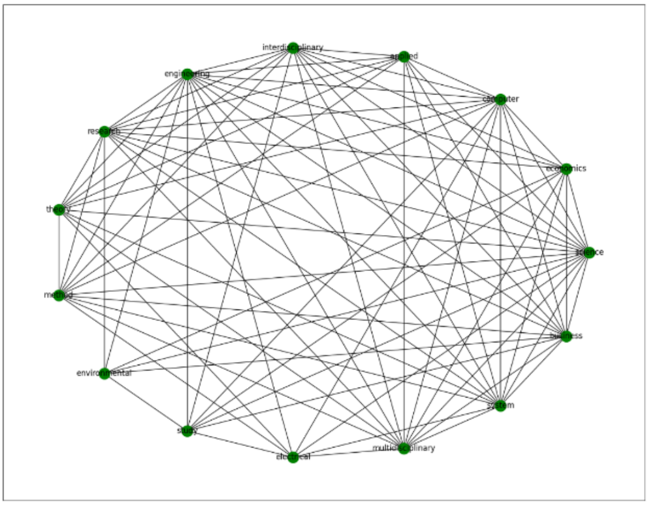


Figure 8. Visualization of Academic data Keywords Unigrams (own research)

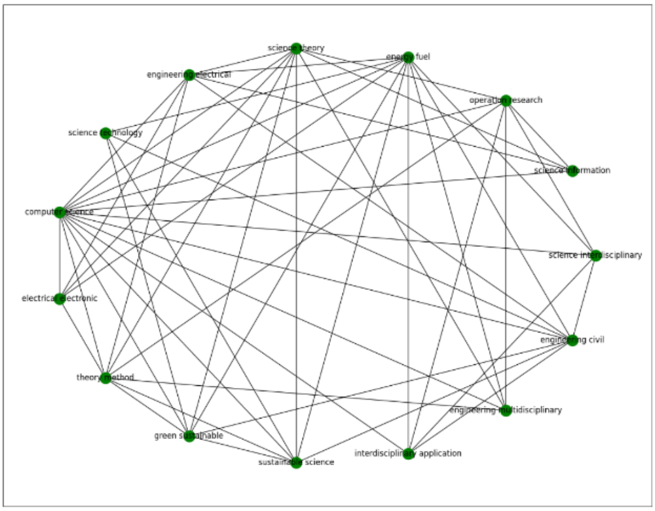


Figure 9. Visualization of Academic data Keywords Bigrams (own research)

3.5. Network analysis

Upon obtaining network and graph visualizations, the next step is to engage in a comprehensive analysis of these graphs in order to extract valuable insights from the underlying networks. In the current study, common analysis such as, Degree Centrality, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality, PageRank Centrality, Hubs and Authorities, Clustering Coefficient, Assortativity Coefficient, Load Centrality, Coreness Centrality, Harmonic Centrality, Average Neighbor Degree, and Closeness Vitality. In addition to this analysis, I applied other non-common analysis such as Sum of Analysis, ICCO ranking, and I-C indicator. In this context, the integration of Kuhn's theory with network analytical tools like the I-indicator, C-indicator, and ICCO ranking offers a novel lens to evaluate the evolution and potential of nodes within a network, metaphorically reflecting the progression of scientific paradigms.

3.6. Relation between, Kuhn’s philosophy and indicators

Integrating Thomas Kuhn's philosophy of science with the I-indicator, C-indicator, and ICCO ranking involves understanding the stages of scientific development and how these indicators can identify and facilitate paradigm shifts. In the pre-paradigm phase, these indicators can assess the potential of various theories. During normal science, they can monitor standard research practice within the paradigm and prioritize areas of research based on key term frequency.

In a crisis, these indicators can identify which alternative theories are gaining traction and which are robust against the anomalies that challenge the old paradigm. If a crisis leads to a paradigm shift, these indicators can be applied to the new paradigm to assess its growth potential and the relevance of its key concepts. After a paradigm shift, they can ensure that the new theories are solidifying their position within the scientific community and continue refining the paradigm as new data and ideas emerge. These metrics can quantify and analyze the structure and dynamics of scientific knowledge, playing a role in identifying conditions that lead to a paradigm shift.

