CMPE 239 – Web and Data Mining

Fall 2013

Group Project

Project Deadlines

- 9/27: Project proposal DUE
  - Students must have picked the type of project they want to do by this date – no change of project type will be allowed.
- 10/2: Project type A – Student Q&A (organized by IBM)
- 11/8: Project type A (IBM Watson TGMC) – implementation phase ends
- 11/21: Project type C – implementation phase ends
  - Project type B – deadline depends on competition
- 11/21: All projects – project report & presentation slides due on Canvas
- (Section 1) 11/26, 12/3: Project presentations
- (Section 2) 11/21, 11/26, 12/3, 12/5: Project presentations

* This is a group project (groups of 2-3 students). However, each student will be graded individually (please refer to the green sheet for details).

* A list of online resources (links to data mining tools, data sets, etc.) is posted on Canvas.

* All project documentation (project proposal, project report, datasets) should be submitted to the appropriate folder in Canvas (one per team is enough).

Project proposal

Should be 1-2 pages long and include:

- Student names
- Project type (A, B, C)
- Project title
- Project description
- Proposed methodology/techniques/resources/datasets to be used

CMPE 239 Project

The project should be a web or data mining application. You should integrate/combine one or more web/data mining techniques (possibly using available tools) into a Web or data mining solution for a specific problem.

You may use publicly available (open source) code in your application, but this should be explicitly stated in your report. You can choose any programming language to implement your project. For project type C you are encouraged to use Mahout on AWS or develop your own code.

Project groups may require access to AWS. Please send the instructor an email including your names. There are strict rules on the usage of AWS credits – *misuse might result in no grade in this class!* Please refer to the green sheet for more details on the policy.

If you intend to use parts of a previous or existing project from other class (or your MS project), you need to talk to the instructor in order to get approval (not guaranteed).

Project Type A: IBM Watson - The Great Mind Challenge (TGMC)

As discussed in class, this is a very strictly defined machine learning project. You have to work in groups of 2 and are free to use any data mining tools/algorithms you deem appropriate for the problem of confidence ranking of potential answers to a Jeopardy! Question (this includes data preprocessing, as well as fine-tuning and combining multiple algorithms and extensive experimentation).

IBM will provide to the students the dataset and access to their system for submissions. They will also offer a Q&A session before the beginning of the competition. More details can be found in the FAQ document provided under the Resources' page on Canvas.

Given the competition's timeline, students will have to complete their project's implementation by the set date. They can then spend the rest of the time writing the report and preparing for their presentation.

The final placement in the competition will only affect the grading process in helping the instructor identify teams who did not spend too much effort in this project. If, for example, all teams place around the middle except for one team that places closer to the bottom, this is a good indication that this team did not try enough!

Project Type B: Kaggle Competition

This project involves developing a data mining solution that will be submitted to one of the Kaggle competitions that have deadline before the 11/21 deadline of our project.

It is the students' task to identify a competition and include the details in their proposal. The instructor will judge whether a) the project's topic is relevant, b) the project's difficulty level is within scope, and c) the number of people in the team is adequate for the proposed project. If any of the above is not true, the group will need to work on a different project type.

All other details of this project in terms of deadlines, placement, deliverables, etc. are similar to project type A.

Project Type C: Web mining application using cloud technologies

In this project the team is responsible for proposing a topic to work with. Even though I am providing you with a list of suggestions, it's usually preferable (and more fun!) to pick a problem "closer to your heart". Just make sure that there's data available that you can use!

An alternative would be to first look at the available data sets (posted on Canvas) and then try to formulate a problem/application around them.

A few sample projects are (these are provided mainly to give you an idea of the scope of the project):

- Design a program that will classify review comments (on products, blogs, etc.) as spam/fake using sentiment analysis
- Design a recommendation engine that will generate recommendations taking into account other factors (e.g. social network/connections)
- Design a program that will use/analyze data from social networks. Note that there are so many possible ways of doing this – e.g. analyze tweets in different US states and find the "happiest" folks, then show results on a map.
- Develop your own system that will mine the social graph of a network and generate interesting patterns (e.g. important users).

Depending on the scope of the project and the number of persons in the team, this project involves implementation of the back-end, extensive experimentation and/or development of a front-end application (GUI).

Deliverables

The projects will be evaluated based on significance, design, correctness, documentation, and appropriate evaluation/testing. You will need to submit your project distribution files and documentation, *both electronically and in hard copies* (CD & printout of the report). Your project documentation should contain the following components:

1. A description of your system, including general architecture, interaction between the components, and specific techniques and algorithms you used. A project report template will be provided to you later in the semester.
2. An evaluation of your system demonstrating its correctness and functionality. If you used any outside sources in your implementation, please clearly indicate which sources, and how and where they were used.

Your project distribution files should contain the following:

1. Complete source code (be sure that your source code is fully documented and easy to read).
2. Binary files (e.g., executables, DLLs, Class files) or other components necessary to run your program.
3. Readme file containing instructions on how to compile, install, and/or run your program.
4. Any test data used for evaluation of your system.

Evaluation

Each student will be evaluated individually. Your final grade will be calculated based on:

- The overall quality of the project report
- The overall quality of the project prototype
- The individual presentation
- The individual participation in all project aspects

APPENDIX – Sample of previous CMPE 239 project topics

Please note that this is for your reference, in order to understand the expected scope and extend of work (most of the projects below were developed by groups of 3 and included a GUI and experimental evaluation). However, I won't approve projects that cover the exact same topics (combination of idea & dataset) as listed below.

- Movie recommendation system using Matrix Factorization (dataset: MovieLens).
- Clustering analysis of Bay Area restaurants (dataset: Yelp)
- Social information based Recommender system (project used social connections to make restaurant recommendations to users. Datasets: Facebook and Yelp)
- Personalized news feed using Facebook data (datasets: Facebook and Yahoo! News)
- Sentiment analysis on Twitter data (dataset: Twitter)
- SMS classification (dataset: SMS Spam from UCI repository)
- Identification of spam/fraudulent advertisments on craigslist (dataset: craigslist)
- A Facebook-based movie recommendation system (dataset: Facebook and MovieLens)
- Song recommendation engine (dataset: Last.fm)
- Crime recognition and prediction system (dataset: SFPD crime incident database)