CSE 4334/5334 – Data Mining
Spring 2015 – Course Project

Due Dates:

- Project Proposal: 11:59pm Central Time, Thursday, March 5th, 2015
- Progress Report: 11:59pm Central Time, Thursday, April 2nd, 2015
- Final Deliverables: 11:59pm Central Time, Tuesday, May 5th, 2015

Demonstration:

- Sign-up sheet will be provided to you for scheduling demos during May 6-8 in front of the instructor.

Academic Honesty:

- You are encouraged to learn from projects/papers on the same dataset, and you have the freedom to use source codes publicly available.
- You must note explicitly, in both source codes and reports,
    - If a piece of source code is taken from and/or adapted from other sources;
    - If an idea is borrowed from and/or adapted from other sources.

  Missing such notes constitutes plagiarism.
- You are encouraged to discuss the project with other students/teams, but you are not allowed to share detailed ideas and source codes with each other.

Dataset:

All teams will use the same dataset from the Yelp Dataset Challenge: http://www.yelp.com/dataset_challenge. Click the button "Get the Data", fill out the form, and download the dataset.

You are free to use any other publicly available datasets together with the Yelp dataset, but the focus should be on the Yelp dataset and other datasets can augment it.

According to their website, the Yelp Dataset has:

- 1.6M reviews and 500K tips by 366K users for 61K businesses in 10 cities across 4 countries;
- 481K business attributes, e.g., hours, parking availability, ambience;
- Social network of 366K users for a total of 2.9M social edges;
- Aggregated check-ins over time for each of the 61K businesses.

Tasks:

You are required to participate in the Yelp Dataset Challenge. You are not required to make a submission to Yelp, although you are highly encouraged to do so.

You are required to analyze and mine the dataset and present the design, implementation, and results of your study. You have the full freedom to define the topic of your study. It can be classic data mining related tasks discussed in our course (classification, clustering, association rule, link analysis, graph mining, …) or those not discussed much in our course (collaborative filtering, prediction and regression analysis, sequential and time-series patterns, spatio-

temporal data mining, sentiment analysis, named entity recognition and disambiguation, record linkage, data cleaning, …). It can even be new types of analysis that gains knowledge and insights from data.

You can use any programming languages and techniques.

Required Submissions:

The course project can be done individually or in a team of 2 students. Each team must create a public GitHub repository (https://github.com/) and use the repository to host all materials related to their course project. We use the latest version of files before the deadlines, according to timestamps in GitHub, to evaluate your projects.

You are required to produce a report, a website, and a demo for your project. You are required to submit all source codes, documents, files, and data (except the original Yelp dataset).

1. Project Proposal: due by 11:59pm Central Time, Thursday, March 5th, 2015

1.1. Create a GitHub account for your team and create a **public repository** for your team's course project. The repository must be named YourGitHubUserName/6339Project.

1.2. Create a thread (and thus the first post of the thread) in the Blackboard discussion forum "Course Projects". The first post should simply provide YourGitHubUserName so that we can access the public repository YourGitHubUserName/6339Project you created for your project.

1.3. Place your project proposal as a PDF or Word file in the aforementioned GitHub repository. The project proposal should have 400-600 words. It should provide the following information (EVERY aspect needs to be covered):

- Project Title
- Team information: member names.
- Objective and overview of the project. Why is it interesting and significant?
- What are the data mining tasks you will perform on the Yelp dataset?
- What do you plan to deliver at the end of the semester? How would you present the outcome of your study? How would you demo your work? What will you place on your website?
- What are the challenges in this project?
- How do you plan to address the challenges? How would you design and implement the solution?
- How would you evaluate the efficacy of your solution?
- How would you partition the tasks and coordinate among team members?

2. Progress Report: due by 11:59pm Central Time, Thursday, April 2nd, 2015

Each team needs to place the following files into their GitHub repository, by the deadline.

2.1. Source codes, documentation, data, website, figures/charts, and all relevant materials that have been produced so far for the project.

2.2. A report in PDF or Word. The report should have 700-900 words. It should provide the following information.

- Current status, with the following information clearly identified and explained:
  - project objective and tasks
  - deliverables
  - challenges of the project

- methods and algorithms designed and documented in details
- initial implementation
- evaluation plan
- change of plan since the project proposal
- difficulties encountered and how you addressed them (or failed to address them), other issues

❖ Tasks to be accomplished:

- what are left to be done?
- how do you plan to finish the project?
- expected challenges ahead

3. Final Deliverables: due by 11:59pm Central Time, Tuesday, May 5th, 2015

Each team needs to place the following files into their GitHub repository, by the deadline.

3.1. Report in PDF or Word. The report should have 3000-5000 words. The report should elaborate on at least the following aspects of the project:

❖ Motivation and objectives
❖ Data mining/analysis tasks tackled
❖ Design of methods
❖ Implementation of methods
❖ Results and Evaluation
❖ Presentation/Visualization of the Outcome
❖ URL of your project website

3.2. Files of the website of the project. The website should elegantly and interactively present the study in the project.

You can host your website on GitHub (by GitHub Pages https://pages.github.com), UTA's Omega server, or any other server.

3.3. Source codes, documentation, data, website, figures/charts, and all relevant materials that have been produced for the project.

Evaluation:

Your projects will be evaluated on the following aspects:

❖ Topic: novelty, interestingness, significance
❖ Technical approach: soundness, thoroughness, originality
❖ Execution: quality of implementation, quality and significance of results
❖ Presentation: website, report, and demonstration

Resources:

You can check out what other students have done using the Yelp dataset: http://www.yelp.com/dataset_challenge. On this page, you will find a list of possible tasks under "The Challenge", notes on the dataset, some examples from Yelp (https://github.com/Yelp/dataset-examples), projects of previous winners, and a list of student papers using this dataset (https://scholar.google.com/scholar?q=citation%3A+Yelp+Dataset&btnG=&hl=en&as_sdt=0%2C5).