AMA546: Mid-Project

Students are required to complete a group project focused on text data science, utilizing computer programming in R or Python. At the project's conclusion, each team is required to deliver a presentation outlining the main findings. Additionally, teams must submit a detailed report along with the corresponding R or Python code files. The report should be no longer than 7 pages, excluding references, tables, and figures, and should include no more than 8 figures and tables in total. All team members are expected to actively participate in every aspect of the project, including statistical analysis, coding, and writing. To ensure fair contribution, students will be asked to provide confidential peer evaluations regarding their teammates' contributions.

1 Project Requirement

The project is related to the **text data analysis**. Students are required to conduct the project under the following datasets and use at least one method (e.g. ridge, lasso, kernel, tree) we mentioned in this course. The students need to choose one dataset and formulate a question by themselves.

- 10-K report data: http://www.cs.cmu.edu/~ark/10K/
  - This dataset is collected by

    Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics (pp. 272-280).*
  - The aim of this work is to investigate the influence of words in 10-K report on stock volatility.
  - It is raw data (text data). The students need to do data processing first.
- Your valid dataset. The students are encouraged to use other text dataset satisfying the requirements:
  - Text data.
  - The number of data should be larger than 2000.

2 Report

2.1 Sections for Report

The report should include the following sections.

1. **Introduction/Background**
   - What specific question are you aiming to answer?
   - What makes this question interesting to you?
   - The source of dataset (In case you use other dataset).
2. **Methods**
   - Summarize and explore the data.
   - What analyses are most appropriate to answer the question of interest?
   - Describe the analyses used.
3. **Result**
   - Present relevant graphics.
   - Interpret results of the analysis.
4. **Conclusion/Summary**
   - What are your conclusions?
   - The contribution of your team members.
5. **Reference**

2.2 Code

You are required to upload the `R`/`Python` code together with your report. Please write clean and readability code. Document your code with comments so that the grader can understand the link between your code and your report.

2.3 Submission

You need to submit a report and a `R`/`Python` code in one `.zip` file on Blackboard before Mar 24, 11:59 am. The file name of your zip file should be `Group#YOUR_GROUP_NUMBER.zip`, where you need to replace YOUR_GROUP_NUMBER with your group number. Only one team member needs to submit the zip file. The file should contain the following:

1. Your report in `.pdf` format. The report should be 11 point Times New Roman, and double spaced. The report should be **no more than 7 pages** (excluding references, tables and figures). The total number of both figures and tables is **at most 8**.

2. Your `R`/`Python` code needs to be ready to run on any computer with `R`/`Python` installed, not just yours. For instance, all dependent data files must be included (and in the folder that is referenced in your code); Your code needs to check if the required libraries/packages are installed (and install them if needed); There should be absolutely no run-time error.

2.4 Grading

The grading of the report depends on the following.

1. Whether you follow the instruction for different sections of the report. (65%)
2. English writing of the report (typos and grammars). (15%)
3. Well documented code. (20%)

3 Presentation

There is no special guideline about the presentation, but the major components should be similar as the components mentioned in Section 2.1. For example, it should contain the proposed problem, data, methodology, outcomes, and your conclusion. Here is something important about the presentation you should know.

- The presentation will hold in the class **on Mar. 19/20**.
- You need to submit your presentation slide on Blackboard before **Mar 19, 11:59 am**.
- The presentation of each group cannot exceed 15 minutes. The schedule will be announced latter. Presenters should arrive at their sessions on time.
- The presentation slides will be uploaded to Blackboard. Therefore, do not include your private information.