Project Report

Team 10

CS105

12/10/23

Project Description

Our project will utilize text mining to understand what makes a good quality Korean Barbecue restaurant. Using web scraping and tokenization of Yelp reviews from a sample of different Korean Barbecue restaurants in Los Angeles County, we will look for keywords for restaurants both individually and as a collective. Each restaurant will have a sample of their reviews vectorized. Recurring words from other vectorized samples of other restaurants will then be removed, leaving us with our keywords. The vectors for each restaurant will then be compared with the overall rating of the restaurant and the number of keywords taken into account. Words used to describe the atmosphere, price, and quality will make up most of our text processing. These keywords will allow us to distinguish the quality of the different Korean Barbeque restaurants and show what components make up a good or bad Korean barbecue restaurant. We will then try to pinpoint the words used the most in these reviews to describe what makes a place and to see what customers value the most when creating a review. We can see how customers may hold certain components through the text processing of all the reviews. The qualities of a restaurant that the customers believe that restaurants must possess to give a higher review will be uncovered. Our hypothesis concerning this project will involve which of the quality, atmosphere, price, or variety of food being offered contributes the most to a person's review.

Data Collection/Cleaning

Web scraping Yelp.com was our main practice when it came to finding and collecting the necessary data for our project. This data consists of restaurant names and the hundreds of reviews accompanying them. Due to all reviews on Yelp sharing the same HTML/CSS class tag,

"css-1q2nwpv," it became much easier to simply call our function to collect all the reviews and place them in their data frame. More than ten reviews per restaurant needed to be read into our dataset as all Yelp pages only have ten reviews a page, this considered, we had to create a while loop that iterates through the following consecutive pages of the restaurant and removes all information that is not required. All the reviews are placed into an array, which is then turned into a data frame. Then, we use pd.concat to combine the new data frame with the previous data frame consisting of reviews from the previous restaurants. The use of the library BeautifulSoup was essential in our study as this gave us access to the ability to parse the reviews from Yelp and append them to our data frame. These Yelp reviews were grabbed through our requests library which was also imported. Once these data frames were properly cleaned by removing extra entries like Yelp prompts such as "Q:" and "Select your rating," we then output the final data frame into one CSV file with the following simple function.

```
df.to_csv('yelp_reviews_KBBQ.csv', index = False)
```

With our data combined, it could now be preprocessed and read for exploratory data analysis. In the end, we decided to analyze 10 restaurants with 100 reviews each.

	Hanu Korean BBQ - 4.8 Stars	Baekjeong - 4.3 Stars	Hae Jang Chon - 4.1 Stars	Road To Seoul - 3.9 Stars	Moodaepo - 3.3 Stars	Bud Namu Korean BBQ - 3.3 Stars	King Chang LA - 4.8 Stars
0	Such an amazing place, great food, great servi...	This place is delicious. Won was our server t...	My party of 7 and I ended up here after being ...	My favorite place to get food for a lower pric...	Great service... nice atmosphere .. food was ...	Price: 10/10Service: 10/10Staff: Super friendl...	Best Small Instance Soup ever. They offer vari...
1	The staff were nice and kind and they cooked f...	I've been here a number of times over the year...	Experience (5/5 stars):*Food: the meats were g...	I don't leave yelps, not that kinda guy.... Bu...	Combo A Lunch Special for only $25.99! I'm not...	I used to go to this AYCE KBBQ a lot in colleg...	Been here twice and it's been great! Service i...
2	Best Korean bbq in LA request for Michael Jack...	On 11/22/2023 in the evening, I showed up 20 m...	4.5 stars. A solid and tasty restaurant with n...	Prices are good. They seated us pretty quickly...	Moodaepo has one of the cheapest AYCE KBBQ opt...	I recently had the pleasure of dining at Bud N...	This place has great service and plenty of opt...
3	Hanu was absolutely delicious!!! From someone ...	Food: Ordered the Large Beef Combo that includ...	We're a party of 7, put our name down at 7:15 ...	Been coming here since I was a kid. There are ...	Got there at 11:30m. Business advertising that...	Hole in the wall, decent place with decent mea...	Our server Wayne, was really kind and fun. The...
4	What a great find in K-town! Came here for the...	6:30pm on a Wednesday is probably the charm be...	Still my favorite Korean BBQ in LA and happy t...	This spot is poppin' on a weekend night! I cam...	Great KBBQ place. Perfect for large parties. I...	This restaurant located in the heart of K-town...	Saturday 5pmParking/Wait: They have a valet-on...
...	...	...	...	...	...	...	...

