Progress Report

Team:

Mudit Pradhan (1001168746)

Jvalin Dave (1001115870)

Out of the various data mining tools we are learning Orange data mining tool. Along with that we are also doing data cleaning and other tasks in IPython notebooks using different tools of Anaconda. We might use Orange just for data visualization.

We tried various ways to open the files and store it. One of the way was to convert it to csv and then do data cleaning. The other way to open the json file was to convert it to python dictionary and then to make data frames using pandas and use pandas function to process it further.

First task of ours is to get the review text of restaurants from all the businesses. We have made dictionaries that contains review items and business items. Now we are working on getting all the review text using this two dictionaries. We would be removing stop words from the reviews. We might also stem them.

The other task that we are working individually are:

1. Mudit Pradhan:

   Based on choice of the location and restaurant type desired by the owner, the trend in the popularity of that kind of restaurant in the location specified. This could be accessed based on number of useful reviews, and can be represented on X axis as time and Y as popularity measure. With this we would also plot popularity of other successful restaurant businesses in that area to briefly show business competition.

   We will also find the localized tastes, flavors, products that are most liked by the people in that area. This would be done by separating the useful reviews based on occurrence of topics related to tastes, flavors and products. These reviews found will be mined further to find out local tastes, flavors and products. This would be done by finding the relative subtopics through Latent Dirichlet allocation model (LDA) which is based on Bayesian inferences. We will represent the data found through pie chart. This data could be used by the user to decide on restaurant menu.

2. Jvalin Dave:

   My goal is to analyze what are the sure shot things that will lead to more stars.

   I am working on getting the nltk libraries that can help me classify positive and negative reviews from the review text of review.json file and also for further analysis. I am implementing tagging and categorizing the words using nltk. I am trying to get names of food items along with adjectives from the text for each positive review and negative review. The food items might also be useful for trends in the market. I am also trying to get the nouns from the reviews that are associated with the adjectives and then represent them as sparse representation using sklearn tools. I think that might be useful for the importance of the common terms that help in predicting positive review.

The other thing is to classify using price range and location as a set of parameter for the prediction of stars of the particular review.

Further I would try the combinations of few of these pair of words or combination of review text along with location and price range for particular rating of stars. To do this I would apply Naïve Bayes or LDA (algorithm) for each combination of adjective and noun to pred+ict the stars. Suppose for (good food |location | 4 stars) and (bad service| price range| 4 stars) .The final result would be pie charts of different stars with the different combinations (as mentioned above) constituting the part of it. It will be helpful to analyze the must requirements for getting more stars/business.

Our findings should be able to help the user to establish a new business, assess the success of proposed business in the area, check taste or product priorities of the people in that area and also be aware of best practices and reviewer criticism on well-established similar businesses in that area.