Capstone Project 2 - Milestone Report 1

Sentiment Analysis of Yelp Restaurant Reviews

Shirley Zhu

Problem Statement

When people go to a new restaurant, they usually don't know what to order. Yelp reviews can usually provide some ideas about what they might like, but there are often too many reviews and it takes a long time to read the reviews and make a decision. This project leverages machine learning to figure out what everyone prefers to eat based on the restaurant"s reviews.

My client will be Yelp and the users of this feature will be the restaurants that list their business on Yelp and the customers who go to eat in these restaurants. Yelp can build this feature because users would love using Yelp to figure out what to order. Yelp can help restaurants to roll out its "popular dishes feature" to help the customers in deciding what to order from the restaurant with whom they aren't familiar. Good rated restaurants can have bad dishes and ok restaurants can have really good dishes. The restaurants can use this information to advertise their popular dishes and improve their not so good dishes, so the customers will be more satisfied after eating there. This will boost the business of the restaurants. Yelp can charge the restaurants for this feature, and Yelp takeout business will grow too.

The first part of this project is the prediction of review stars using Natural Language Processing. The second part is the recommendation of highly rated food items for a restaurant.

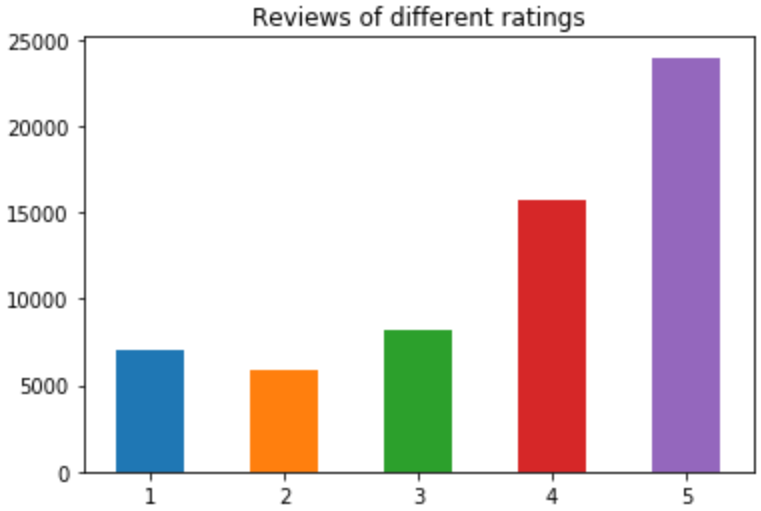
The data source is from the Yelp Open Dataset https://www.yelp.com/dataset. The dataset is in the json format. Business.json and review.json were used in this project.

Data Wrangling

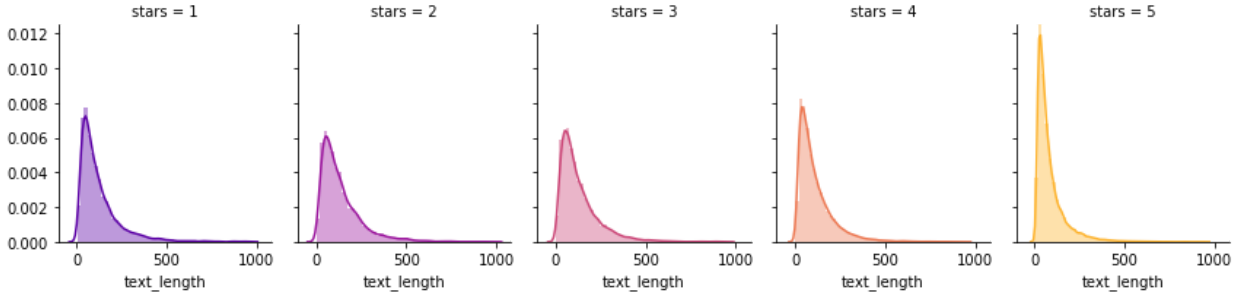
- Read the json file into a pandas dataframe. Because the size of the json is too big, the json was read by data chunks.
- Plot the distribution of review counts for businesses.

- Load the business json. Filter out the businesses that have “Restaurants” in their categories. Sample 2000 restaurants because the original dataset is too big.
- Create a dataframe that contains only restaurants and these restaurants have reviews in the review dataframe.
- There are 188593 businesses in the yelp_academic_dataset_review dataset. There are 57173 businesses that are restaurants. In order to extract the restaurants from the review dataset, merge the restaurants dataframe with the reviews dataframe on business_id.
- A sample size of 30000 was taken to make it faster.

Exploratory Data Analysis



The 5 star reviews have the most counts, followed by 4, 3, 1 and 2.



It shows that people who tend to review a business as good (4 or 5 stars) have shorter reviews (86 or 113 words), and the reviews that have poorer ratings tend to be longer words.

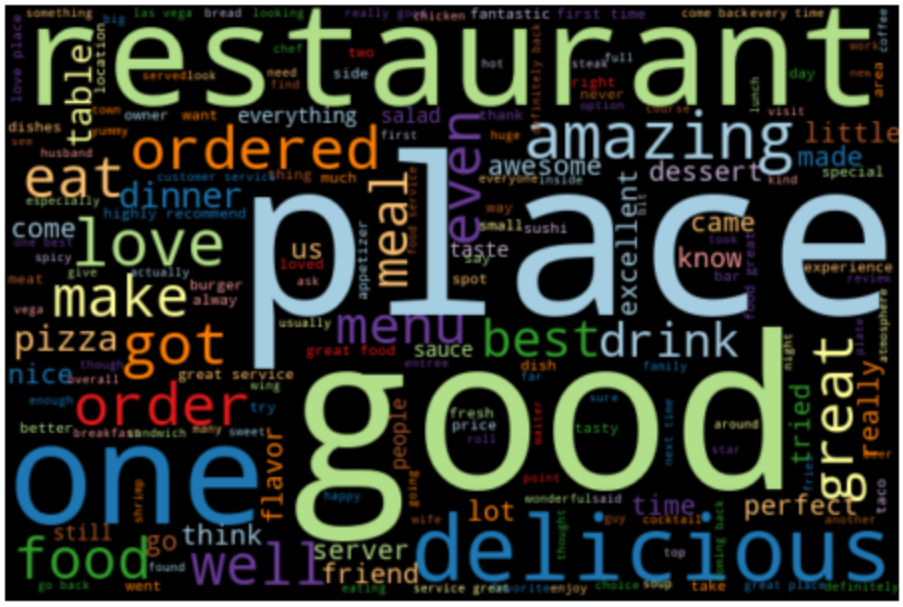
Inferential Statistics

One-way ANOVA was used to tests the null hypothesis that the distribution of the text lengths of all stars have the same mean length. The conclusion is the null hypothesis should be rejected because the p-value is 0, meaning the mean text lengths for different stars are significantly different.

Text Mining

Steps:

- Import the nltk, string, wordcloud libraries to do the text mining.
- Define a “cleaning” function to remove the punctuations, lowercase all case-based characters, and remove common English stopwords.
- Use the function “cleaning” to process the reviews.
- Create a bag of words by joining the words in cleaned text for each star.
- Generate the word clouds.



Word Cloud of Reviews of 5 Stars

Word Cloud of Reviews of 4 Stars

Word Cloud of Reviews of 1 Star

The 5 star reviews use positive words like good, great, love, delicious, amazing. The 4 star reviews have similar words as 5 stars, but not as many "love", "amazing", "best" as in 5 star reviews. The most frequent word for all reviews would be a neutral word "place". The most request words in 1 star reviews are neutral words such as place, food, time, said, total, table, order, service, never.