MATH6183 – Final Group Coursework

Abstract Analysis

Coursework G2 (Text Mining) 27

Group Members:

Kishore Rajendra - 34812636
Sachin Suresh - 34812598
Umesh Uddar - 34884157
Vyom Khanna - 28965736

Highlights

A dataset of 4385 research papers published between 2000 and 2022 is explored in this study. The first step is to organize the data and create visual summaries like word clouds and yearly trends. As a next step, advanced techniques are used, such as Topic Modelling, Regression, Classification, Association Rule Mining, Clustering, and Dimensionality Reduction. As a result of these methods, we are able to gain insights into topics, citation predictors, journal classification accuracy, word associations, and more. Furthermore, the findings have practical applications for researchers, journals, and academic decision-makers.

INDEX

1. Introduction

In this analysis, we dive into a dataset containing information about over 4385 research papers published between 2000 and 2022 by three major journals. These journals are the Journal of the Operational Research Society, Health Systems, and the Journal of Simulation. Dataset details include the paper's title, its journal, when it was published, the number of pages, authors, views, citations, and more. The main goal is to dig into these abstracts and associated details, using different methods in order to uncover interesting things. The goal is to see how papers have changed over time, whether certain topics stand out, and what factors might influence attention and impact. By analysing these papers, we can learn about their trends and characteristics.

2. Data Pre-Processing

To initiate the analysis, we began by loading the dataset from the "journal_data.csv" file. This involved reading the CSV file into R and creating a data frame to work with. Subsequently, we delved into data preprocessing to ensure the dataset's cleanliness and readiness for analysis. Several key steps were undertaken during this phase:

1. **Handling Missing Values:** We addressed any missing values in the dataset, utilizing the `na.omit()` function to remove rows containing null or undefined values. This step ensured a more complete dataset for subsequent analyses.

2. **Text Data Transformation:** Focusing on the abstracts, we transformed the raw text data into a structured format suitable for analysis. This involved creating a corpus, converting all text to lowercase to maintain consistency, removing punctuation, numbers, and common English stop words to isolate meaningful terms, and stripping unnecessary whitespaces.

3. Document-Term Matrix (DTM) Creation: In text mining, we structured raw abstracts by converting them to lowercase and removing unnecessary characters. The resulting Document-Term Matrix (DTM) captured term frequencies across documents, serving as a foundation. TF-IDF analysis emphasized key terms, distinguishing abstracts and revealing significant themes. The creation of the DTM was a crucial step, forming the basis for subsequent tasks by capturing term frequencies.

4. TF-IDF Transformation: The TF-IDF (Term Frequency-Inverse Document Frequency) transformation was applied to the DTM. This step aimed to highlight the importance of terms within the corpus, considering both their frequency in individual documents and their rarity across the entire dataset.

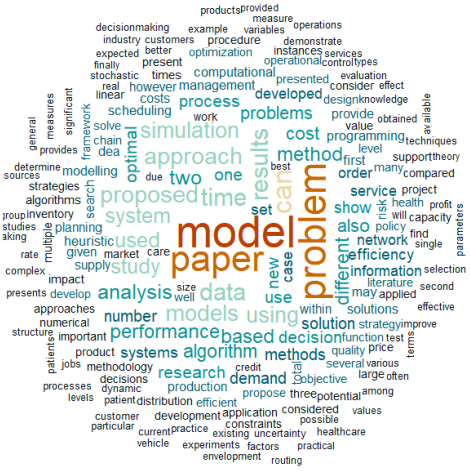
These preprocessing steps ensured a clean, standardized dataset for further analysis, enabling us to delve into meaningful patterns and insights within the abstracts and associated metadata.

3. Exploratory Data Analysis (EDA)

In exploring the dataset, we conducted a thorough examination and visualizations to gain a comprehensive understanding of its characteristics.