Derived from the ratio of a node’s betweenness centrality to its node degree, the I-indicator provides a comprehensive understanding of a node’s role in the network’s structure and dynamics. It identifies nodes with growth potential, enhances node ranking, facilitates network optimization, and aids in predictive modeling.

The I-indicator’s utility is multifaceted, offering a nuanced perspective on a node’s potential to exert influence and expand its reach in the network. It’s instrumental in pinpointing influential nodes, optimizing network resources, and forecasting future behavior and growth trajectories of nodes. Its versatility makes it a valuable tool across various domains, enabling researchers to gain deeper insights into network structures and make informed decisions.

3.7. Application of machine learning algorithms

I tried to use machine learning algorithms (linear regression, LSTM, CNN, DGNN, Random Forest, XGBOOST, and SVM) as metrices for network analysis. My preliminary step involves the identification of the top 100 most frequently occurring keywords. This specific subset of words warrants my highest level of attention and forms the cornerstone of my subsequent analysis. Labeling is conducted in the following manner, contingent on the movement of the keywords between quarters:

To illustrate the labeling process, we present examples of how nodes were classified based on ranking changes between consecutive periods. The classification criterion focused on the top 100 nodes. We tracked nodes' presence within the top 100 across successive periods, assigning labels based on their ranking dynamics, which facilitated the creation of a meaningful target variable for classification. Nodes were labeled as follows:

Label -1: A node in the top 100 in the first period but not in the second.

Label 1: A node outside the top 100 in the first period but in the top 100 in the second.

Label 0: A node consistently in or out of the top 100 across both periods.

Example Nodes:

“Free” Node: The node "free" remained within the top 100 rankings, indicating sustained relevance, likely due to ongoing interest in concepts like liberty or free resources. **“Policy” Node:** The node "policy" dropped significantly, reflecting a decreased focus on policy-related discussions, possibly due to shifting priorities or emerging new issues. **“Engineering” Node:** The node "engineering" rose dramatically, suggesting increased attention to technological advancements or infrastructure developments (Table 1 and 2.).

Table 1. Node ranking 1. (own research)

	1st Quarter of 2021 Rank	2nd Quarter of 2021 Rank	Label
“Free” Node	6	**13**	**0**
“Policy” Node	54	**1325**	**-1**
“Engineering” Node	1806	**25**	**1**

“Background” Node: The node "background" fell out of the top 100, indicating a shift away from foundational discussions to other emerging themes.

Table 2. Node ranking 2. (own research)

	3rd Quarter of 2022 Rank	4th Quarter of 2022 Rank	Label
“Background” Node	10	**Doesn’t exist**	**-1**

The final result of these algorithms is to classify and predict the trend of a specials keyword in the timeline. The results of the algorithms metrices presented in the following (Tables 3-10):

Table 3. MLP Imbalanced Class Results (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1
Actual -1	4	60	0
Actual 0	1	6368	12
Actual 1	0	61	14

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.8	0.06	0.12	64
Class 0	0.98	1	0.99	6381
Class 1	0.54	0.19	0.28	75

Table 4. MLP balanced Class Results (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1
Actual -1	54	14	0
Actual 0	6	92	4
Actual 1	2	0	68

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.87	0.79	0.83	68
Class 0	0.87	0.9	0.88	102
Class 1	0.94	0.97	0.96	70

Table 5. SVR balanced Class Results (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1
Actual -1	51	14	3
Actual 0	6	92	4
Actual 1	0	0	70

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.89	0.75	0.82	68
Class 0	0.87	0.9	0.88	102
Class 1	0.91	1	0.95	70

Table 6. Random Forest balanced Class Results (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1
Actual -1	64	4	0
Actual 0	4	94	4
Actual 1	0	0	70

