Tracy Oguni

Class Project Progress Report

Data Mining

April 2, 2015

# Current Status

The following is the current progress on the class project.

## Project Objectives

As stated in the project proposal, the project objective is as follows:

The objective of the project is to survey the different kinds of businesses in the Yelp Data Set and see which are more likely to get positive or negative reviews. For example, do people generally rate Dentists lower than restaurants? If so, why? This involves sorting and classifying the data from the Yelp dataset, and predicting the type of rating a business will get based on its type. The results will be presented in different categories, including business type and city.

## Tasks

As shown in the project proposal, the following data mining tasks will be 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. This will involve simply putting the data into different software arrays or parts of an array based on the type of business they refer to as described by their categories.
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 to be used include:
      1. Number of each star rating
      2. Percentage of each star rating in the total of a business type’s ratings
5. Predicting the type of a business from its ratings.
   1. Several businesses will be 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. The number of ratings they have
      2. The proportion of each rating (one to five stars) that they have
      3. Their average rating

## Deliverables

The deliverables for this project will be a website, text report, and graphs showing

1. Average rating and percentage of each type of rating (one to five stars) of each business
2. Average rating and percentage of each rating type by state
3. Average rating and percentage of each rating type by city
4. Total accuracy rate in predicting a business type from ratings
5. Total accuracy rate of predicting a business type from ratings by city and state.

## Challenges of the project

The major challenge of the project is in finding a way to accurately predict a business’ type from its ratings.

## Methods and Algorithms

The only important algorithm is the classification method. It will be an implementation of the Naïve Bayes algorithm. The algorithm will use number and type of ratings, as well as city, state, and business categories if they improve the prediction accuracy of the algorithm.

## Initial implementation

No implementation has been done beyond the initial set up of the working environment and creating a java class. All the work so far is on the most important part of the project: deciding on and working out the specifics of a classification algorithm

## Evaluation Plan

As explained in the project proposal, the evaluation plan involves dividing the data into training and test sets and using the test set to confirm the performance of the algorithm.

## Changes since Project Proposal

The only change since the project proposal has been in including a description of the results by city and state.

## Difficulties Encountered

The only difficulty encountered was in choosing a classification algorithm. The Naïve Bayes algorithm was chosen for its relative speed, simplicity, and the fact that the student has experience implementing it for a previous class assignment.

# Tasks to be accomplished

## What is left to be done

The planning phase of the project has been completed. In the coming weeks, the implementation will commence as listed in the project tasks.

## Finishing the Project

The project will be finished before the final report, and packaged as per the delivery instructions

## Expected Challenges

There are two expected challenges. Firstly the Naïve Bayes implementation might provide a very low prediction accuracy. If this happens, a different classification algorithm, (like Nearest Neighbor classification), might prove better.

The second challenge is that the student has five different projects this semester, including two for this class. The requirements for the projects have been kept simple as a result, so the workload is not too much, but there is always a risk that a project will require more time than can be spent.