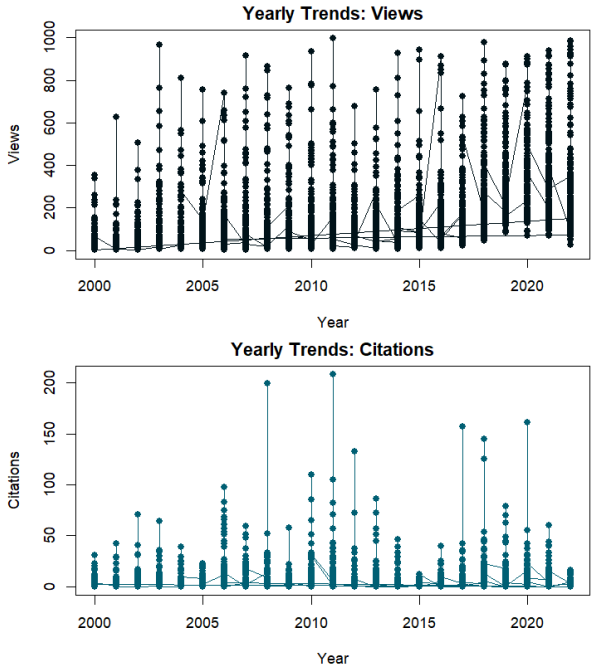
1. Visualizations:

- **Word Cloud:** A Word Cloud was generated to visually represent the distribution of popular words within the abstracts. Larger and bolder words in the cloud indicated higher frequency, offering an intuitive overview of prevalent terms.



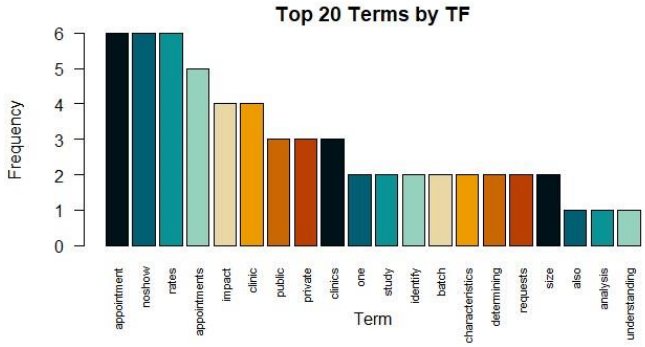
Word Cloud for Bag-of-words

- **Yearly Trends:** Plots were created to illustrate how certain metrics, such as views or citations, vary across different years. This allowed us to identify any noticeable trends or shifts over time.

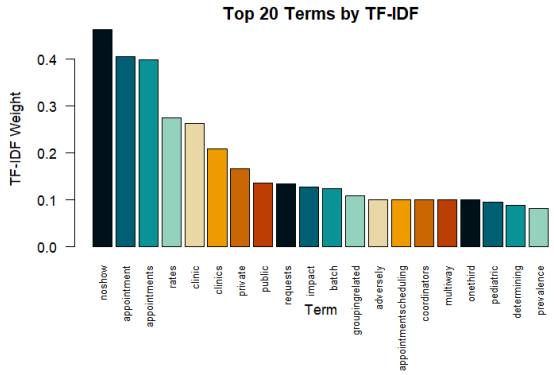


Yearly Trends for Views and Citations

- **Top Terms Visualization:** Plots comparing Term Frequency (TF) and TF-IDF in document 10 reveal raw term frequencies and their corpus-wide significance, offering valuable insights into keyword relevance.



Top 20 Terms by (TF)



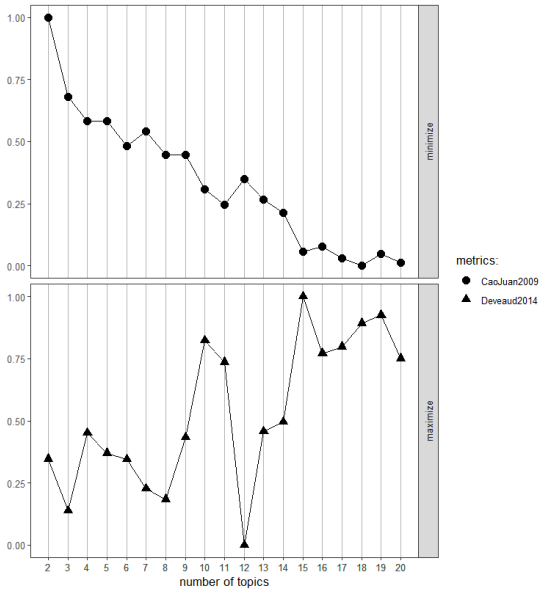
Top 20 Terms by (TF-IDF)

These exploratory analyses served as a foundation for deeper investigations, guiding subsequent tasks such as topic modelling, regression, classification, and clustering. The visualizations and descriptive statistics allowed us to identify potential areas of interest and formulate hypotheses for further exploration.

