Data Mining Class Project Final Report

## Motivation and objectives

## The aim of the project is to attempt the prediction of a business' category by using its rating. Consequently, a high prediction accuracy would suggest that the type of a business correlates positively with the kind of ratings it gets.

## Data mining/analysis tasks tackled

As shown in the project proposal, the following data mining tasks were performed:

1. Extracting the information about Businesses form the dataset.
   1. This will involve writing a program to open the dataset files, read the information needed on the businesses and store that information
   2. The information to be extracted includes
      1. For Businesses
         1. Business ID
         2. Zip Code
         3. Categories
         4. City
         5. Review count
         6. State
         7. Stars
         8. Type
      2. For Reviews
         1. User ID
         2. Review ID
         3. Stars
         4. Business ID
2. Grouping the businesses based on business types.
   1. For each business named in the dataset, a class called Business was instantiated and the results were stored in a java HashMap for effieicent retrieval.
   2. For each category of businesses, a class called Category was instantiated and the results were stored in a java HashMap for efficient retrieval
3. Sorting and evaluating the different rating types for each business type.
   1. The rating for each business type will be divided into five groups based on the number of stars they give.
4. Producing numerical measures of how each business type is being rated.
   1. The numerical measures used include:
      1. Number of each star rating
      2. Percentage of each star rating in the total of a business type’s ratings
      3. Mode of each star rating
      4. Median of each star rating
      5. Mean of each star rating
5. Predicting the type of a business from its ratings.
   1. Several businesses (a total of one quarter of all the surveyed businesses) were set aside for evaluating the efficacy of the classification. Using Naïve Bayes algorithm, these businesses will be classified as one of several business types based on:
      1. Their average rating
      2. Their median rating
      3. Their mode rating

## Datasets collected and Used

* + - 1. The list of businesses from the Yelp Dataset
      2. The list of reviews from the Yelp Dataset

## Design of methods

All methods were part of one of three classes:

* + - 1. The Naïve Bayes class, which controlled the overall flow of the program
      2. The Business class, which stored information about each business surveyed
      3. The Category class, which stored information about the different categories that businesses in the dataset belonged to.

## Implementation of methods

The most elaborate method, called classify(), belonged to the Naïve Bayes class and was responsible for the running of most important logic of the program. From its viewpoint, all other methods were merely supporting its function. It implemented a Naïve Bayes classification algorithm with data collected from the dataset and processed by other methods.

## Results and Evaluation

The results show that the data in the UT Arlington SFS and RateMyProfessor datasets have graphs that are roughly alike when the mean of the clusters are plotted against the number of professors in that cluster. Both datasets show the majority of professors belong to the second largest cluster. However, while the clusters for the SFS dataset range from 63 to 96 percent of the maximum score, the clusters for the RMP dataset range from 38 to 90 percent of the maximum score.

This greater range in scores for the RMP dataset suggests that scores for measures used by RateMyProfessor.com produces a different kind of result from those used by the UT Arlington SFS survey. In other words, the less descriptive terms like "clarity" used in the RMP dataset produces a greater range of responses than the more descriptive prompt: "The instructor used teaching methods that facilitated my learning" used by the UT Arlington SFS dataset.

The result of treating both datasets together is less significant due to a marked difference in the number of professors in the RMP dataset compared to the SFS dataset. The gap between the smallest and largest clusters remains the same as that of the RMP results, but the look of the graph changes and largest number of professors belong to the highest cluster. This is likely due to the fact that a large amount of SFS data points belong in that cluster.

## Presentation/Visualization of the Outcome

The presentation was cancelled so there is no presentation information for the project.

## URL of your project website

The website for the project can be found at: https://sites.google.com/site/tracyogunidataminingclass/