**Wesley Upham**

**Text Mining Where You Eat**

A good restaurant requires a collaborative effort of multiple variables. The restaurant must offer a well-suited atmosphere, fair pricing, and most importantly good food. Yelp tries to grasp these variables in its data, which I will use to try and accurately output a good restaurant recommendation.

I will use several types of data mining methods to analyze this data including simple descriptive statistics and keyword text summarization.

All code referenced in this report section refers to the file ‘final.py’ which is my completed yelp recommendation application. First, I will import the necessary libraries, connect to the database and then prompt the user with several input questions.

These input questions provide the set up for my descriptive statistics. I will take the user inputted information, then run SQL queries to find out the price range score and the star rating.

People who are looking for a new restaurant to try want something that will be on par with there expected food quality and around the same price they normally spend on a meal out. This is why I took the average of the both star rating and price. I then put these conditions into a query using a where clause to only extract restaurants that met the criteria that the user would expect from his/hers favorite restaurants.

I then grabbed all of the categories that the inputted restaurants were listed under in Yelp. I want to take into account the users cuisine preferences. So once I returned all of the categories from a specific restaurant, I added them to a list that I could later reference. I will only output restaurants that have one of the category types of the user’s favorite restaurants.

I then used nltk (Natural Language ToolKit) to find the 20 most frequently used words in the aggregate text review for the inputted restaurants. Finding these words allows me to find what topics are important to the customers who frequent the same restaurants as the customer. These topics are most likely important to the user who I am recommending a restaurant for so it is important to find these out.

Once I find these words I can perform another basic descriptive statistic , counting, to figure out which of these most frequently appears in our potential restaurants. Again the potential restaurants are called in a query which outlines several conditions including the users requested city, state, average star rating, and average price range. We then rank them by most frequent to least frequent. And recommend the most frequent.

Overall my project has some interesting findings. I think that using text reviews has a lot of potential for recommending food on yelp. Using a more advanced text summarization method or delving into text sentiment could be fruitful endeavors in finding more accurate potential restaurant recommendations. The nltk keyword summarizer does a good job of getting basic topics, which is why I choose it. With such a large amount of data I thought that this method would be best to avoid irrelevant words and would extract better topics by shear volume of data.