Christopher Yung

Homework 6

Project Outline

For my project, I plan to use the data provided by Yelp in the Yelp Dataset Challenge. In the dataset, we are given multiple JSON files to work with: *business, review, user, check-in,* and *tip*. These datasets are organized and documented very well on http://www.yelp.com/dataset_challenge. In addition, the datasets provided are very clean and dense, so there is not much noise or many missing values to cope with. There are no irregularities in the data in the *user* and *review* datasets.

My first hypothesis is that users with a broader vocabulary will have a larger spectrum of ratings as opposed to those who have a smaller vocabulary. This will be done by first going through the *user* data and filtering to get the user information of those who have more than 10 reviews by using the *review_count* property. Afterwards, I'll use the filtered *user_id* values and grab all the reviews for each respective user in the *review* data. From the *review* data, I can access the *text* of each review. For each user, I will read their review text and store their unique words, word count, and rating. From there, I will calculate the standard deviation of ratings for each user and be able to plot the data for unique word count versus rating standard deviation.

Another hypothesis to test the ratings of franchises in different types areas ranging from rural to suburban to cities. For example, pizza franchises like Papa John's, Domino's, and Pizza Hut can be found throughout the country in all three of those types of areas.

Specialty pizza places, however, are densely populated only in cities and decrease as the population density decreases. My hypothesis is that the closer you get to an area of local specialty shops, the lower the ratings are for franchises in the same food category. Papa John's may be popular in rural or suburban areas simply because it is one of the few pizza places nearby. However, their ratings may be much lower in the cities due to much wider selection.

To test this hypothesis, I will use the *business* dataset from Yelp in conjunction with the 2010 Census dataset that can be found at census.gov. The *business* data contains the *name* of the business along with its *state* and *city*. Using these properties, I can use the *census* dataset to look up the population for each city. After that, it is just a matter of plotting the average rating for a specific franchise in each city and plotting the average rating versus population density.