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.94	0.94	0.94	68
Class 0	0.96	0.92	0.94	102
Class 1	0.95	1	0.97	70

Table 7. XGBoost balanced Class Results (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1
Actual -1	62	6	0
Actual 0	1	97	4
Actual 1	0	0	70

Classification Report

	Precision	Recall	F1-Score	Support
Class -1	0.98	0.91	0.95	68
Class 0	0.94	0.95	0.95	102
Class 1	0.95	1	0.97	70

Table 8. SVR Imbalanced Class Results (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1
Actual -1	0	64	0
Actual 0	0	6381	0
Actual 1	0	75	0

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.00	0.00	0.00	64
Class 0	0.98	1	0.99	6381
Class 1	0.00	0.00	0.00	75

Table 9. Random Forest Imbalanced Class Results (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1	
Actual -1	7	57	0	
Actual 0	10	6349	22	
Actual 1	0	23	52	

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.41	0.11	0.17	64
Class 0	0.99	0.99	0.99	6381
Class 1	0.70	0.69	0.70	75

Table 10. XGBoost Imbalanced Class Results

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1	
Actual -1	4	60	0	
Actual 0	4	6357	20	
Actual 1	0	20	55	

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.50	0.06	0.11	64
Class 0	0.99	1.00	0.99	6381
Class 1	0.73	0.73	0.73	75

In my pursuit of model validation, I adopted a multifaceted approach. The first validation technique involved randomly selecting data points from within the dataset. This approach allowed us to gauge the model's ability to replicate observed patterns within the data it had been trained on. By comparing the actual outcomes with the model's predictions, I gained valuable insights into its performance in scenarios resembling those encountered during training.

My second validation method, which is elaborated upon in the subsequent section, delved into the use of previously unseen data. This approach is of paramount importance as it simulates real-world scenarios where the model encounters data instances that it has not been explicitly exposed to during the training phase. The validation results presented in Tables 11-14.

Table 11. Random Forest Validation Results (Random Validation) (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1	
Actual -1	34	2	0	
Actual 0	2	44	1	
Actual 1	0	0	37	

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.94	0.94	0.94	36
Class 0	0.96	0.94	0.95	47
Class 1	0.97	1.00	0.99	37

Table 12. XGBoost Validation Results (Random Validation) (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1
Actual -1	65	5	0
Actual 0	1	135	6
Actual 1	0	0	87

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.98	0.93	0.96	70
Class 0	0.96	0.95	0.96	142
Class 1	0.94	1.00	0.97	87

Table 13. XGBoost Validation Results (Unseen Validation) (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1
Actual -1	334	19	0
Actual 0	1	494	5
Actual 1	0	0	342

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	1	0.95	0.97	353
Class 0	0.96	0.99	0.98	500
Class 1	0.99	1.00	0.99	342

Table 14. Random Forest Validation Results (Unseen Validation) (own research)

Confusion Matrix:

	Predicted -1	Predicted 0	Predicted 1
Actual -1	347	6	0
Actual 0	4	492	4
Actual 1	0	0	342

Classification Report:

	Precision	Recall	F1-Score	Support
Class -1	0.99	0.98	0.99	353
Class 0	0.99	0.98	0.99	500
Class 1	0.99	1.00	0.99	342

4. CONCLUSIONS AND RECOMMENDATIONS

In the early phase of this research, I focused on collecting data through a keyword-centric approach centered on 'project management', drawing from both scholarly and practical sources to encapsulate a comprehensive view of the field. This approach facilitated the accumulation of a wide-ranging dataset that informed my network analysis. My exploration of network analysis methodologies was broad and nuanced, incorporating a variety of techniques to unravel different facets of the network.

Thomas Kuhn's philosophy revolutionized the perception of scientific progress by introducing the concept of "paradigm shifts" to describe fundamental changes in scientific disciplines. Contrary to the idea of a smooth, continuous advancement, Kuhn proposed that periods of "normal science" operate under a consensus of beliefs and practices until accumulating anomalies trigger a revolutionary phase, leading to a new paradigm.

