CS838 Project Proposal

Team: Xiaodong Wang, Bozhao Qi, Shuang Wu

(a) The title of the project
"A new Semantic approach on Yelp review-star rating classification"

(b) A short description of the problem you propose to solve

In this project, we want to show if our modified classification model could give higher precision and recall for review-star rating based Tips, Users, Reviews Business and Check-in tables from Yelp dataset. We may crawl additional info from some web pages and try out different machine learning algorithms to show how well it would boost the overall prediction performance. In addition to the machine learning mission, we will dig into data sets and show more insights and interests of the restaurant industry locally.

The traditional approach is using the sentiment analysis of reviews in terms of the number of positive, negative, negation words in the comments and total length of the review to test the rating star prediction accuracy. In our project, we will consider that each user has his/her own preference which might be extremely skewed from each others. Say one person regularly give statically "five-star" comments but with "four-star" rating owing to his/her specially high connoisseurship. Under such situation, we want to write our own optimizer on the individual level to weigh down/up users' rating according to their historical records in the effort to improve the precision and recall value while reducing the RSS value. we will use lexicons compiled by UIC professor Bing Liu to score the sentiment of words in the reviews. Furthermore, there are couple of subtasks we want to illustrate in this project.
1. Building up the word cloud for each star reviews
2. Figuring out the most successful business model for several different major cities across different countries.
3. Presenting some business and seasonal trends.

(c) a brief outline of how you will approach the problem, and how you will evaluate your results. Do not forget to include a list of references!

We will firstly join a new data set derived from the original Yelp datasets, and then split it into training and testing subsets by 60/40. We plan to use stratified sampling and continue to apply 10-fold cross validation using six different classifiers on training set. Classifiers in Python sklearn package will be used to run Random Forest, Decision Tree, Support Vector Machine, Naive Bayes, Linear Regression and Logistic Regression on our datasets. The residual sum of square errors for the precision and recall will be used to reduce the effect of skewed rating distributions. Our own functional learner will be applied onto those machine learning algorithms. During the whole process, we plan to use Spark to process all the big data.

Reference: "Oversampling with Bigram Multinomial Naive Bayes to Predict Yelp Review Star Classes" Kevin Hung and Henry Qiu, University of California, San Diego.