SI 618 Project Proposal —Zoey Li

Motivation:

- Using Yelp business and Yelp review data, we would like to understand if the star rating attributed to a restaurant or a cuisine business is actually trustworthy? If a restaurant just opened, it may be just out of luck that 4 out of its first 5 customers like this restaurant, and so its stars distribution has a large variance. To take into account of large variance, we use instead the coefficient of variance (CV), defined as the ratio of the standard deviation to the mean, as a measure of the quality of restaurants.

Datasets:

- Yelp business data and Yelp reviews data will be used in this project.

Dataset manipulation:

- We will use sparksql to join the Yelp business data and the Yelp review data on unique businesses ids. This way, we can explore variables relationships between the reviewers and the restaurants they reviewed.

Tasks:

- First we need to identify businesses in categories indicative of cuisine. From the Yelp business data, I will use mrjob to decompose the `categories` attribute and count the number of occurrences of each category. This way, we will be able to see which type of cuisine is most popular, that is, if you go out, which type of cuisine will you most likely to encounter.
- We will use sparksql to shrink the dataset down to cuisine businesses, and join this dataset with Yelp review data. This will give us reviews of cuisine businesses.
- We will use sparksql to calculate, for each restaurant, the mean and the standard deviations of stars given by the reviewers, and then calculate the coefficient of variance, by taking the ratio of the mean and the standard deviation.
- We will examine the distribution of CVs, and decide on a threshold that separates good vs bad restaurants.
- We will use sparksql to count the number of restaurants with CVs above the threshold in each city, and rank these cities according to the counts.
- Lastly, we will use sparksql to produce a ranking of cities by the number of restaurants with average star ratings above 3 (which we normally consider as above average quality), and then we compare this ranking with the ranking of cities by CVs.

Visualization:

- Display star distributions of different types of cuisine restaurants.
- Display a comparison plot consist of one map of best restaurants as measured by CV vs one map of best restaurants as measured by average stars; using longitude and latitude of each restaurants.