Individual Coursework Report

MATH 6183

Text Mining and Analysis of Academic Abstracts

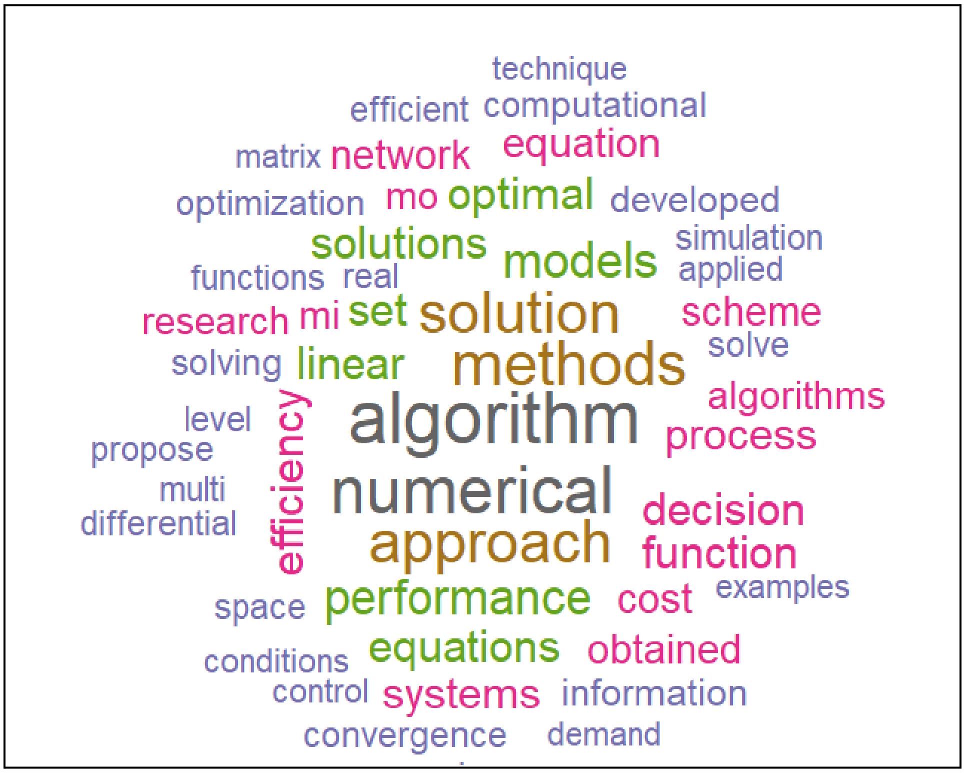
Student No. 37630504

1. Introduction

This report presents a comprehensive analysis of the text abstracts to uncover underlying patterns in academic output. The objective is to apply text mining techniques to characterize the distinctiveness of journal scopes and identify thematic overlaps.

Although the assignment brief states the dataset contains 4,385 entries, on checking the provided file we see that it actually contains **9,385 entries**. This analysis proceeds with the full dataset of 9,385 papers. The dataset corresponds to papers published between 2000 and 2022. The data features papers from four specific publications: the ***Journal of the Operational Research Society (JORS)***, the ***International Journal of Computer Mathematics (IJOCM)***, ***Applied Artificial Intelligence (AAI)***, and the ***Journal of Interdisciplinary Mathematics (JOIM)***. Each entry provides metadata for a research paper including the title, journal name, publication year, page count and the text of the abstract.

The primary data story driving this analysis investigates **the blurring of disciplinary boundaries in modern research.** I hypothesize that while these journals historically served distinct academic fields, ranging from pure mathematics to applied operations research, the integration of ML and AI in most fields has created **topical convergence** in recent years. Also, these high-interest topics are drivers of citation counts, regardless of which journal they belong to.



To test these hypotheses, this report dives into various techniques such as exploratory data analysis using TF-IDF statistics, identifying themes through Topic Modelling (LDA) and applying ML methods (multiple regression, classification) to predict citation impact and journal categories.

