CMPS 242 Fall 2016: Project Plan
October 13, 2016

Greeshma Swaminathan (gswamina@ucsc.edu)
Neha Ojha (nojha@ucsc.edu)
Jianshen Liu (jliu120@ucsc.edu)
Alex Bardales (abardale@ucsc.edu)

Focus Area:

The key idea of our project is to recommend restaurants to users based on the observations, user ratings and reviews provided by the Yelp dataset. We will be using Python as a coding language.

Data Description:

We plan to base our analysis on five datasets, namely, business, user, review, tip and checkin. We have performed preliminary analysis on these datasets by dividing each of the datasets among the group members. We will briefly describe what we analyzed in this process:

Business dataset - There are 85901 individual business details in the dataset. There are 16 attributes out of which 'type' can be dropped. From a preliminary analysis there are 26729 restaurants out of which 20125 are 'open'. Initial idea is to just work with these restaurants and ignore the other businesses. The 'review_count' field may be used to further reduce this dataset or it can act as a weight to estimate the confidence in 'stars'.

User dataset - There are 686556 unique users' information in the dataset. There are 11 attributes to describe these users. The data has users who have started yelping since October, 2004 and range upto July 2016. Since some of the attributes are nested like "compliments" and "votes", features have been generated and new columns have been added to the dataset. We haven't dropped any columns in this dataset yet but, have identified some attributes which can be dropped before preprocessing the data further. The maximum number of reviews given by any user is 10897 and 13 users have given no reviews.

Review dataset - There are 2685066 reviews in this dataset, 1132610 reviews with 5 stars, 674636 reviews with 4 stars, 321700 reviews with 3 stars, 224334 reviews with 2 stars and 331786 reviews with 1 star. Here is the part of detail of what we have get from initial analyzing.

- After sorting business id by the number of reviews, the most popular business has 5558 reviews, the second most popular business has 4531 reviews, and third most popular business has 4333 reviews.
- After sorting users by number of reviews on the same business, the most loyal user to the same business has 31 reviews, the second one has 19 reviews, and third one has 18 reviews.

- Each review has three votes attributes, they are “funny”, “useful” and “cool”. We tried to sort business by number of votes from reviews, and we found the hottest business gains the most 12676 votes from its reviews, the second hottest business gains 10572 votes and the third hottest business gains 10118 votes. Surprisingly, the third hottest business isn’t belongs to the top 10 businesses with the most number of reviews.

The Tip and Checkin datasets are relatively smaller with fewer attributes. We plan to incorporate any useful information from these datasets into the above mentioned datasets.

Machine Learning Algorithm and Evaluation Criteria:

We intend to use an algorithm to offer restaurant recommendations to users based on the similarity of previous user review pattern. Our initial idea is to implement Matrix Factorization, as it known to be effective in capturing latent interactions provided in the dataset. We will split our dataset into test, train and validation sets. We plan to use cross verification on the test data as the evaluation criteria for the algorithm.