Joseph Insalaco

Professor M. Iyengar

CPE 551

May 18, 2021

CPE 551 Final Project

Github Link: <https://github.com/JInsalaco/CPE551>

Web scraping and sentiment analysis are two very popular topics for data analysis. The ability to scrape web page data and analyze the sentiment and reviews for restaurants, movies, etc. is very useful in data science. The overall goal of this final project has two parts, both of which allowed me to better learn the language and how to analyze data.

The first part of this project was to scrape review data from a website. For this project, I chose to scrape reviews from Yelp.com, a popular restaurant review website. Additionally, for the sake of getting interesting reviews, I chose a restaurant that had a lot of mixed reviews, and one that was overall very interesting. For this project, I chose to scrape the reviews for a popular college bar called Kenka in New York City. Using the requests and beautifulsoup python modules, I was able to pull the html for the web page and parse out the review and the star rating for everyone. I was also able to clean up the data and removed extra information from the extracted reviews. After the data was in a useable state, I exported the data into a .csv file.

The second part of this project was to clean the data and perform sentiment analysis on the reviews. I also wanted to see if there was a correlation between the sentiment and the star rating the customers gave. I was able to import the .csv file into a pandas dataframe, and clean the data by removing all stop words, and lemmatizing the words. This will help us better analyze the sentiment of the reviews. I performed sentiment analysis using TextBlob, and I grouped the calculated sentiment for each row into 5 different bins. These bins ranged from 1-5 and represented a scale of negative, somewhat negative, neutral, somewhat positive, and positive. I was also able to create a word frequency distribution, and word clouds based off the positive and negative words in our reviews.  
The results showed that many of the five-star or one-star reviews had sentiments that fell within the neutral or somewhat positive bin. Overall, my program took 100 reviews, of which 98 were neutral or somewhat positive according to their sentiment. I was also able to find that when we placed them in these bins, the average sentiment was a 3.44 and the average star rating was a 3.52 both of which are very close to a neutral score.