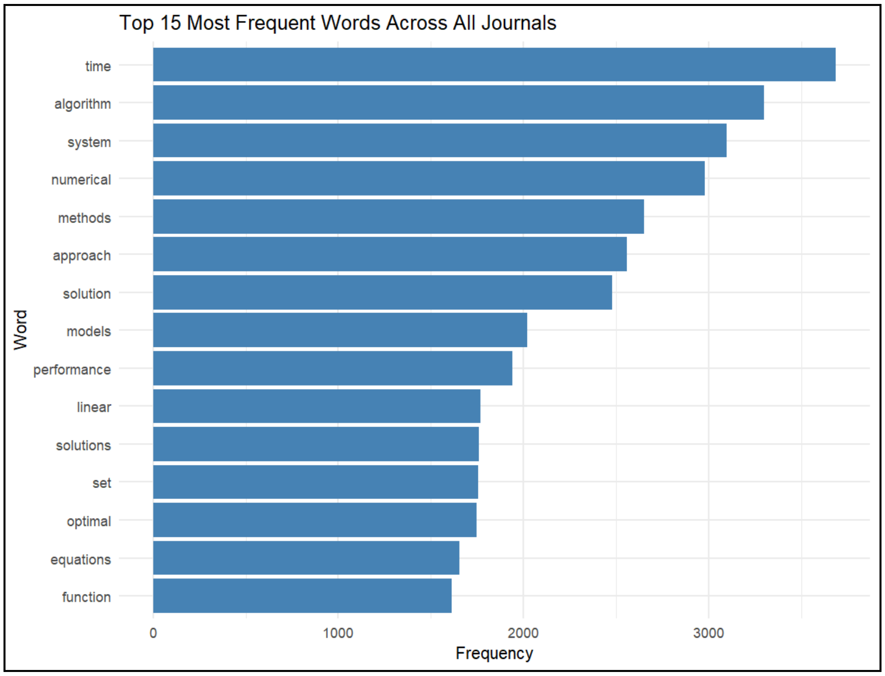
2. EDA and Pre-processing

Before we applied any advanced modeling techniques, the raw textual data required preprocessing to reduce noise and standardise the data. To do so, we used the *tidytext* and *tm* libraries in R, to tokenise the abstracts into individual words. Standard English stop words were removed to focus on content based terms.

Additionally, we also removed some wide domain specific words. Words such as “paper”, “abstract”, “study”, “results”, “proposed”, “method”, “model”, etc. were excluded. While these words are frequent, they are generic to academic writing and don’t offer a lot of significance when it comes to segregating topics based on the specific content of the research. Numeric digits were also stripped from the text to make sure years or parameter values were not being treated as keywords.

2.1 Visualizing the Corpus

To understand the vocabulary of the dataset, a word frequency analysis was conducted.

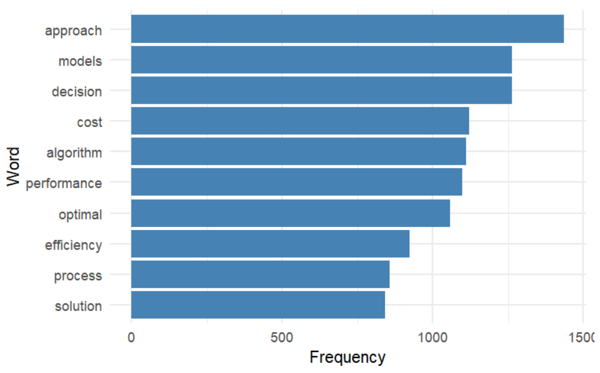


This figure illustrates the most frequent terms across all journals after cleaning. The terms appearing here seem to align with the overall themes of the journals, which focus heavily on quantitative methods and computational systems.

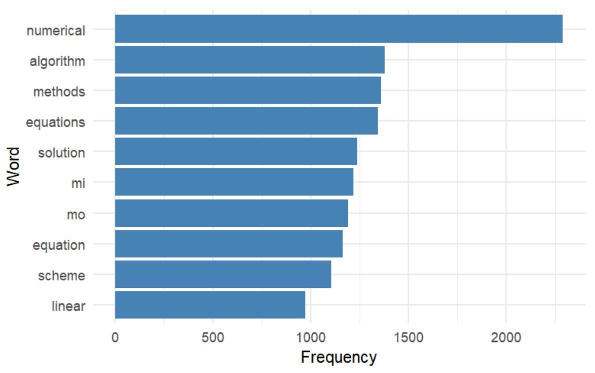
Once we carry out pre-processing of the abstract data, we generate a corpus consisting of a set of documents, each corresponding to a single journal. In doing this, we can then analyse the set of words pertaining to each of the four journals separately. For the purpose of an initial visualisation, we will also create two subsections of the data, pre 2015 and post 2015.

As seen in the figures below, even though there is some topical overlap, there are some distinct top words that can be used to segregate the papers into distinct themes.