After converting the data frame to a readable CSV, we must conduct preprocessing with tokenization before conducting exploratory data analysis. This means we have to turn all relative terms into comparable tokens by editing the reviews. To meet this criteria we must remove all punctuation, capitalization, and stop words from the reviews.

Exploratory Data Analysis

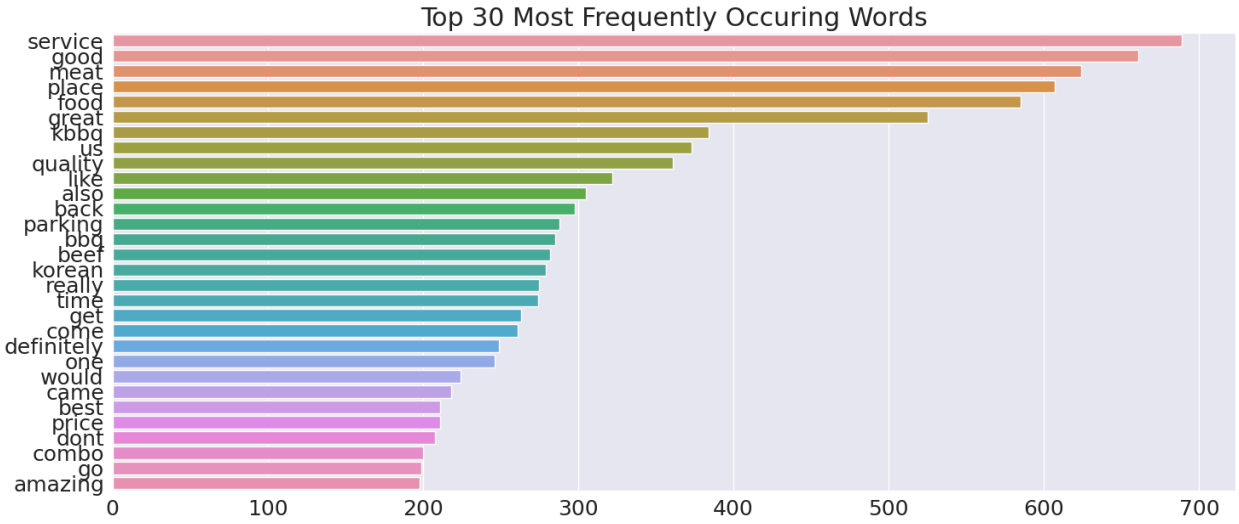
Frequency Analysis:

By analyzing the most frequently used we wanted to see the trends, patterns, and relationships between a restaurant rating and specific words. For our dataset, the words we wanted to look at included words relating to different qualities of a Korean barbeque restaurant. To do this, we first created a corpus with the different reviews and words.

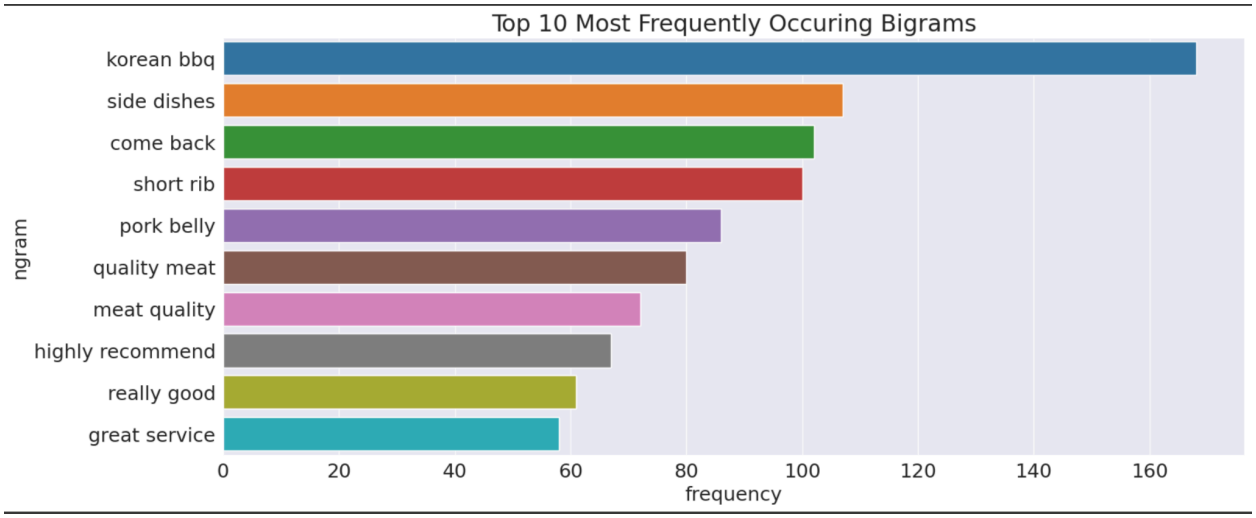
```
[ ] def corpus(text):
        text_list = text.split()
        return text_list
```

```
dataframes = [hanu_kbbq, prime_k, king_chang_LA, pzkbbq, woo_hyang_woo,
              j_bbq, castle_bbq, moodaepo, bud_namu_kbbq, moon_bbq]

merged_df = pd.concat(dataframes, ignore_index=True)

merged_df['Review_lists'] = merged_df['Review'].apply(corpus)

# Display the merged DataFrame
print(merged_df['Review_lists'].head(n=1000))
```

