Group Project

INT3086 Data Structures and Algorithms for Data Mining

2023-24
Weighting 40% of Final

1. Introduction

This project aims to apply what you have learned in this course to a real-world application. You are required to conduct a data analysis using data mining techniques with a real-world dataset base on the sources to carry out some studies and analysis of a self-defined meaningful application.

2. Group Formation

Each group should consist of maximum to 5 members. A group less than 3 members is not recommended.

3. Process

Study the following resources from the Internet to find a data set that your group is interested to base on it to preforms an analysis.

Dateset (select and download to your project)

https://www.kaggle.com/datasets?datasetsOnly=true

GitHub - awesomedata/awesome-public-datasets: A topic-centric list of HQ open datasets.

Other advance data mining project

https://favtutor.com/blogs/data-mining-projects

https://sites.google.com/site/datathinkingpractice/assignments/student-projects

[Note: The project indicates in these resources are provided for reference only. You are not required to complete your work at this professional standard level]

Your general process

1. Select a dataset your group is interested. Define a topic you want to focus.
2. Write python code to analyze and visualize your findings based on the focus of the project.
3. Use python program to extract suitable data and present the data in respect to your analysis focuses.
4. Provide a report supported with your program outputs to discuss any findings. Limit your report to not over 2000 words but not include the appendix. (e.g., put not essential data as Appendix)

It is not compulsory to use the dataset provided in the references. You may search other websites to select data set. However, your result needs to be original and different from other groups. The relevance of the results should be explained clearly.

4. Evaluation Criteria

Data Analysis Results (20%)

You should work on a problem based on some real-world data sets. The results of your project should be supported by evidence and should provide interesting, useful and meaningful information that is related to the outcomes of your group collected data. (NOTE: prediction is optionally provided in this project)

Programming Skills (20%)

The program needs to be effective to achieve the analysis purpose. While the outcome is clear and meaningful. It is able apply good data structure in your design with evidence (via analysis and discussion) effective performance on time complexity.

Report (20%)

Content all essential components about the project with well and consistent formatting layout. High readability and without major grammatic problems. Understandable and provide suitable graphics, tables and screenshot captures where appropriated.

Difficulty (30%)

Your project will be evaluated based on its difficulty and your knowledge that you can demonstrated. To achieve a higher score, you may explain why your data analysis work is complicated and what techniques you have used to solve the problems you have encountered. You may also explain the principle behind the data mining techniques to demonstrate your understanding.

Group Presentation (about 20 minutes recording) (10%)

The presentation should explain and demonstrate your project outcomes clearly and concisely. Therefore, try to explain clearly in your presentation *why* you think you deserve a high score for your project. The instructor may not give you adequate credits for your project if you have not given relevant reasons and evidence in your presentation.

5. Suggested Report Components

You may consider including some of the components below in your project: (Note it is not limit to these components. You may design your report based on your project while it may be different form the followings).

- Explain the information provide by your selected dataset. For example, how it can give value for a real-world application.
- Provide an implementation of the data mining algorithms to analyze the dataset so as implement your idea on data mining on that selected dataset.
- Explain the design principles behind your data mining algorithms.
- Report the findings from your outcomes of data-mining algorithm.
- Analyse the time complexity of your algorithm design. If any, identify the pitfalls and suggest solutions on your algorithm design.
- Make conclusion and recommendation for improvements.

6. Submission

Each group is required to submit the following on or before the due date (which will set on the Moodle). **Only one member in a group** needs to submit. Late submission may not catch for mark submission.

Summary of submission components:

1. The selected dataset (e.g., CVS file)
2. Source code (e.g., all python program files)
3. Project report (e.g., pdf file format)
4. Presentation powerpoint, and at the first slide, provide a link to access the presentation recording (provide password if any).

NOTE:

Zip all files (not include the presentation video), and folders **Submit to Moodle under the link of [Group project submission].**

For the recording, you may store in Google Drive or youTube with the share/access link in report. DO NOT upload the recording to Private Zone of Google, while it needs permission to access.

Your report needs to be in pdf format and your source code should be Python source code file. If there is more than one Python program files, they need to be arranged in a folder named Python Source Code.

7. Plagiarism Policy

Your project should represent the work of your own group. Do not include in your project any code not written by your group members. You may use any external Python libraries and data sets, provided that you have adequately acknowledged the source and clearly explained which part of your project is not done by your group.

[A general concept is applicable to plagiarism, for example, you import a Python library for use, it would not see as plagiarism. However, you copy and paste a set of code from some existing programs, you may need to explain why and how it is used, and declare they are not your codes.]