4. Topic Modelling (LDA):

In the realm of Topic Modelling using Latent Dirichlet Allocation (LDA), we employed a statistical model to uncover hidden topics within the abstracts. Determining the optimal number of topics involved a careful exploration using the Find Topics Number function, considering various metrics such as CaoJuan2009 and Deveaud2014. The goal was to find a balance between coherence and interpretability, ultimately arriving at the most suitable number of topics for our dataset.

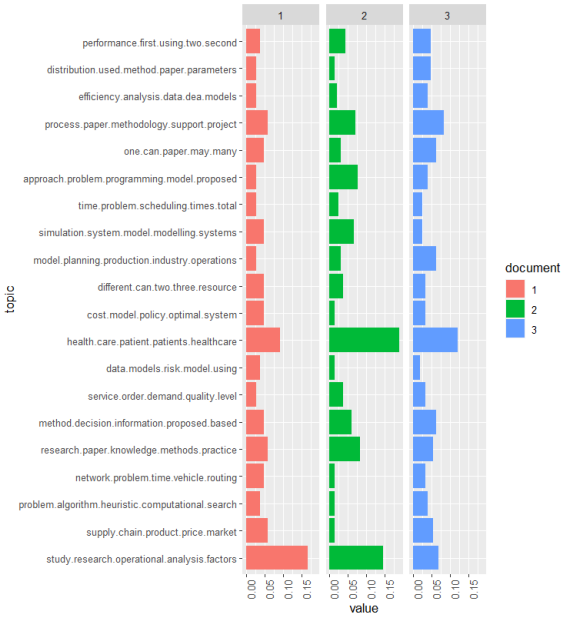
Once the optimal number of topics was determined, we trained the LDA model on the abstracts. The results were presented in the form of topics, each characterized by a distribution of words. Interpretation of these topics involved examining the most prominent words within each topic and understanding the overarching theme they represented. This qualitative analysis shed light on the latent structure and content categories present in the abstracts.



