National University of Singapore
School of Computing

SWS3023 Web Mining
Group Project

To study and research an interesting business analytics problem. In your group, you are to discuss and decide on an interesting problem that requires some form of analytics. Since this course is on web mining, **the chosen topic must involve mining your datasets from the web and/or combining multiple datasets**. A list of possible topics can be found at the end of the document but you are encouraged to propose your own topic.

Tasks

- Form groups of 4-5.
- Choose an interesting business analytics problem.
- Submit a project proposal so that the teaching team can better advise you.
- Apply the CRISP-DM process, focusing on the first 5 steps:
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
- Prepare the presentation slides for the project presentation.

Deliverables

1. Project Proposal
2. Poster + video (optional) + presentation during Project Showcase Day

Project Proposal

- Decide on a topic. (Refer to the bottom for a list of possible topics or suggest your own topics)
- Choose a group name and indicate the members of the group on the project proposal.
- Write a **1-page** project proposal describing the topic you have chosen.
- **One** person in the group is supposed to submit the proposal in pdf format to LumiNUS Files. (**Project Proposal** folder)

Poster Presentation

- Each group has to prepare a poster
- The poster should be similar to a poster used in academic presentation and has to be uploaded to SWS_ALL to be viewed by all students of SWS. You can also prepare a video to showcase what the project is about (optional).
- During the Project Showcase Day, a timeslot will be allocated for each group for evaluation by the lecturer and TA. Each team will be given some time (roughly 10-20 mins) to present their project (slide presentation).
- In the presentation you should include the following items (where applicable):
  1. Introduction
     - Introduce what this project is about.
  2. Problem/Data
     - Describe the business domain and the data used.
  3. Analysis Process
     - Describe the steps to undertake this project.
     - Discuss any challenges and the steps taken to address these challenges (if applicable).
  4. Data Collection
     - Describe the approach of mining the web content (if applicable)
     - Describe the approach of combining multiple datasets (if applicable)

5. Data Exploration
    - Describe the various data exploration tasks performed (e.g. data visualization, descriptive analysis, etc).
6. Data Preparation
    - Describe the data preparation process such as data cleaning (if applicable).
7. Experiments
    - Describe the modeling techniques used (e.g. regression, classification, clustering).
    - Compare and contrast different modeling algorithms.
    - Explain and discuss the experimental results.
    - Discuss any limitations of the approach.
8. Discussion and Conclusion
    - Discuss any learning points.
    - Suggest possible future improvements.

Consultation

- There are 4 planned consultations with the lecturer
    - 1st consultation: **Project formation and project scoping**: to give suggestions of potential topics and scope of the project taking into consideration the limited time.
    - 2nd consultation: **Formulate strategies to mine sites**: discuss the strategies on how to go about mining relevant data.
    - 3rd consultation: **Project fine-tuning (analysis):** to advise on the experiment setup and experimental results.
    - 4th consultation: **Project fine-tuning (presentation)**: to advise on how to present your finding and your project.
- Ad hoc consultations with the lecturer
    - Teams can further arrange ad hoc consultations with the lecturer if you require additional advice along the way.
- Ad hoc help from TA
    - Teams can also approach the TA for additional guidance.

Topics

The following are just some suggestions of possible topics. **You are free (and encouraged) to propose any interesting topics and to use any techniques not covered in the course. Try to propose a project that has practical applications.**

Email Spam Classification

Build a system/propose an approach to classify emails to be either spam or not spam.

Opinion Spam Detection

Nowadays, people are using product reviews to aid them in purchase decisions. Some companies have intentionally fabricated fake reviews to either boost the rating/sentiment of their products or to damage the rating/sentiment of their competitor's products. Build a system/propose an approach to detect opinion spam.

Sentiment Analysis for Weibo Posts

Build a system/propose an approach to classify Weibo posts to be either positive/negative/neutral.

Sentiment Analysis for Reviews

Build a system/propose an approach to classify product reviews (or movie reviews etc) to be either positive/negative/neutral.

Movie Rating Prediction System

Build a system/propose an approach to predict the rating of a movie.

Election Prediction System

Build a system/propose an approach to predict which candidate will win an election.

Foreign Exchange Analytics (when to buy/sell)

Build a system/propose an approach that will make use of historical data to decide when to buy and sell on the foreign exchange market and attempt to maximize earnings.

Stocks Exchange Analytics (when to buy/sell)

Build a system/propose an approach that will make use of historical data to decide when to buy and sell on the stock exchange market and attempt to maximize earnings.

User Profiling of Weibo Users

Build a system/propose an approach that will classify the demographics of a Weibo user by gender, age group, etc.

Social Network Analysis Classification

Build a system/propose an approach that will make use of the social network linkage to provide useful analytics of a user.

Forecasting/Prediction Systems

Build a system/propose an approach that will perform forecasting/prediction.

Classification Systems

Build a system/propose an approach that will perform classification.

Recommender Systems

Build a system/propose an approach that will produce recommendations.