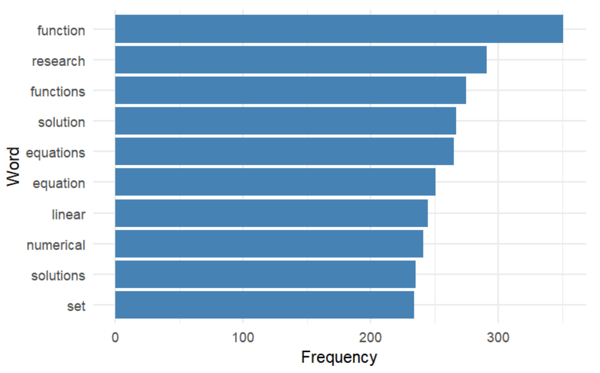
Journal of the Operational Research Society (JORS)



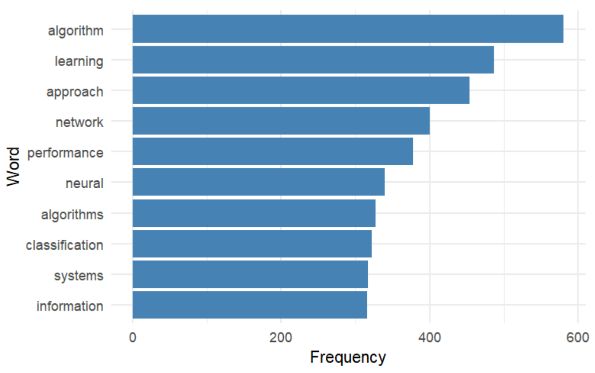
International Journal of Computer Mathematics (IJOCM)



Journal of Interdisciplinary Mathematics (JOIM)



Applied Artificial Intelligence (AAI)

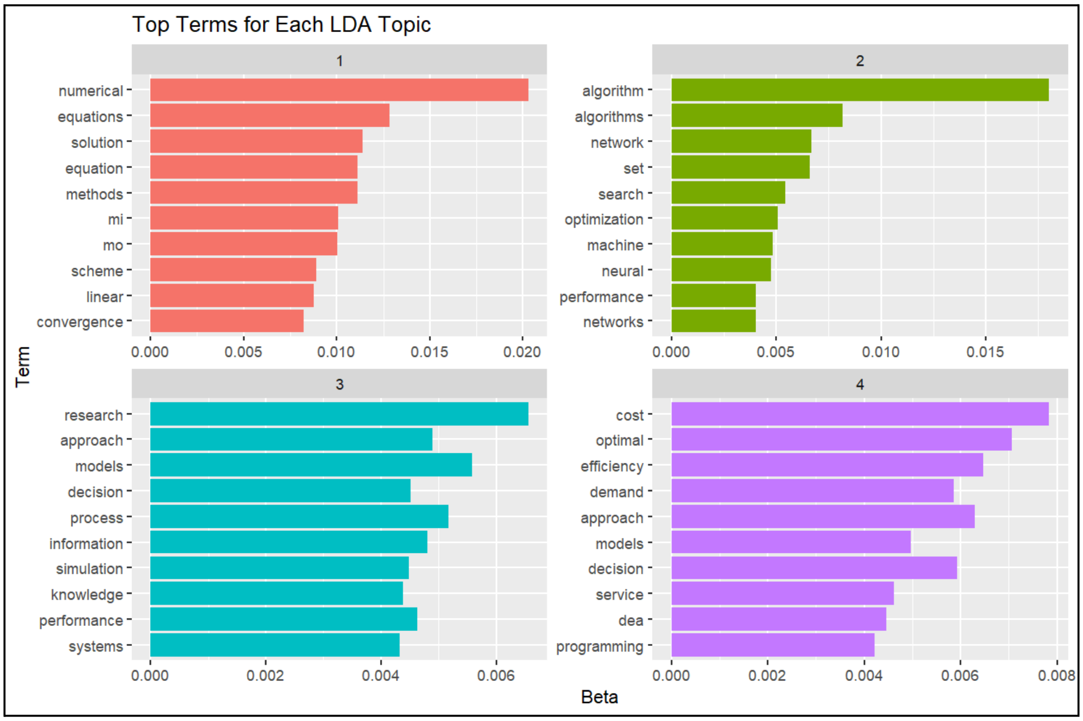


3. Unsupervised Learning: Topic Modelling

To test the hypothesis of thematic overlap, Latent Dirichlet Allocation was applied to the document-term matrix. Given that the dataset broadly falls into three main domains (Mathematics, Operations Research and AI), we run the LDA with both k=3 and k=4 to test whether there is some redundancy in certain topics or if the dataset can truly be divided into 4 distinct topics as per the journals given.