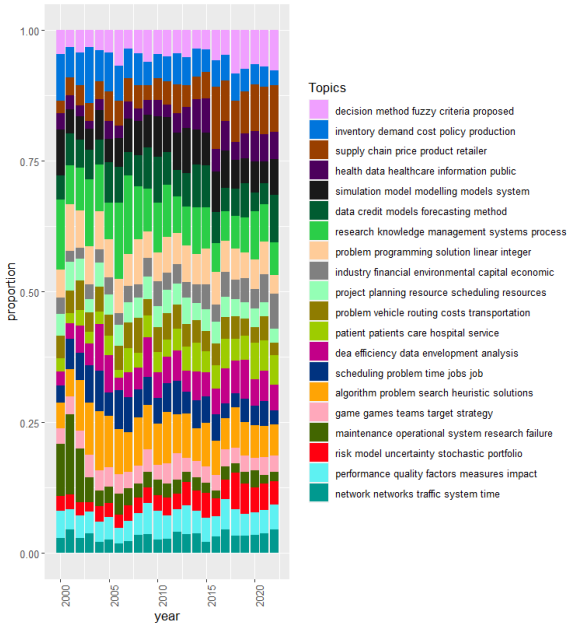
Visualisation optimal number of topics

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
[1,]	"study"	"supply"	"problem"	"network"	"research"	"method"	"service"	"data"
[2,]	"research"	"chain"	"algorithm"	"problem"	"paper"	"decision"	"order"	"models"
[3,]	"operational"	"product"	"heuristic"	"time"	"knowledge"	"information"	"demand"	"risk"
[4,]	"analysis"	"price"	"computational"	"vehicle"	"methods"	"proposed"	"quality"	"model"
[5,]	"factors"	"market"	"search"	"routing"	"practice"	"based"	"level"	"using"
[6,]	"social"	"strategy"	"solutions"	"number"	"systems"	"criteria"	"can"	"credit"
[7,]	"-"	"products"	"results"	"location"	"management"	"methods"	"results"	"methods"
[8,]	"technology"	"profit"	"proposed"	"transportation"	"work"	"fuzzy"	"capacity"	"financial"
[9,]	"change"	"optimal"	"solution"	"new"	"use"	"decisionmaking"	"study"	"forecasting"
[10,]	"findings"	"retailer"	"problems"	"facilities"	"literature"	"group"	"customer"	"results"

Supporting our findings, visualizations showcased the distribution of topics over time and across different journals. These visual aids provided a clear understanding of how topics evolved and whether certain themes were more prevalent in specific years or journals. Additionally, tables were utilized to present the top words associated with each topic, offering a more detailed insight into the content encapsulated by each identified theme. Overall, the LDA analysis contributed valuable insights into the underlying topics present in the abstracts, enriching our understanding of the dataset.



Visualise Topic Proportion



Visualise Topic Proportion per year

5. Regression Analysis:

For the Regression Analysis, we implemented a multiple regression model with the aim of predicting the number of citations a new abstract might receive. This model considered various predictors such as the number of views and the altmetric score to understand their influence on the response variable, which is the number of citations.

The summary of the regression model provided insights into how well the predictors explained the variation in the number of citations. Key information included coefficients, which indicated the direction and strength of the relationships between predictors and citations. A positive coefficient suggested a positive impact, while a negative coefficient indicated a negative impact.

In examining the key predictors, we assessed their significance and impact on the response variable. The discussion focused on understanding which factors, such as views and altmetric scores, played a substantial role in influencing the number of citations. This regression analysis aimed to uncover the relationships between these variables and provide insights into the factors contributing to the citation impact of the abstracts.

```
Residual standard error: 8.049 on 2637 degrees of freedom
Multiple R-squared:  0.6131,    Adjusted R-squared:  0.514
F-statistic: 6.189 on 675 and 2637 DF,  p-value: < 2.2e-16
```

6. Classification Analysis:

For the classification analysis, we employed a model to predict which journal a paper belongs to base on its features. The purpose of this classification model was to understand how accurately we could determine the journal category of a paper.

After training the model, we evaluated its accuracy, which is a measure of how often the model's predictions were correct. This accuracy score provided a clear indication of the model's performance in classifying papers into their respective journals.

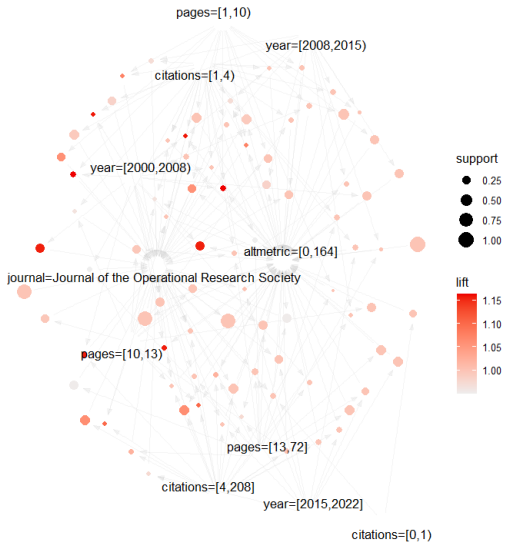
The insights gained from the classification model were significant. Understanding the accuracy helped us assess the reliability of the model in correctly assigning papers to their appropriate journals. This information is crucial for identifying patterns and distinctions between different journals, contributing to a more nuanced understanding of the dataset. The classification analysis served as a valuable tool in exploring the characteristics that distinguish papers published in the Journal of the Operational Research Society, Health Systems, and the Journal of Simulation.

