Stat 315 Final Project Report

Joseph Headley

May 2021

1 Introduction

My final project for Stat 315 involved combining Natural Language Processing's "Sentiment Analysis" technique and recommendations given through a User Based Collaborative Filtering system on a dataset of reviews provided by Yelp to recommend businesses to new users. The motivation behind this project was to acquire a better understanding of how entities such as Netflix and YouTube curate content for their users to such a fairly high degree of precision. Being able to accurately pinpoint new customers' content of interest is a vital part of enabling a business to properly market themselves and succeed in their industry and amongst their competitors. Thus finding a way to capture this kind of predictive power is a major topic of interest for many businesses and services.

Although before I can get into the details of this project, I would like to take some time now to explain the two essential concepts that form the crux of my project, Sentiment Analysis and User Based Collaborative Filtering. First, as previously stated, Sentiment Analysis is a technique that falls under the umbrella of Natural Language Processing. Sentiment Analysis utilizes a lexicon of words that are given either positive or negative associations as a basis upon

which to determine the emotional opinion of a user based on the words appearing in the user's statement that also match the words present in the aforementioned lexicon. It is essentially a computer's equivalent of having a human's capacity to understand the emotions and feelings inherent within a statement through a fundamental understanding of the words and the context in which the words are used.

Second, a User Based Collaborative Filtering system is a type of Recommender System that provides recommendations to users based on the preferences of other users. As the name suggests the key component of User Based Collaborative Filtering is that it provides a user with recommendations based on the similarity between the preferences of said user and other users. Should these users have similar likes and/or dislikes the User Based Collaborative Filter would then recommend the new user some of the content that other users with similar preferences liked but the new user has yet to experience. By combining the Sentiment Analysis process and the User Based Collaborative Filtering system, I coded a Recommender system in R to make recommendations based on a subset of a dataset of businesses published by Yelp.

2 Data

The data used in this project was a subset of a business dataset made available by Yelp. The data was initially parsed from two json files, the first containing information about businesses and the second containing information about reviews of those businesses. I had parsed the json files to acquire data for 100,000 businesses and 1,000,000 business reviews. The most important information for building the recommender system in the data that I parsed were the businesses' business ids and the reviewers' user ids, ratings, and reviews.

3 Methods

First, as I stated prior, I parsed two large json files to extract data on 100,000 businesses and 1,000,000 business reviews and stored them in two separate dataframes, called "businesses" and "reviews" respectively. In order to be able to properly analyze the data and use them for my recommender system, I decided to further truncate the data by making table data structures of the business_ids from the "businesses" dataframe and the user_ids from the "reviews" dataframe in R. These tables, named business_table and user_table, store every unique business_id and user_id and a tally of how often these ids appear in their respective dataframes. I, then, sort these ids in decreasing order of their tallies and store the 500 businesses and 500 users with the largest number of reviews given within two separate character vectors, "top500businesses" and "top500users".

I further reduce the amount of data to use by choosing the first 60 users from top500users and choosing the first ten businesses that each user reviewed that also appears in top500businesses. Through the use of several control flow statements (3 for loops and an if-else statement), I created a dataframe whose rows were comprised of each of the 60 users' user_ids, the business_ids for each of the first ten businesses that each user reviewed that also appears in top500businesses, and a quantity that I refer to as each business' "sentiment_score".

I compute this sentiment_score through a function I wrote, called GetSentiment(), which runs Sentiment Analysis on the reviews given by these 60 users for their first ten businesses and counts the number of words with positive and negative associations in the review that are present within the lexicon that is being matched to. For this project I chose to use the "Bing" lexicon for sentiment analysis. The computation of the sentiment_score is then made by dividing the

number of words that had a positive association by the number of words that had a positive or negative association and then averaging this quantity with the numeric rating given by the user.

Next, the aforementioned dataframe is recast into a matrix and split up via a train-test split, with 40 users being used to train the User Based Collaborative Filtering Recommender model and 20 users being used to test that model. A for loop is employed to make predictions (recommendations) for each user in the test set and these recommendations are finally placed into one final dataframe that appears as such.

