MATH6183 – Final Group Coursework

- This coursework will count for 40% of the assessment for MATH6183.
- Details of how to submit the coursework are provided below.
- The deadline for the coursework is **15:59 on 9 January 2024**. The time of electronic submission is taken from Blackboard.
- The deadline is strict. Please refer to the course handbooks for the penalties applied to late assignments.
- Remember to back up your work regularly. No additional time will be given for lost computer files.
- This work must be your own. Any references used must be cited and any help received from others should be duly acknowledged. You are reminded that cheating and plagiarism are treated very seriously by Mathematical Sciences and by the University.

Electronic submissions

- You are expected to submit the final project electronically on **9 January 2024**. Details of what to include are given below.
- You will work *individually*. Copying from another person or solution set is not permitted. You are reminded that cheating and plagiarism are treated very seriously by Mathematical Sciences and by the University.
- Please name your files using using the final eight digits of your student ID number only, e.g.12345678.pdf, where 12345678 should be replaced by your student ID number. Do not include your name.
- Your files should be uploaded to the MATH6183 Blackboard site in the Assignments section.

1 Introduction: Analyzing Abstracts

The file `journal_data.csv` can be downloaded from Blackboard and contains 4385 entries corresponding to papers published by the following journals between 2000 and 2022.

- Journal of the Operational Research Society (`https://www.tandfonline.com/journals/tjor20`)
- Health Systems (`https://www.tandfonline.com/journals/thss20`)
- Journal of Simulation (`https://www.tandfonline.com/journals/tjsm20`)

Each row corresponds to a published paper and contains the title, journal name, year, number of pages in the published version, a list of authors, the number of views of the paper on the website, the number of citations of the paper, an altmetric score and the abstract. The altmetric score provides a guide to the amount of attention a paper has received. See this article for more details.

2 Tasks

Your objective in this assignment is to carry out some analysis on these data to look for interesting patterns and draw some general conclusions about what you have found. Below we include some tasks for you to work through.

1. **Bag of Words and TF-IDF:** carry out pre-processing of the abstract data to generate a corpus consisting of a set of documents each corresponding to a single abstract. Investigate the distribution of popular words in the corpus and draw plots to help the reader understand your findings. You may wish to consider how these vary between different years and between different journals.
2. **Topic modelling:** use LDA to search for different topics in the data and investigate how the topics vary between papers. You may wish to consider whether the distribution of topics varies over time or between different journals.
3. **Regression:** Train a multiple regression model that can be used to predict the number of citations of a new abstract submission.
4. **Classification and association:** Using the data in the spreadsheet and results from step 1, build classification models to determine how accurately it is possible to predict which journal a paper belongs to and association models to determine how often particular words appear together in different abstracts. It may also be interesting to see how this differs by journal.
5. **Clustering:** Identify the clusters in the data using an appropriate clustering method. For each cluster, identify top words (e.g. by looking at the features with the highest mean TF-IDF scores in the cluster).
6. **PCA:** Visualise the clusters you have found in 2D by applying PCA to reduce the dimension.

Although these tasks may all be relevant and performed on all or parts of the dataset, you are not expected to perform these one after another, and then present a sequence of results. You should think about each step carefully and customise the task and its analysis in order to answer a few interesting questions that **you hypothesise** on your own. For example, you may suspect that some of these journals are focused on a few topics only, while others publish a broader range of articles. Or perhaps you suspect that there has been a change in the themes or citation behaviour since 2015. How can you customise the steps above to find out if your hypotheses are correct or not? How can you convince the reader that you are indeed correct using appropriate graphics and performance measures? In other words, can you tell an interesting **"data story"** using elegant graphics and well-written English that others will enjoy reading? We very much hope so!

3 What You Should Submit

- R code that works with the data file provided with the assignment brief.

- A written report on your analysis of a maximum of 2000 words. The word count does not include text inside tables or figures or text in references. The report should be saved as a pdf file and should include all of the plots and tables that you need for explaining your results.
- A single page saved as a pdf listing 3-5 "highlights" of your study. A highlight is a sentence of not more than 85 characters (including space) conveying the core findings and provide readers with a quick textual overview of the article. Highlights describe the essence of the study (e.g., key results, take away messages) and highlight what is distinctive about it. (You may be tempted to use a automated AI tool for this, but please do not! This is against the learning objectives of the module and the results of such tools are not always reliable and may indeed result in lower marks.)

4 Mark Scheme

The assignment will be marked out of 100 with the allocation of marks between different aspects as follows.

- R code: clarity and functionality. **20 marks**
- Presentation of report including formatting and the quality of graphics used. You are free to use any format here. You are welcome to feel creative! **20 marks**
- Quality of writing including logical structure, clarity and preciseness of explanations. **20 marks**
- Scope of the work in terms of the methods used in the analysis. **20 marks**
- Quality of interpretation of results, demonstrating an ability to understand the procedures that have been used and the outputs obtained. **20 marks**