7. Association Rule Mining:

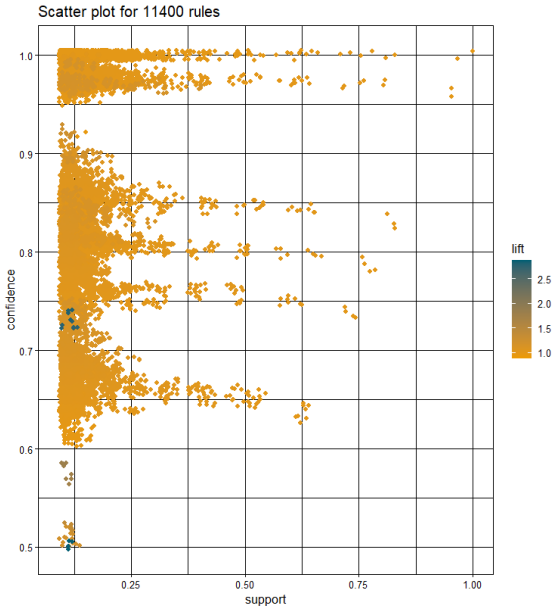
For Association Rule Mining, our approach involved converting the abstracts into a binary matrix, where the presence or absence of specific terms was highlighted. We then applied the Apriori algorithm, considering a support threshold of 0.1 and a confidence threshold of 0.8 to identify meaningful associations between words.

The most significant association rules were derived from this analysis, showcasing patterns of words frequently occurring together in different articles. These rules provided insights into the relationships between terms, unveiling potential connections and themes within the dataset.

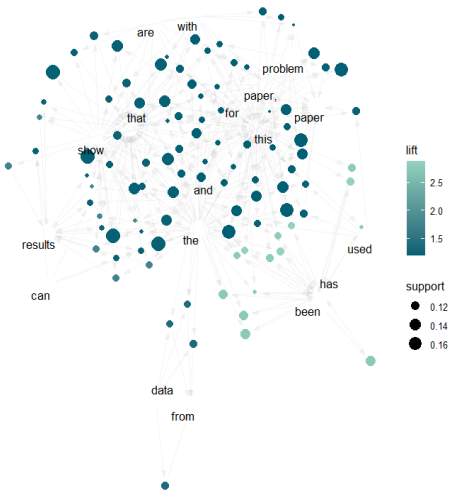
Discussing actionable insights, certain association rules could guide future research directions or highlight key concepts that often co-occur. For example, if terms related to a specific methodology consistently appear together, it might suggest a prevalent research approach in the field. Exploring these associations adds depth to our understanding of the content and potential collaborations within the published papers.



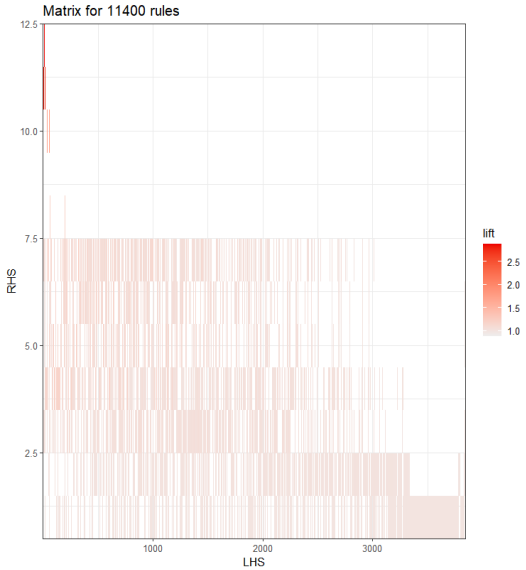
Association Rule Plot



Scatterplot of support vs. confidence



Graph of the rules



Matrix-based plot

8. Clustering Analysis:

In the domain of clustering analysis, we applied a method to group similar abstracts together. The approach utilized key parameters to define how abstracts should be grouped based on common characteristics. For instance, the number of clusters and similarity measures played crucial roles in this process.