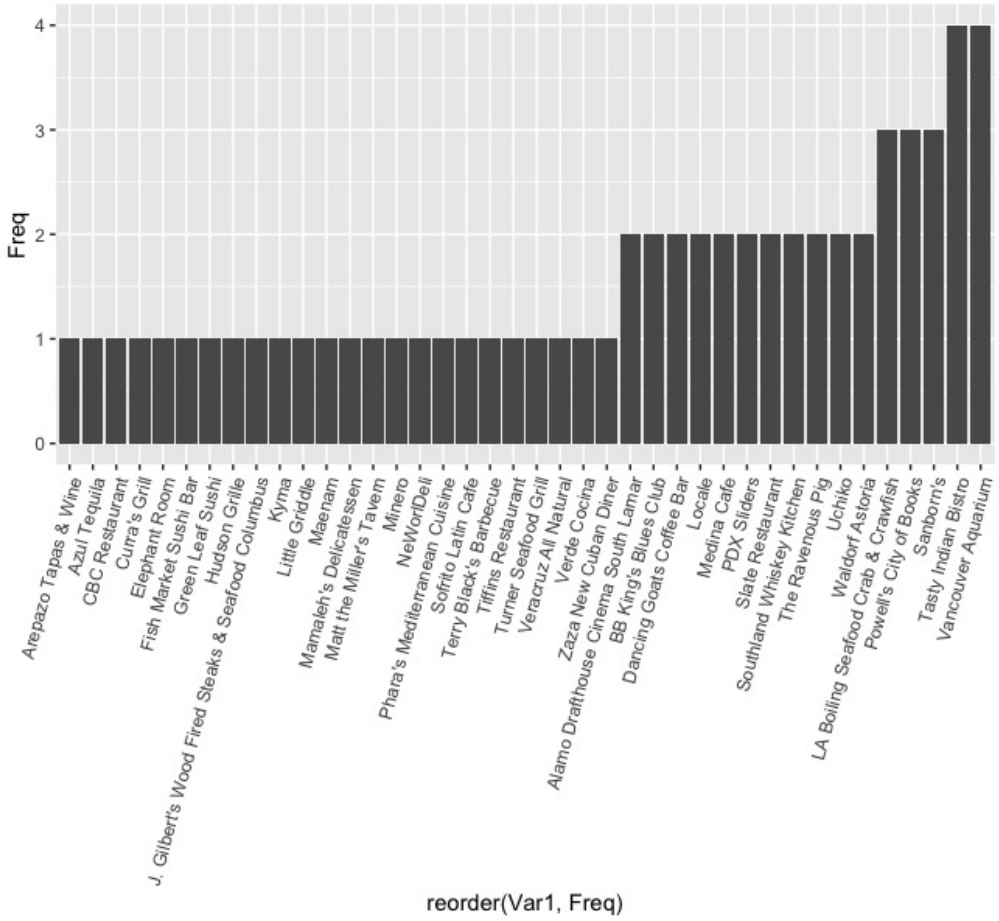
	V1	V2	V3	V4	V5
User 41	Dancing Goats Coffee Bar	LA Boiling Seafood Crab & Crawfish	Uchiko	Elephant Room	Powell's City of Books
User 42	Hudson Grille	Dancing Goats Coffee Bar	LA Boiling Seafood Crab & Crawfish	The Ravenous Pig	Slate Restaurant
User 43	Tiffins Restaurant	Locale	Phara's Mediterranean Cuisine	Azul Tequila	NeWorlDeli
User 44	Little Griddle	Tasty Indian Bistro	Sanborn's	Green Leaf Sushi	Alamo Drafthouse Cinema South Lamar
User 45	Vancouver Aquarium	Powell's City of Books	Tasty Indian Bistro	Maenam	Medina Cafe
User 46	Tiffins Restaurant	Uchiko	Locale	Mamaleh's Delicatessen	Turner Seafood Grill
User 47	J. Gilbert's Wood Fired Steaks & Seafood Columbus	Arepazo Tapas & Wine	Terry Black's Barbecue	Matt the Miller's Tavern	CBC Restaurant
User 48	Little Griddle	Powell's City of Books	Southland Whiskey Kitchen	Tasty Indian Bistro	Medina Cafe
User 49	Tasty Indian Bistro	Powell's City of Books	Southland Whiskey Kitchen	Green Leaf Sushi	Vancouver Aquarium
User 50	Alamo Drafthouse Cinema South Lamar	PDX Sliders	Fish Market Sushi Bar	Verde Cocina	Curra's Grill
User 51	Tiffins Restaurant	Veracruz All Natural	Locale	Sanborn's	PDX Sliders
User 52	Waldorf Astoria	Uchiko	Locale	BB King's Blues Club	Slate Restaurant
User 53	Little Griddle	Vancouver Aquarium	Powell's City of Books	Sanborn's	Alamo Drafthouse Cinema South Lamar
User 54	Dancing Goats Coffee Bar	LA Boiling Seafood Crab & Crawfish	Powell's City of Books	Sanborn's	Alamo Drafthouse Cinema South Lamar
User 55	Medina Cafe	Vancouver Aquarium	Powell's City of Books	Maenam	Sanborn's
User 56	LA Boiling Seafood Crab & Crawfish	Kyma	The Ravenous Pig	Slate Restaurant	Minero
User 57	Hudson Grille	Waldorf Astoria	BB King's Blues Club	Sofrito Latin Cafe	Zaza New Cuban Diner
User 58	Dancing Goats Coffee Bar	LA Boiling Seafood Crab & Crawfish	Tasty Indian Bistro	Vancouver Aquarium	Powell's City of Books
User 59	Dancing Goats Coffee Bar	LA Boiling Seafood Crab & Crawfish	Powell's City of Books	Southland Whiskey Kitchen	Medina Cafe