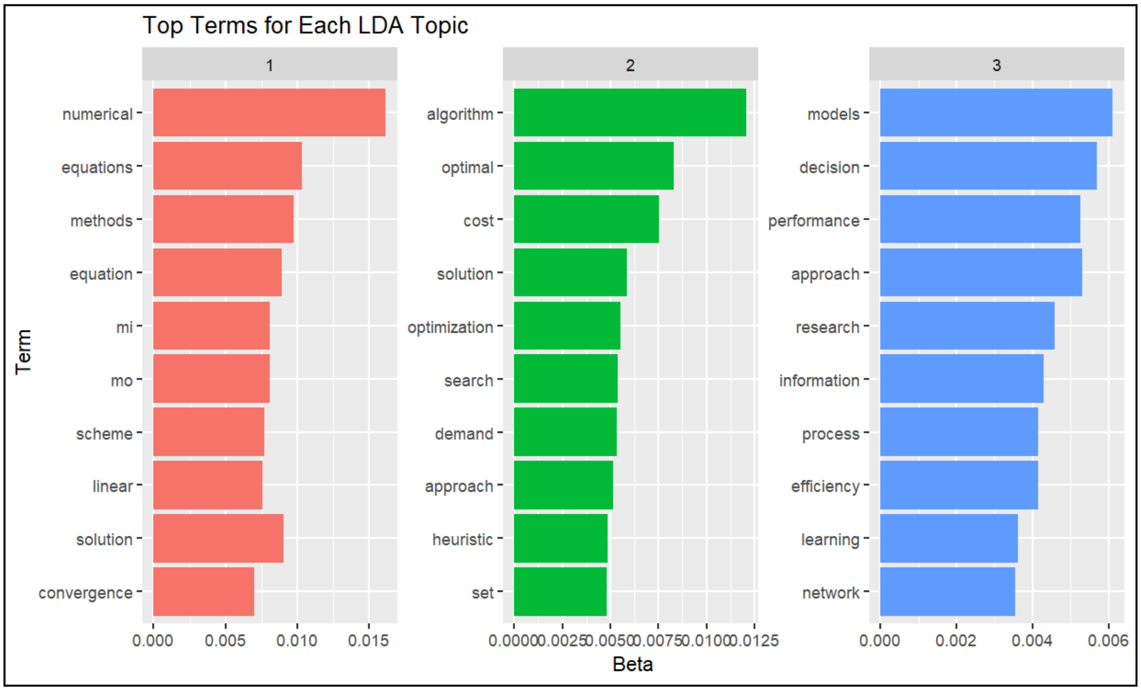
From the figure below, it can be observed that **topics 3 and 4 show a certain degree of thematic and semantic overlap**. However, before dismissing it entirely, we would like to quantitatively test out the performance of both these models. In order to do that, we look towards the **perplexity score of each model**.

Although k=4 yielded a lower perplexity score of **2702.29** compared to the perplexity score of **2884.63** for k=3, we **select k=3 for the final analysis**. While k=4 provides a better statistical fit, it resulted in semantically redundant topics (e.g. splitting ‘mathematics’ into two indistinguishable sub-groups and similar top words such as ‘approach’, ‘models’, ‘decisions’).



This confirms the considerable overlap noted in the assignment brief, suggesting the data is best described by three core themes (**AI, Operations Research, and Mathematics**) rather than four distinct journal identities.

3.1 Interpretation of Topics



The figure above displays the top terms contributing to each of the three topics (beta values).

- **Topic 1:** This topic is characterized by words such as “numerical”, “equations”, “convergence”, etc. This strongly aligns with the scope of *Theoretical or Computer Mathematics.*

- **Topic 2:** The top terms here include “optimal”, “cost” and “demand”. These terms are indicative of *Operations Research*.
- **Topic 3:** Words like “models”, “performance” and “network” are only present in this topic, thus clearly capturing the *Artificial Intelligence* domain.

The clear distinction between these topics validates the use of LDA for feature extraction. These Topic probabilities for each document were extracted and used as numerical features for the subsequent supervised learning tasks.

4. Supervised Learning

Having quantified the text data into topic probabilities, we now test the hypothesis that specific topics drive academic impact (Number of citations).

4.1 Regression Analysis: Predicting Impact of various factors

In this section, we construct a multiple linear regression model to predict the number of citations a paper receives. As citation counts are right-skewed (with many papers having few citations and a few highly-cited papers), the dependent variable was **log-transformed**.

The independent variables include the publication **year**, the number of **pages** and the **topic probabilities** derived from the previous LDA model. Once the gammas from the 3 topics are joined to the original data, we are prepared to run the multiple regression model and observe the results.