Kuhn's stages of scientific development can be related to the use of network analysis indicators like the I-indicator, C-indicator, and ICCO ranking in recognizing and fostering paradigm shifts. In the pre-paradigm phase, these indicators could assess the potential and connectivity of competing theories. During normal science, they could monitor the centrality and clustering of dominant theories.

In my investigation, the comparative performance of various machine learning models revealed only marginal differences. However, XGBoost distinguished itself as the preferred model due to its predictive accuracy and computational efficiency, making it ideal for time-sensitive tasks. It performs well even with smaller datasets, an important consideration when data is scarce. I selected XGBoost for its balanced attributes, confirmed through validation and testing, making it a cornerstone for my research and ensuring its relevance in practical applications.

I employed a thorough evaluation strategy, using separate training and testing datasets to ensure models could generalize well beyond the data they were trained on. By testing these models over ten periods and on new, unseen data, I confirmed their robustness and consistent predictive power, solidifying their potential for practical application in varied scenarios.

For the study contribution can present topics such as: Identification of Research Gaps, Comprehensive Data Collection and Network Construction, Development of Network Analysis Methodologies, Practical Application and Trend Forecasting, and Incorporation of Kuhn's Scientific Philosophy. In the context of future expansion and recommendation I can mention to the Theoretical Expansion (e.g., Algorithmic Evolution, Interdisciplinary Approaches, and Longitudinal Studies) and Practical Expansion (e.g., Decision Support Systems, Educational Tools, and Toolkits for Practitioners).

Despite significant contributions, this study has several limitations. Data source limitations include reliance on available data, which may not capture all project management practices, and variability in data quality, affecting analysis robustness. Methodological constraints involve algorithm selection, where other algorithms might perform better, and initial keyword label imbalance affecting model performance. The scope of analysis is limited by a keyword-centric approach, potentially overlooking broader contextual factors, and the temporal scope, which may not capture long-term trends or recent changes.

Model generalizability is a concern, as findings are tailored to project management and may not apply to other fields without adaptation. Cultural and organizational differences were not extensively considered. Predictive limitations include the probabilistic nature of forecasts and the impact of unanticipated external factors. Technological and practical implementation challenges involve ongoing validation of model efficacy and the need for significant computational resources and expertise. These limitations highlight areas for further investigation to enhance the study's contributions and practical impact.

The study's insights and methodologies offer practical applications for project managers across industries. General recommendations include integrating models with established methodologies like Agile or Six Sigma and developing user-friendly toolkits. For construction projects, predictive models can be used for risk management and network analysis for stakeholder communication. In IT projects, enhancing Agile frameworks with predictive analytics and optimizing resource allocation with algorithms are beneficial strategies.

Healthcare projects can implement decision support systems for compliance and optimize patient pathways using network analysis. Marketing projects can adjust strategies with trend forecasting and apply machine learning for targeted marketing. Manufacturing projects can use predictive models for supply chain optimization and network analysis for quality control. R&D projects can track emerging trends and identify key partners through collaboration networks. Public sector projects can assess policy impacts with predictive models and improve community engagement by mapping stakeholders. Educational projects can adjust programs based on skill predictions and develop interventions for student performance.

Implementing these strategies involves training on machine learning and network analysis tools, developing software solutions, and establishing feedback loops to refine tools based on user feedback and trends. By focusing on these practical applications, project managers can leverage advanced analytical tools and insights to improve project outcomes, drive innovation, and stay ahead of industry trends.