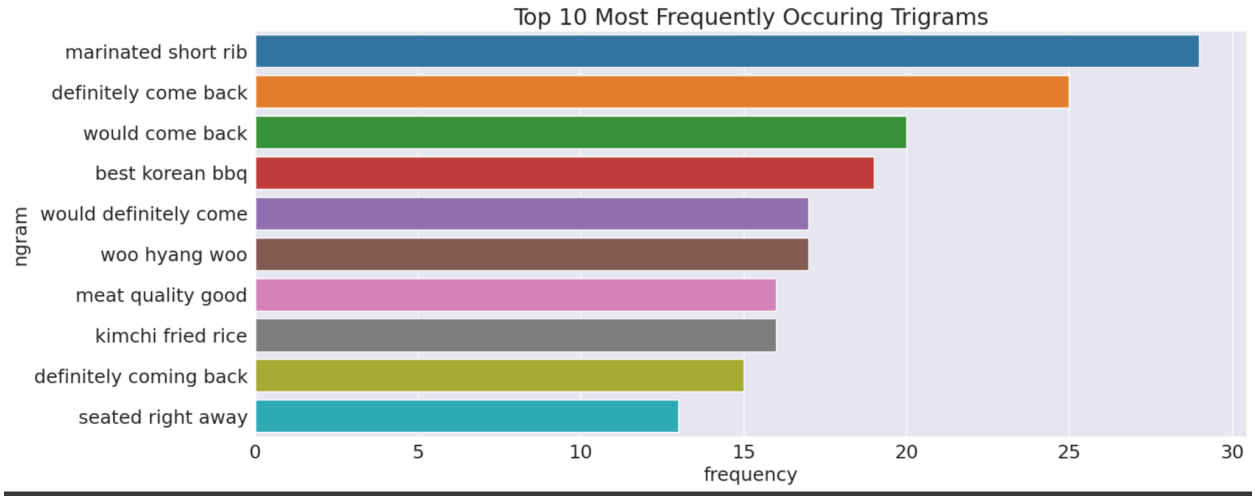
The code above shows us creating a corpus by splitting the reviews and merging all the reviews for each restaurant into one merged data frame to work with. After this, we moved on to finding the frequencies for each term and creating data visualizations to analyze.



The purpose of collecting the term frequency is to better understand the terms used by customers to describe the quality of a Korean barbeque restaurant. From this visualization of term frequency, we were able to see that some of the most commonly used terms used to describe a restaurant's rating. A lot of these terms are also related to food places in general, however, we see from the histograms a lot of the terms that are unique to Korean barbeque restaurants relate to customer service, the variety of meats they have, the quality of their food, and location.



To go deeper into our term frequency analysis, we created n-grams and looked into the most frequent bigrams and trigrams. By analyzing we can get a better contextual understanding of common phrases used by reviewers. In the bigram above, we see that the most common phrases relate to the variety and quality of meats. With terms such as “short rib”, “pork belly”, and “meat quality” something that separates these Korean barbeque restaurants from one another are the different meats along with customer service and the other factors to have a successful restaurant.



To go even further we did trigrams, and the results from the most frequent phrases were all related to the variety of food items and customer service. From these visualizations of term frequency, we were able to determine terms that uniquely relate to the Yelp ratings of Korean barbeque restaurants. Our group determined that along with customer service, the variety of a restaurant menu means a lot to their Yelp ratings and reviews.