4 Results

In having created a functioning recommender system through a combination of Sentiment Analysis and User Based Collaborative Filtering, I decided to expound upon two possible results from this recommender system. The first result was uncovering any biases or tendencies in the recommendations given to the 20 users in the test set. The second result was to use this recommender system to make actual recommendations to users other than those present in

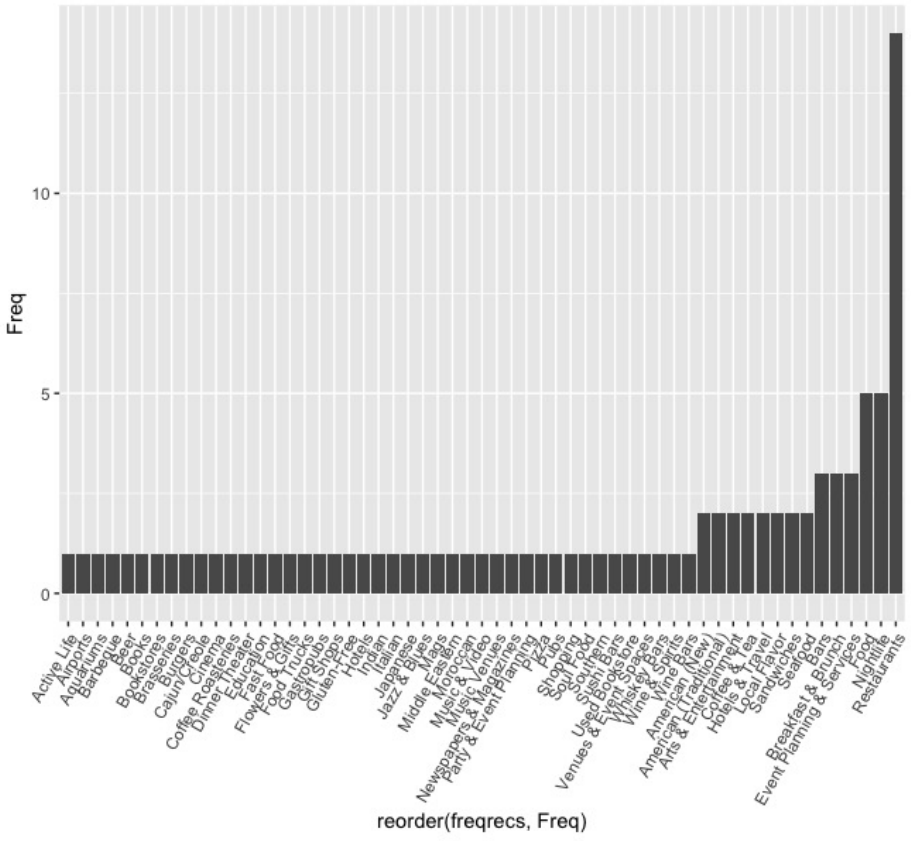
the initial Yelp dataset based on reviews that these users give for businesses among those that were included in the data that the recommender system was trained on.

For this first result, I tabulated the number of times specific businesses were recommended to all of the 20 users in the test set and then I plotted this frequency on a bar plot in the following plot.



