THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

STAT4609 Big Data Analytics

Second Semester, 2019-2020

Group Project

Purpose:

This group project aims to provide students with more practical experience of using big data analytics learned from the class on a real-life problem. You will learn how to formulate a problem and apply relevant big data analytics tools in practice. The project will account for 30% of your final grade.

Project Teams:

Each project can be done in a team of up to 4 students. Each team must appoint a team leader and should choose a team name. You should form a team yourself and inform me by completing the online registration via Moodle on or before **April 9, 2020**.

Details of Projects:

1. Identify a project topic and determine the objectives of the group project. The team should use one of the Kaggle datasets listed below:
   - TripAdvisor Image Restaurant
   - 1002 short stories from project Gutenberg
   - MUSAE Facebook Page-Page Network
   - Social Network Fake Account Dataset
   - Autistic Children Data Set
   - Fruits 360
   - COVID-19 X-ray Dataset (Train & Test Sets)
   - Coronavirus (COVID-19) TV Coverage
   - WebMD Drug Reviews Dataset
   - Title and Headline Sentiment Prediction

   You can collect raw data by web scrapping if you do not find any of the above datasets interesting.

2. Study and understand the dataset by exploring it. Pay attention at the quality of data (e.g. any missing values), the meaningful features, data distribution, and the types of variable values. Perform necessary data cleansing and transformation.

3. Choose appropriate big data analytics tools and develop the necessary algorithms/models upon the dataset. At least one of the big data analytics tools must be recommender system, social network analysis, deep learning or text analytics.

4. Find tune the algorithms/models and try to explain the outcomes of the big data analytics as much as possible with regard to the project objectives.

5. Each team must submit a project progress report (in .pdf format) and python code (including outputs) (in .ipynb format) via Moodle by **April 27, 2020**, and a presentation file (in ppt/pptx or pdf format) via Moodle by **May 10, 2020**. All file names must begin with team name. eg. WineTaster_Report.pdf

Project Progress Report (10%)

The progress report should comprise 7 to 15 pages of A4 size paper (single line spacing), including:

- A cover page including
  - Title of the project (at most 15 words)
  - Name of the team
  - List of team leader and team members (with student UID)
- Objectives of the project (presenting the background of the project, the problem of the study, and project objectives).
- Data Sources: Description of data and data preprocessing (including the source of data, the description of major features/variables, the quality of the data, and appropriate data preparation).
- Preliminary Findings (including the preliminary results of your analysis)
- Conclusions and Future Plans (describing the problem to be encountered and how they might be solved, etc.)
- References (such as research articles, books, book chapters, websites, etc.).
- Appendix (excluded from the 15 pages)

Grading of project progress report will be based on problem formulation, data description and analysis, use of big data analytics tools, interpretation of findings, originality and creativity, content and organization.

Project Presentation (20%)

In the oral presentation, each team will spend 12 minutes for presentation and 8 minutes for question period using Zoom. The team leader will be responsible to share the screen for presenting slides and to control the slide show in the Zoom meeting. Each team member should contribute in the presentation. Basically, the presentation should introduce what the project is about, how the big data analytics tools work to address the project objectives, interpret the results of the analysis, and finally the conclusions and reflections from the project. Grading will be based on the content of presentation, oral presentation skills, visual aids of presentation, and questions & answers.

Important Dates:

- Team registration deadline — April 9, 2020
- Submission deadline of project progress report and python code — April 27, 2020
- Feedback from instructor — The week of May 4, 2020
- Submission deadline of project presentation slides — May 10, 2020
- Project presentation — May 11, 2020