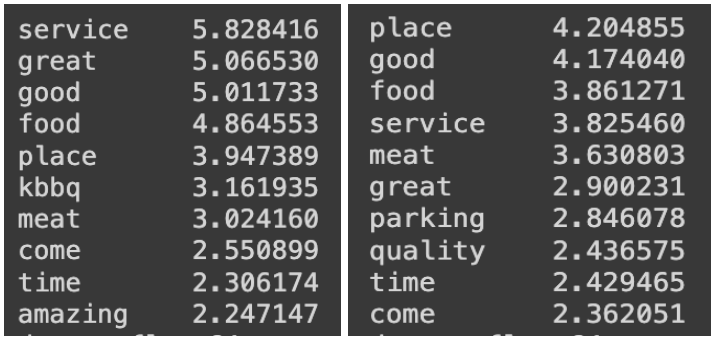
TF-IDF

Due to our use of term frequency, we decided that TF-IDF was the optimal way to retrieve extra information regarding the tokens in our dataset. With this considered, we conducted extra cleaning through the removal of NaN entries and computed the TF-IDF matrix for each restaurant. Through our computations, we were able to determine the importance of each word relative to their frequency. We noted that despite the reviews for the restaurants being drastically different from the lower-rated to the higher-rated reviews, people who dine at K-BBQ restaurants typically look for the same qualities regardless of location. These qualities are food quality, food variety, and quality of service. Surprisingly, not much data appeared regarding the atmosphere, cleanliness, or hours/availability.

Higher rated:

food	5.723386	meat	5.654619
great	5.393786	quality	5.542337
meat	4.867576	place	5.492609
service	4.561904	great	4.904881
beef	4.494867	service	4.853308
good	4.134559	good	4.413865
place	4.089783	food	3.970838
combo	4.037313	korean	3.309478
kbbq	3.847200	amazing	3.125325
server	3.115545	definitely	3.118008

Lower Rated:



service	5.828416	place	4.204855
great	5.066530	good	4.174040
good	5.011733	food	3.861271
food	4.864553	service	3.825460
place	3.947389	meat	3.630803
kbbq	3.161935	great	2.900231
meat	3.024160	parking	2.846078
come	2.550899	quality	2.436575
time	2.306174	time	2.429465
amazing	2.247147	come	2.362051

(Despite the differences in rating, these four places shared similar scores regarding TF-IDF.

Sentiment Analysis

A sentiment score closer to 1 represents a more positive sentiment, 0 is a more neutral statement, and -1 is a generally negative sentiment. In our case, all restaurants had a score above 0.5, showing a relatively positive sentiment from the reviewers. Sentiment scores from the higher-rated group were higher than those of the lower-rated group, (which was expected,) with the highest and lowest from the higher-rated group being 0.7684 and 0.911 respectively, while the lowest and highest in the lower-rated group were 0.5709 and 0.8133. This shows that restaurants that have more of the "important words" as listed above are also able to achieve a higher sentiment score, or essentially a more positive reaction from reviewers.

In sentiment analysis, we used the WordNet database, which is based on the Natural Language Toolkit, and the pre-trained sentiment analysis model Sentiment Intensity Analyzer, similarly based on the NLTK. These two resources combined with some simple text cleaning (removing contractions and unimportant words) allowed the simplification of each review's meaning, giving us the ability to give them a sentiment score that accurately reflected their original intentions.

```
lemmatizer = WordNetLemmatizer()
```

```
sia = SentimentIntensityAnalyzer()
```

Conclusion

Through our different frequency methods and data analysis, we determined that food variety and quality were the main focus of reviewers when it came to determining a good Korean Barbecue restaurant. Despite reviews and businesses of varying star levels, this desire for variety and quality remained the same as for all restaurants, these qualities were the most discussed and

sought out. Besides the food, service was also highly discussed as this token possessed the highest frequency in our first EDA frequency method. Despite their perceived importance, the atmosphere and cleanliness of the restaurants did not appear in the top 10 or top 30 for terms in any of our methods.

Contributions

Eric Gordeyev: Led data collection and cleaning, and aided in both EDA and TF-IDF development and analysis.

Paul Nguyen: Aided in data collection and cleaning, developed slides/presentation and project report, and assisted in EDA and TF-IDF development.

Nate Natividad: Aided in data collection and cleaning, developed slides/presentation and project report, and assisted in EDA and TF-IDF development.

Jenny Zhang: Led TF-IDF and sentiment analysis, and contributed analysis in the EDA portion and project report.

Eric Yang: Led EDA portion developing and visualizing term-frequency and n-grams, also contributed towards TF-IDF analysis.