As seen below, since Topic_3 was dropped, it becomes our reference category against which the other factors are judged. Thus, the intercept represents the baseline log-citations for a paper that is **100% Topic 3** and **the coefficients for Topic 1 and Topic 2 represent the change in citations relative to Topic 3**.

```
Coefficients: (1 not defined because of singularities)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	51.566648	3.361788	15.339	< 2e-16 ***
year	-0.025140	0.001673	-15.023	< 2e-16 ***
pages	0.026921	0.001652	16.300	< 2e-16 ***
Topic_1	-0.048762	0.028225	-1.728	0.08409 .
Topic_2	-0.099335	0.033262	-2.986	0.00283 **
Topic_3	NA	NA	NA	NA

The results of the regression model reveal several key insights:

1. **Age Effect**
    - The coefficient for year is **-0.025140** which is statistically significant ($p < 0.05$).

   - Since the coefficient is negative and the value of ‘Year’ increases for new papers, it indicates that newer papers have fewer citations.
   - This confirms **the expected trend that older papers have more citations due to the time-dependent nature of accumulation**.
2. **Length Effect**
   - The coefficient of pages is +0.027 while also being statistically significant.
   - Thus, the number of pages has a positive relationship with citations which suggests that **longer papers tend to have a high impact on the citation count**.
3. **Topic Effect**
   - The coefficients for the topics show which research themes are currently more in demand based on citations.
   - Papers focused on **Topic 1** have slightly fewer citations than papers focused on **Topic 3**, but the difference is not statistically significant (p=0.08) and they perform roughly the same.
   - Papers focused on **Topic 2** have **significantly fewer citations** than papers focused on **Topic 3**.
   - This indicates that, holding age and length staying constant, **papers focusing on AI topics seemingly receive more citations than the OR topics.**

That being said, our model has a **very low R-Squared value of 0.043** indicating that **it explains only 4.3% of the variation in citations**. Thus we can conclude that while the identified trends are statistically significant, **the low R-Squared value indicates that abstract words and metadata alone are poor predictors of citation numbers**. Factors external to the dataset, such as author reputation, journal prestige or social media promotion, likely play a much larger role.

4.2 Classification Analysis: Predicting Journal Content

As seen from the topic modelling section as well as the Regression analysis, the 3 different topics are relatively distinguishable. We now test the predictive ability of a classification model based on these discovered topics to check whether these topic groups would be sufficient in classifying abstracts to the required journal. To achieve this, we employ two algorithms to balance interpretability and predictive power: Logistic Regression and Random Forest.

Logistic Regression