Through this plot we can see that certain businesses tend to get recommended more than once and may seem to imply either a specific set of similar

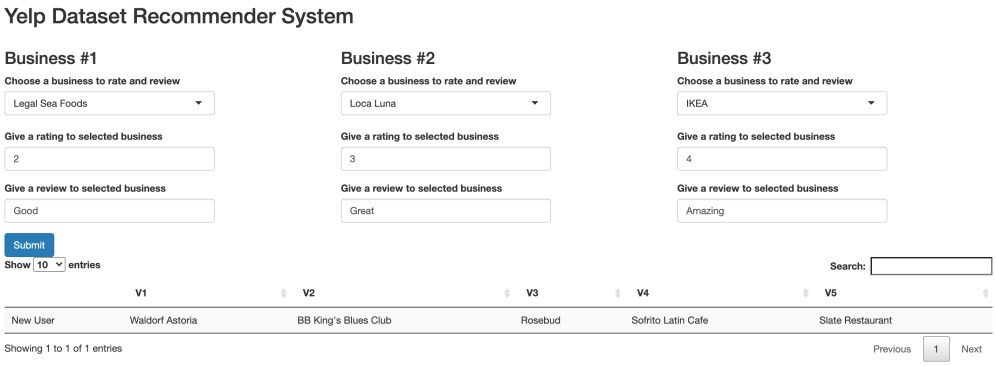
tastes among the users in the test or a very broad appeal among the relatively frequently recommended businesses. Taking this data exploration one step further, I parsed through the categories that Yelp associated with each business that was recommended to the users in the test set. After having once again tabulated the frequency with which certain categories were associated with the recommended business, I also plotted the frequency of these categories on a bar plot in the following plot.



The insight provided by this plot is that almost the entirety of businesses that were recommended were either restaurants or had some restaurant component associated with them. The prevalence of the "restaurants" category

would indicate that because restaurants remain among the most popular/ most reviewed businesses on Yelp that Yelp is a business geared towards sharing customer sentiments on restaurants as it's known for.

Lastly, I created an R Shiny web application to allow people outside of the initial Yelp dataset to give reviews for available businesses. The application can be accessed from the following website link: (https://josephheadley.shinyapps.io/deliverables). The application works by allowing the user to pick 3 businesses and give a numeric rating and a text review for each of the 3 businesses. After submitting said information the application will run an internal R script that then provides the user with 5 business recommendations based on the information that the user provided as input. This web application appears as follows:



Some shortcomings present in my recommendation system was that as a result of the Sentiment Analysis component of the reviews, certain words that were either not in the Bing lexicon or had neither positive nor negative associations attached to them by the Bing lexicon would have no bearings on the result given in our Sentiment Analysis. As such reviews without any charged words would not contribute at all to the final decision of the Recommender System. Luckily the component that really matters, the sentiment_score, is an average of the rating, which the user must provide, and the review_score, which is given

by the Sentiment Analysis of a review. As such there will always be at least one user determined criteria to help judge a user's business preference.

5 Conclusion

In summary, through the combination of Sentiment Analysis and User BAsed Collaborative Filtering, I was able to develop a Recommender System trained on a dataset of Yelp business reviews. From this Recommender System's results on the test dataset, I carried out exploratory data analysis through data visualization of the frequency of specific businesses that were recommended to users in the test set and of the frequency of the categories that these businesses fell under in order to better understand any trends or tendecies that were present in the recommendations given. I, then, took the Recommender System that I built and reframed it as an R Shiny application that I deployed to the free R Shiny Server, which would allow users to make their own recommendations among the available businesses and receive recommendations based on the preferences gleaned the information the user provided. With the high potential for improvement that this Recommender System has some next steps would be to implement a way to evaluate the accuracy of the system in providing a recommendation for businesses that the users in the test set might actually have very positive sentiment towards and most likely choose to frequent or to improve upon the system's capacity to more efficiently handle more users and more business reviews.

6 Citations

- **Dataset**: https://www.kaggle.com/yelp-dataset/yelp-dataset
- **Sentiment Analysis Tutorial**: https://www.kaggle.com/rtatman/tutorial-sentiment-analysis-in-r
- **Recommender System Tutorial**: https://www.data-mania.com/blog/how-to-build-a-recommendation-engine-in-r/
- **R Shiny Tutorial**: https://shiny.rstudio.com/tutorial/