WQD7009: Course Assignment (40%)

Deadline for Part 1 and Part 2 is on the 14th of January 2022 (0000hr)

Table of Contents

1 Part 1: (30%)

1.1 Market Basket Analysis (15%)

Method of Evaluation: From Report

Perform Market Basket Analysis on the provided dataset. To provide a report containing the following **compulsory** sections:

- Introduction to Frequent Itemsets & Association Rules Mining.
  To include explanations on the followings:
  - Support
  - Confidence
  - Lift
  - Conviction
- Introduction to the dataset
  - Perform data Analysis & Data Exploration
- To extract the Association Rules from the dataset using A Priori algorithm using Python
  - Hint: Usage of mlxtend package available in Python
- To evaluate and discuss the association rules extracted using the metrics presented (i.e. Support, Confidence, Lift and Conviction)

1.2 Using Singular Value Decomposition (SVD) in a Recommender System (15%)

Method of Evaluation: From Report

Study the tutorial provided in the following link:

https://analyticsindiamag.com/singular-value-decomposition-svd-application-recommender-system/#:~:text=In%20the%20context%20of%20the,given%20to%20items%20by%20users.

Based on the tutorial, write a report for Section (B) of Part (1) that contains the followings **compulsory** sections:

- Introduction to the different techniques involved
  - SVD
  - Recommender System
  - Collaborative Filtering
- Elaboration on the theoretical concept with specific examples taken from the tutorial with specific examples presenting the concepts behind
  - SVD

    - Recommender System
    - Collaborative Filtering
- To explore the usage of evaluation metrics for the developed recommender's system
    - E.g., usage of Mean Squared Errors (MSE) or Mean Average Precision (MAP)

- Working demo with GUI using either
    - https://www.streamlit.io/ OR
    - https://blog.jupyter.org/and-voil%C3%A0-f6a2c08a4a93

1.3 Format for Part 1: Section (a) and (b)

The reports for Section (a) and (b) should follow the following format

- Title page
    - To provide the name of all the members together with the matric ID
- Task distribution among members
- Introduction
    - To provide the required introduction as mentioned earlier in the different sections
- Objective of the report
    - To spell out clearly the objective of the report and what should be your end goal that can be measured as part of your experiments
- Methodology
    - To report the methods used to complete your experiments
    - To discuss the structure of the approach, code, and the whole organization of the solution
- Results & Discussions
    - To present your results and discussion on your results
- Conclusion & Future works
    - To present the conclusion of your work and how the work can be further improved.
- **Please submit the report by the stipulated deadline, i.e., on the 14th of January 2022**

2 Part 2: (10%)

Method of Evaluation: Presentation

The assignation of papers for the different groups are as follows:

*Note: Groups distribution are available via the following link:
https://docs.google.com/spreadsheets/d/1elmNWx-JBzA5eSBwqyeLgySr_AqWjy51/edit?rtpof=true#gid=598842754

i. Random Forests for Big Data
(Groups I and 2)
https://www.sciencedirect.com/science/article/abs/pii/S2214579616301939

ii. A Novel Clustering Method Using Enhanced Grey Wolf Optimizer and MapReduce
(Groups 3 and 4)
https://www.sciencedirect.com/science/article/abs/pii/S2214579618300297

iii. Variations on the Clustering Algorithm BIRCH
(Groups 5 and 6)
https://www.sciencedirect.com/science/article/pii/S2214579617300151

iv. Train Delay Prediction Systems: A Big Data Analytics Perspective
(Groups 7 and 8)
https://e-tarjome.com/storage/panel/fileuploads/2019-05-11/1557562237_E11086-e-tarjome.pdf

v. Anomaly Detection and Repair for Accurate Predictions in Geo-distributed Big Data
(Groups 9 and 10)
https://www.sciencedirect.com/science/article/abs/pii/S2214579618302119

vi. Hadoop MapReduce Performance on SSDs for Analyzing Social Networks
(Groups 11 and 12)
https://www.sciencedirect.com/science/article/abs/pii/S221457961730014X

vii. Lossless Pruned Naive Bayes for Big Data Classifications
(Group 13)
https://www.sciencedirect.com/science/article/abs/pii/S2214579616301320

Based on the paper selected, each group needs to prepare a presentation deck containing the followings:

- Introduction
- Problem Statement
- Research Objectives
- Methodology
- Results
- Conclusion

**** The evaluation will be solely based on the presentation and based upon the followings**

- Clarity of the presentation and the paper presented **(3%)**
- Understanding of the paper (evaluated from Q&A) **(3%)**
- Creativity in presenting the paper **(4%)**

**** Please submit the presentation deck by the stipulated deadline, i.e., on the 14th of January 2022**

3 Rubric

3.1 Rubric for Part 1

Criteria/ Marks Allocations	2-18 marks	19-30 marks	31-50 marks
Report Evaluation 1: (Structure and Clarity, 50 marks)	Report meets the bare minimal standard of structure and clarity	Report is well written in terms of format and presentation	Report is very well written and presented in a clear and concise manner.
Report Evaluation 2: (Technical Content, 50 marks) • **Discusses all processes involved to solve the problem** • **Presents the background theory clearly for the reader for the different techniques used in solving the problem.** • **Discussion on the evaluation metrics** • **Commented source code**	Technical content meets the bare minimal standard required for a technical report. Results are presented at a bare minimal. Source code is commented at a minimal level.	Technical content presents well on all the methodologies, processes and contains enough background knowledge to help user understand the overall solution to the problem. Results are well presented with well accompanied performance metrics Source code is well structured and commented.	Technical content is of highest quality and presents excellently on all the methodologies, processes and contains enough background knowledge to help user understand the overall solution to the problem. Results are well presented with well accompanied performance metrics and limitations of the results are well discussed Source code is of high quality and well commented as well as structured.