To assess how distinct the journals are, we train a logistic regression model to identify papers belonging to each of the 4 different journals (JORS, IJOCM, JOIM, AAI). The model used the LDA topics as predictors as well as pages and years. First off, we analyse the ‘Applied Artificial Intelligence’ using this model and the results are displayed below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.233881	0.129054	-32.81	< 2e-16 ***
Topic_1	-4.931697	0.220541	-22.36	< 2e-16 ***
Topic_2	-2.244557	0.141548	-15.86	< 2e-16 ***

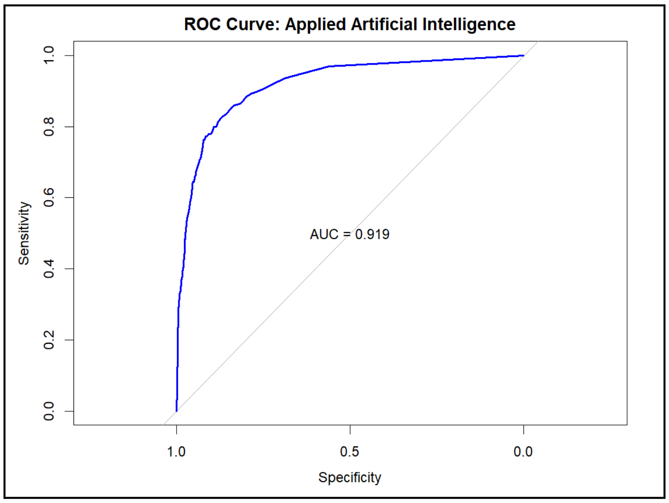
As observed from the table above, the model results show that **Topic 1 and Topic 2 are very weak predictors compared to Topic 3 which is the baseline variable in this analysis.** This confirms that despite the blurring of boundaries, the core vocabulary in the AAI aligns with Topic 3 and it still is distinct enough to statistically differentiate it from pure Mathematics or OR. To further confirm we even see that when the variables in the model are changed to a different baseline, **Topic 3 shows a strong positive coefficient (2.244),** thus proving that as the prevalence of AI keywords increases, so does the likelihood of the paper belonging to this journal.

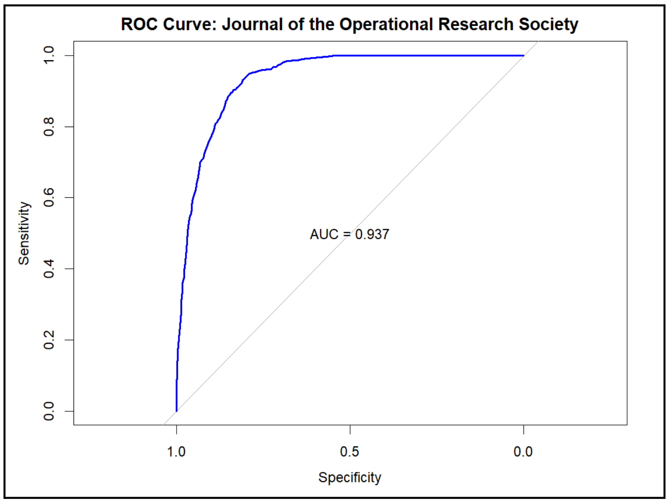
An identical analysis was conducted for the other 3 journals as well. As expected, **Topic 2 was a highly positive predictor for the JORS** and **Topic 1 was a positive predictor for both the other 2 mathematics based journals**. Some of the other major observations are listed below

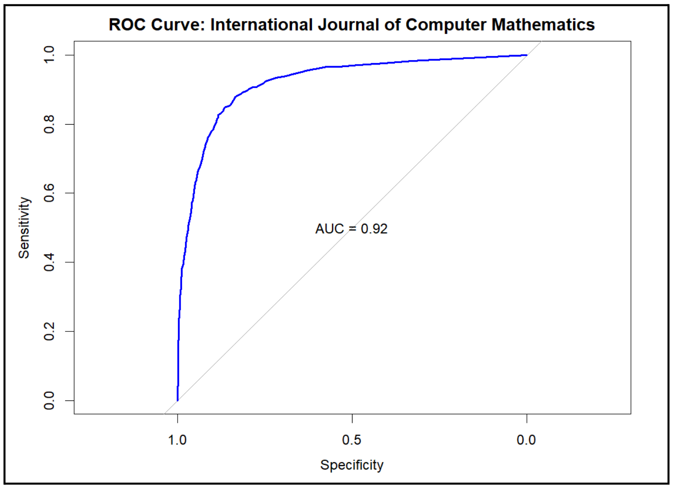
- Papers rich in **Topic 1** are significantly (-6.57) **less likely** to be from JORS than from Topic 3. (There seems to be a much higher topical overlap between JORS and the AI based journal than with the Mathematics based journals)
- The coefficient for pages is negative, which tells us that as the number of pages **increases**, the probability of the paper being in JORS **decreases.** This suggests JORS papers might be shorter than papers in other journals, but this would require further confirmation.

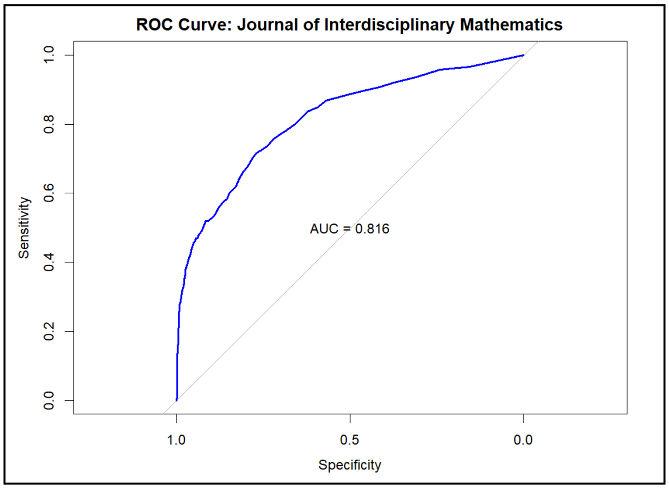
Random Forest

To address the limitations that a logistic regression model might have, we trained a Random Forest classifier to analyse the complex relationship that all the variables might have with each other. As shown in the figures below, the Random Forest model shows strong predictability with 3 of the journals showing distinct enough vocabulary to be successfully classified.


ROC Curve: Applied Artificial Intelligence


ROC Curve: Journal of the Operational Research Society


ROC Curve: International Journal of Computer Mathematics


ROC Curve: Journal of Interdisciplinary Mathematics

It's clear that JOIM is the hardest to tell apart, with an **AUC of 0.816**, which is quite a bit lower than the others. This makes sense because Interdisciplinary Mathematics itself covers many subjects and might be harder to classify.