Future expansion and recommendation

The foundational work presented in this study offers a robust platform for future expansion in both theoretical and practical realms. Here are potential avenues for advancement:

Theoretical Expansion:

1. Algorithmic Evolution: Future research can focus on the development of new algorithms or the refinement of existing ones like XGBoost, tailored specifically for project management data analysis. This could involve machine learning and artificial intelligence to better predict project outcomes.
2. Model Complexity: Building more complex models that can take into account the multi-faceted nature of project management, such as stakeholder dynamics, resource allocation, and risk management, could provide deeper insights.
3. Interdisciplinary Approaches: Integrating concepts from psychology, sociology, and organizational behavior could enhance understanding of team dynamics and decision-making processes within project management.
4. Longitudinal Studies: Further longitudinal studies can be conducted to validate and refine the predictive capabilities of the models over longer periods and across different project types and industries.
5. Paradigm Shift Analysis: Expanding on the application of Kuhn's philosophy, researchers could study historical paradigm shifts in project management to better understand and anticipate future changes.

Practical Expansion:

1. Decision Support Systems: Development of advanced decision support systems based on the study's findings could assist project managers in real-time decision-making, risk assessment, and strategic planning.
2. Educational Tools: The insights from this study can be used to create educational and training programs that focus on the application of machine learning in project management.
3. Performance Monitoring: The development of performance monitoring tools that use the study's model to provide ongoing assessment and predictive insights could be invaluable for project managers.
4. Customization for Industries: The model can be customized for specific industries or types of projects, providing targeted predictive analytics for sectors such as construction, IT, or healthcare.
5. Toolkits for Practitioners: Creation of a toolkit that encapsulates the study's methodologies and findings, offering a user-friendly interface for practitioners to apply these insights in their projects.

Cross-disciplinary Applications:

1. Beyond Project Management: The methodologies developed could be tested and adapted for other fields such as supply chain management, logistics, or business process optimization.
2. Cultural Adaptability: Future studies could examine how these models perform across different cultural contexts and organizational structures, potentially leading to region-specific models.
3. Integration with Existing Frameworks: The models could be integrated with existing project management methodologies like Agile, PRINCE2, or Six Sigma, to enhance their predictive capabilities.
4. By building upon the groundwork laid by this study, both scholars and practitioners can drive the field of project management toward a more data-driven, predictive, and efficient future. The potential for expansion is significant, with opportunities to make substantial contributions to the theoretical underpinnings of the discipline as well as to the practical tools and techniques used in the industry.

5. NEW SCIENTIFIC RESULTS

This chapter outlines the innovative scientific outcomes derived from this thesis. The results, along with their implications and relevance to the field, are enumerated below:

1. Integration of Machine Learning with network analysis: I confirmed the potency of integrating Machine Learning with network analysis, particularly in the realms of natural language processing and machine learning algorithms. Despite being underutilized, this fusion has yielded significant advancements, particularly in the precise labeling and categorization of keywords. Through the seamless integration of machine learning techniques, the analysis of textual data in project management literature has attained newfound efficiency and precision, culminating in results that are not only more accurate but also more substantively insightful.

2. Inclusion of Kuhn's Scientific Philosophy: I demonstrated the profound impact of incorporating Thomas Kuhn's philosophy of science into modern analytical techniques, providing an innovative lens for anticipating paradigm shifts within project management. This integration has enriched the theoretical framework of the study, offering a profound and comprehensive understanding of the intricate dynamics within the field, thus paving the way for more informed decision-making.

3. Model Choice and Validation: I established the superiority of XGBoost among various machine learning models, owing to its exceptional predictive efficiency and computational speed. Through meticulous validation, the efficacy of this model has been unequivocally confirmed, showcasing the immense potential of machine learning in amplifying the accuracy and efficiency of data analysis in real-world applications.

4. Practical Implementation and Trend Prediction: I validated the practical implications of my research findings by offering invaluable insights for project managers to adeptly navigate future industry shifts. By harnessing the power of predictive analysis on terminological trends and patterns, practitioners can proactively anticipate and strategize for forthcoming changes within the sector. This pragmatic approach underscores the tangible relevance and utility of the study, equipping practitioners with indispensable tools and strategies.