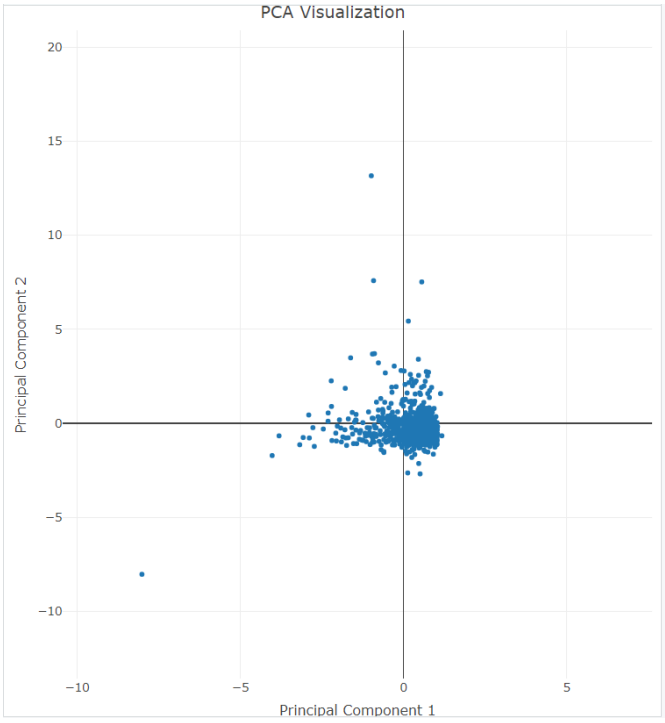
Upon implementing the clustering analysis, we uncovered distinct clusters within the dataset. These clusters represent groups of abstracts sharing similarities in content, potentially revealing common themes or topics. Interpretation of these clusters involved examining the characteristics and commonalities among abstracts within each group.

9. Dimensionality Reduction (PCA):

In Dimensionality Reduction using Principal Component Analysis (PCA), we applied a method to simplify our data while retaining essential information. PCA works by transforming a high-dimensional dataset into a lower-dimensional one, focusing on the most significant aspects.

The results of applying PCA were presented through visualizations, specifically a 2D plot that showcased the clusters identified in the data. By reducing the dimensions, we aimed to capture the key patterns and relationships within the dataset in a more manageable form.

Interpretation of the reduced dimensions involved understanding the contributions of different variables to each principal component. This insight allowed us to grasp which aspects of the data were crucial in forming the identified clusters. The simplicity introduced by dimensionality reduction facilitated a more straightforward understanding of the data's structure and patterns, aiding in drawing meaningful conclusions from the analysis.



PCA Visualization

10. Conclusion:

In wrapping up our analysis, several key findings emerged from each facet of our exploration:

1. Bag of Words and TF-IDF:

- Identified popular words through word cloud and TF-IDF analysis, revealing themes and significant terms.
- Explored how word importance varied across different years and journals.

2. Topic Modelling (LDA):

- Utilized LDA to unveil hidden topics within abstracts.
- Determined optimal topic number, showcasing topics' evolution over time and across journals.
- Extracted meaningful insights from topic distributions and associated word lists.

3. Regression:

- Built a multiple regression model predicting the number of citations based on views and altmetric scores.
- Offered insights into factors influencing paper impact.

4. Classification and Association:

- Developed classification models predicting journal categories.
- Explored word associations, shedding light on common co-occurrences and potential trends.

5. Clustering:

- Identified clusters within the data using appropriate methods.
- Revealed distinctive features through top words within each cluster.

6. PCA:

- Applied PCA to visualize data clusters in a 2D space.
- Enhanced understanding of abstract patterns and relationships.

Overall, our analysis provides a comprehensive understanding of the dataset's content, impact, and categorization. Noteworthy themes in abstracts were uncovered through text mining and topic modelling. Regression insights highlighted factors influencing paper citations, while classification models showcased the predictability of journal categories. Clustering and PCA added a layer of complexity, revealing patterns and relationships within the data. These findings not only contribute to scholarly understanding but also offer practical implications for researchers, journals, and decision-makers in the academic realm. Our data-driven story paints a rich picture of the landscape covered by the examined papers, paving the way for further exploration and nuanced understanding in future studies.