The steep rise of the curve shows us that the model has relatively high specificity and also seen from the results were the error rates. **The journal of AAI had the smallest error rate of 8% compared to 14% error rates for the other 3 journals**. This could be because the AAI journal uses unique words (like 'neural', 'network') that might not appear in the other journals.

The performance of the Random Forest suggests that the boundary between these journals might not be linear but still quite predictable. Logistic regression works reasonably well and gives us a decent idea about the likelihood of a paper falling under a specific journal based on their topics but capturing the complex interactions between page count, year and topic mixtures (RF) provides a robust classification, further confirming our observations.

5. Clustering and PCA

To further investigate the hypothesis of themes overlapping, we applied unsupervised learning to group the papers based on their content, without using the journal labels. While LDA assigned mixed probabilities, clustering forces each paper into a distinct group, allowing us to test if the text data aligns with the four journal titles.

Finally, clustering was performed using the Partitioning Around Medoids (PAM) algorithm. Unlike LDA, which assigns mixed probabilities, clustering forces each paper into a distinct group based on its topic profile. The analysis used the same feature set (Topic Probabilities) to group papers into k=3 clusters.

5.1 Methodology: Partitioning Around Medoids (PAM)

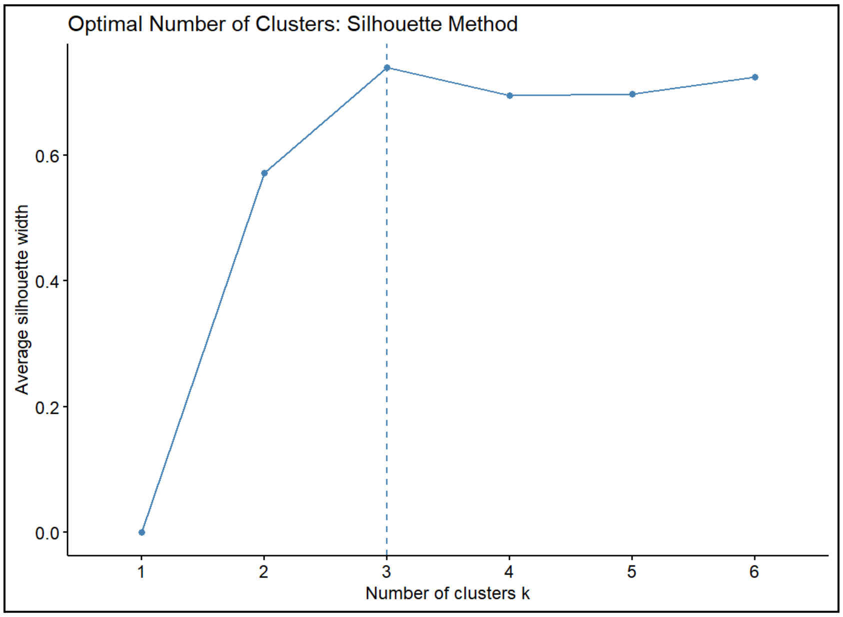
Instead of the standard K-means algorithm, we employed Partitioning Around Medoids (PAM). This is because it is more robust to noise and outliers, as it uses actual data points (medoids) as cluster centers rather than mathematical means.

The input features for clustering were the **Topic Probabilities** (Gamma) derived from the LDA model, rather than the raw word counts. This **dimensionality reduction** step ensures the clustering focuses on the Principal components of thematic similarities (e.g. "AI vs. Math") rather than getting distracted by specific rare words.

5.2 Determining the Number of Clusters (k)

Although the dataset contains papers from four journals, we determined the optimal number of clusters using the Average Silhouette Method.

As shown in figure below, the silhouette width peaked at **k=3**. This suggests that despite the presence of four publication journals, the data naturally forms three distinct thematic groups. Forcing the algorithm to find four clusters resulted in a lower silhouette score, implying that two of the journals likely share common vocabulary.



5.3 Cluster Quality and Journal Association

Comparing the algorithmically generated clusters against the actual journal labels reveals a pattern of convergence as well as distinctiveness.

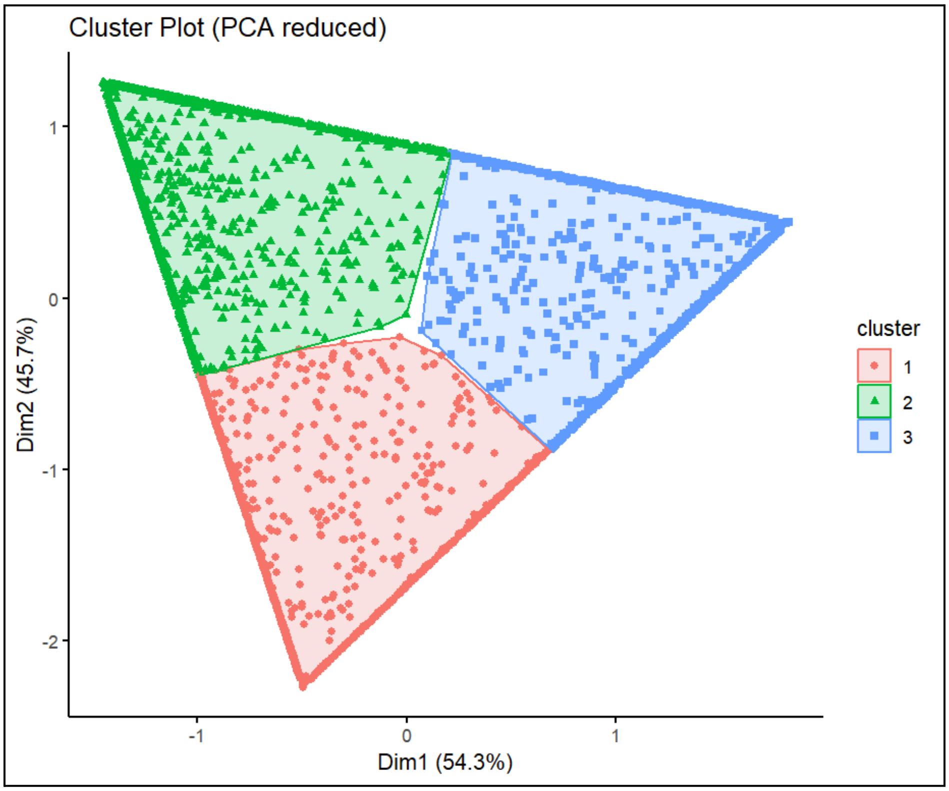
Table: Confusion Matrix of Clusters vs. Journals

Cluster	Applied Artificial Intelligence	International Journal of Computer Mathematics	Journal of Interdisciplinary Mathematics	Journal of the Operational Research Society