5. Longitudinal Examination: I corroborated the efficacy of the longitudinal examination approach employed in this study, characterized by a meticulous examination of data over various time intervals. This methodological rigor is instrumental in ensuring the consistency and reliability of selected models, thereby enabling a comprehensive understanding and forecasting of long-term trends within the field. By incorporating a temporal dimension, the study achieves a nuanced analysis of dynamic shifts, thereby offering valuable insights into future developments in the field of project management.

6. SUMMARY

Finally, the steps of the research process are summarized step-by-step:

1. Keyword Selection: The research journey began with the careful selection of relevant keywords. The keywords were chosen based on their relevance to the research topic.

2. Web Search: The selected keywords were then used to conduct a comprehensive search on the web. This wasn’t a simple task of entering words into a search engine; it required a strategic approach to ensure that the search results were both relevant and comprehensive.

3. Time Frame Division: To enhance the accuracy of the results, each month was divided into three-time frames. This division was based on the understanding that data trends can change significantly within a month.

4. Data Preservation: After the data was collected, it was saved for further analysis. This step was not just about storing data; it was about organizing the data in a way that would make the subsequent steps more efficient.

5. Data Preprocessing: Next step involved preprocessing the saved data. This step was crucial in ensuring the quality of the research findings.

6. Network Creation: The research journey continued with the creation of a network.

7. N-gram Network Creation: The next step was to create different networks based on n-grams. The n-gram networks were created by considering different combinations of n-grams.

8. Network Analysis: Once the networks were created, they were subjected to rigorous analysis.

9. Node Filtering: The top nodes were then filtered based on their degree.

10. Node Labeling: The nodes were then labeled. This step involved assigning meaningful labels to the nodes, which facilitated the interpretation of the network analysis results. The process of labeling nodes was predicated on the top 100 nodes.

11. Machine Learning Application: Machine learning algorithms were then applied to the analysis and labels.

12. Model Selection: The best models were then selected based on their performance. This step involved comparing the performance of different models and selecting the ones that provided the most accurate predictions.

13. Model Validation: The selected models were then validated. This step involved testing the models on a validation set to ensure that they were able to generalize well to unseen data.

14. Model Testing: The final step involved testing the models. This step provided a final check on the performance of the models and ensured that they were ready for deployment.

7. LIST OF PUBLICATIONS

Journal publications

1. **Abdi Khalife, Mehrzad** ; Dunay, Anna ; Illés, Csaba Bálint (2021): Bibliometric Analysis of Articles on Project Management Research.*PERIODICA POLYTECHNICA SOCIAL AND MANAGEMENT SCIENCES* 29 : 1 pp. 70-83.
2. Al-Hanakta, Reham ; Illés, Bálint Csaba ; Dunay, Anna ; Abdissa, Gemechu Shuremo ; **Abdi, Khalife Mehrzad** (2021): The Effect of Innovation on Small and Medium Enterprises: A Bibliometric Analysis. *VISEGRAD JOURNAL ON BIOECONOMY AND SUSTAINABLE DEVELOPMENT* 10 : 1 pp. 35-50
3. Hervie, Dolores Mensah ; Illés, Bálint Csaba ; Dunay, Anna ; **Abdi Khalife, Mehrzad** (2021): Bibliometric analysis of human resource management (HRM) in the hospitality and tourism industry. *VADYBA: JOURNAL OF MANAGEMENT* 37 : 1 pp. 59-80.
4. Shuremo, Gemechu Abdissa; Illés, Bálint Csaba ; Dunay, Anna ; **Abdi Khalife, Mehrzad** (2020): The effect of corporate social responsibility on small and medium enterprise sustainability: Bibliometric analysis. *SELYE E-STUDIES* 11 : 2 pp. 41-58.
5. **Abdi Khalife, Mehrzad** ; Dunay, Anna (2019): Project business model with the role of digitization and circular economies in project management. *MODERN SCIENCE / MODERNI VEDA* 6 (2019) : 6 pp. 106-113.