1	141 (13.7%)	448 (13.8%)	219 (13.9%)	1739 (49.3%)
2	869 (84%)	264 (8.1%)	572 (36.5%)	1745 (49.4%)
3	24 (2.3%)	2540 (78.1%)	778 (49.6%)	46 (1.3%)
Total	1034	3252	1569	3530

The analysis reveals three distinct thematic identities:

1. **The “Mathematics” Cluster (Cluster 3):** This cluster successfully isolated the mathematical discipline. It captures **78.1%** of all *International Journal of Computer Mathematics* papers and **49.6%** of *Journal of Interdisciplinary Mathematics* papers. Notably, it contains almost no papers from *Applied AI* (~2%) or *JORS* (<1.5%), confirming that the vocabulary of theoretical mathematics remains highly distinct from applied operations research and AI.
2. **The “AI-OR Convergence” Cluster (Cluster 2):** This cluster strongly supports the hypothesis of disciplinary overlap. It captures the vast majority (**84.0%**) of *AAI* papers. However, it also contains **49.4%** of the *JORS* papers. This suggests that nearly half of the Operations Research output uses vocabulary indistinguishable from AI, highlighting the rapid integration of AI methods into modern OR.
3. **The “Traditional OR” Cluster (Cluster 1):** The remaining **49.3%** of *JORS* papers formed a separate group (Cluster 1), distinct from both the pure mathematics and the AI-heavy clusters. This likely

represents the traditional scope of Operations Research (e.g. linear programming, logistics) that has not yet converged with machine learning topics.



Conclusion on Clustering: The analysis demonstrates that while Mathematics remains a distinct island (Cluster 3), the boundary between AI and OR has largely dissolved. *Applied AI* papers effectively form a subset of the broader "Modern OR" theme found in Cluster 2.

6. Conclusion

The findings support the initial hypothesis. The LDA successfully identified three distinct "core" topics: Operations Research, Theoretical Mathematics, and AI. However, the clustering analysis revealed significant overlap between the journals. This confirms that boundaries between the themes are not as rigid and the separation in topics may not be as significant as their allocation to separate journals suggests.

Furthermore, the regression analysis demonstrated that the "topic" of a paper is a significant predictor of its future impact and so is the time (i.e. papers gaining more citations over time). Papers heavily loaded with AI-related terms (Topic 3) showed a statistically significant advantage in citation counts, suggesting that current academic attention is disproportionately focused on this domain.

Limitations

- This analysis relied on single words tokenisation. A multiple-word analysis (e.g. treating "neural network" as one term) might capture context better.
- The regression model explains only a portion of the variance in citations, indicating that factors outside the abstract (such as author reputation or institution) play a big role in a paper's success. In the future we could integrate author data to build a more robust and better performing predictive model.