Book of proceedings

6. **Abdi Khalife, Mehrzad ;** Dolatabadi, Amir Hosian Kamali ; Illés, B. Csaba ; Dunay, Anna (2020): Application of big data in Project Management. In: Dimitar, Kirilov Dimitrov; Dimitar, Nikoloski; Rasim, Yilmaz (eds.) Proceedings of XIV. International Balkan and Near Eastern Social Sciences Congress Series on Economics, Business and Management-Plovdiv / Bulgaria, September 26-27, 2020. Plovdiv, Bulgaria, pp. 366-370.
7. Koszty, David ; Fodor, Zita ; **Abdi Khalife, Mehrzad** (2020): Controlling systems and competitiveness at SMEs: A bibliometric network analysis. In: Dimitar, Kirilov Dimitrov; Dimitar, Nikoloski; Rasim, Yilmaz (eds.) Proceedings of XIV. International Balkan and Near Eastern Social Sciences Congress Series on Economics, Business and Management-Plovdiv / Bulgaria, September 26-27, 2020. Plovdiv, Bulgaria, pp. 93-100.

8. Saadi, Shahbaz Ahmad ; Dunay, Anna ; **Abdi, Khalife Mehrzad** (2020): A Bibliometric Analysis and Literature Review of Project Management in Small andMedium Enterprises. In: Dimitar, Kirilov Dimitrov; Dimitar, Nikoloski; Rasim, Yilmaz (eds.) Proceedings of XIV. International Balkan and Near Eastern Social Sciences Congress Series on Economics, Business and Management-Plovdiv / Bulgaria, September 26-27, 2020. Plovdiv, Bulgaria, pp. 427-435. , 9 p.
9. Dunay, Anna ; Illés, Bálint Csaba ; **Abdi Khalife, Mehrzad** ; Daróczi, Miklós (2019): Project team and human resource model in digitization era. In: Illés, Bálint Csaba (szerk.) Proceedings of the International Conference on Management: "People, Planet and Profit: Sustainable business and society" Volume I. Gödöllő, Magyarország : Szent István Egyetemi Kiadó Nonprofit Kft. 385 p. pp. 170-176.
10. **Khalife, Mehrzad Abdi** ; Mahdavi, Iraj (2015): Optimization modeling in construction project risk mitigation, literature review. In: Nenad, Mladenović; Dragan, Urošević; Zorica, Stanimirović 42nd International Symposium on Operations Research- Proceedings, pp. 638-641.

Other

11. **Abdi Khalife, Mehrzad ;** Dunay, Anna (2019): Role of digitization and circular economies in project management, literature review. In: Bálint, Horváth; András, Borbély; Eszter, Fodor-Borsos; Péter, Földi; Amelita, Kata Gódor; Zsombor, Kápolnai (szerk.) V. Winter Conference Of Economics PhD Students And Researchers : Book of Abstracts Gödöllő, Magyarország : Doktoranduszok Országos Szövetsége (DOSZ) (2019) 138 p. pp. 10-10.
12. Dunay, Anna ; Illés, Bálint Csaba ; **Abdi Khalife, Mehrzad** ; Daróczi, Miklós (2019): Project team and human resource model in digitization era. In: Fodor, Zita (szerk.) Book of Abstracts of the 9th International Conference on Management : "People, Planet and Profit: Sustainable business and society" : 9th ICoM 2019, Gödöllő,: Szent István Egyetemi Kiadó (2019) 178 p. pp. 38-38.
13. Illés, Bálint Csaba ; **Abdi Khalife, Mehrzad** ; Dunay, Anna (2019): Agile project management model compatible for the technological advancement and evolution. In: Fodor, Zita (szerk.) Book of Abstracts of the 9th International Conference on Management : "People, Planet and Profit: Sustainable business and society" : 9th ICoM 2019. Gödöllő, Magyarország : Szent István Egyetemi Kiadó (2019) 178 p. pp. 39-39.
14. **Mehrzad, Abdi Khalife** (2018): Project management in industrial digitalization era and industry 4.0, literature review, and analysis the gap. In: VI. International Scientific Conference on Young Science (2018)