



KBBQ Text Mining

Eric Yang, Jenny Zhang, Eric Gordeyev, Nate
Natividad, Paul Nguyen

Project Description

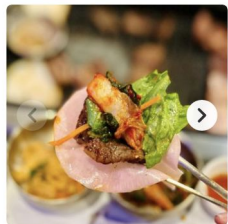
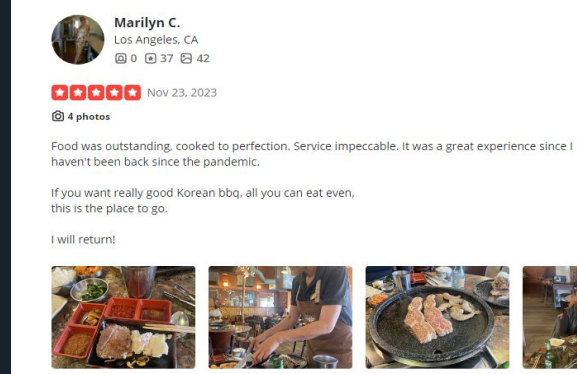


- Web scrape, read, and classify reviews for Korean-BBQ restaurants of all ratings in the LA area.
- Understand what contributes the most to making a quality K-BBQ restaurant (Variety, Service, Food Quality, Atmosphere, etc.)
- Explore the frequency and relationship between relevant terms.



Data Collection

- Web scraped 10 restaurants
- 100 reviews each
- Each review is categorized by business name with accompanying star rating
- Analyzing restaurants of varying star levels from some of the best to the worst in the area.



1. Hanu Korean BBQ

★★★★★ 4.8 (1.3k reviews)

Korean Barbeque Wilshire Center

Open until 11:00 PM

Free parking • Free WiFi

"Thank God, I've found Hanu **KBBQ**! With so many new kbbq joints popping up in and round LA pretty much..." [more](#)

Data Cleaning

- Once the 10 restaurants were loaded in it was necessary to concat all of these into one single dataframe to transform into a csv.
- We then had to clean every instance of pre-generated Yelp prompts that were present.
 - The two being “Q: “ and “Select your rating”
- Once conducted we can convert all of this into a csv file with one line of code.

```
- df.to_csv('yelp_reviews_KBBQ.csv', index =
```

	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
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...
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...
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...



Data Cleaning

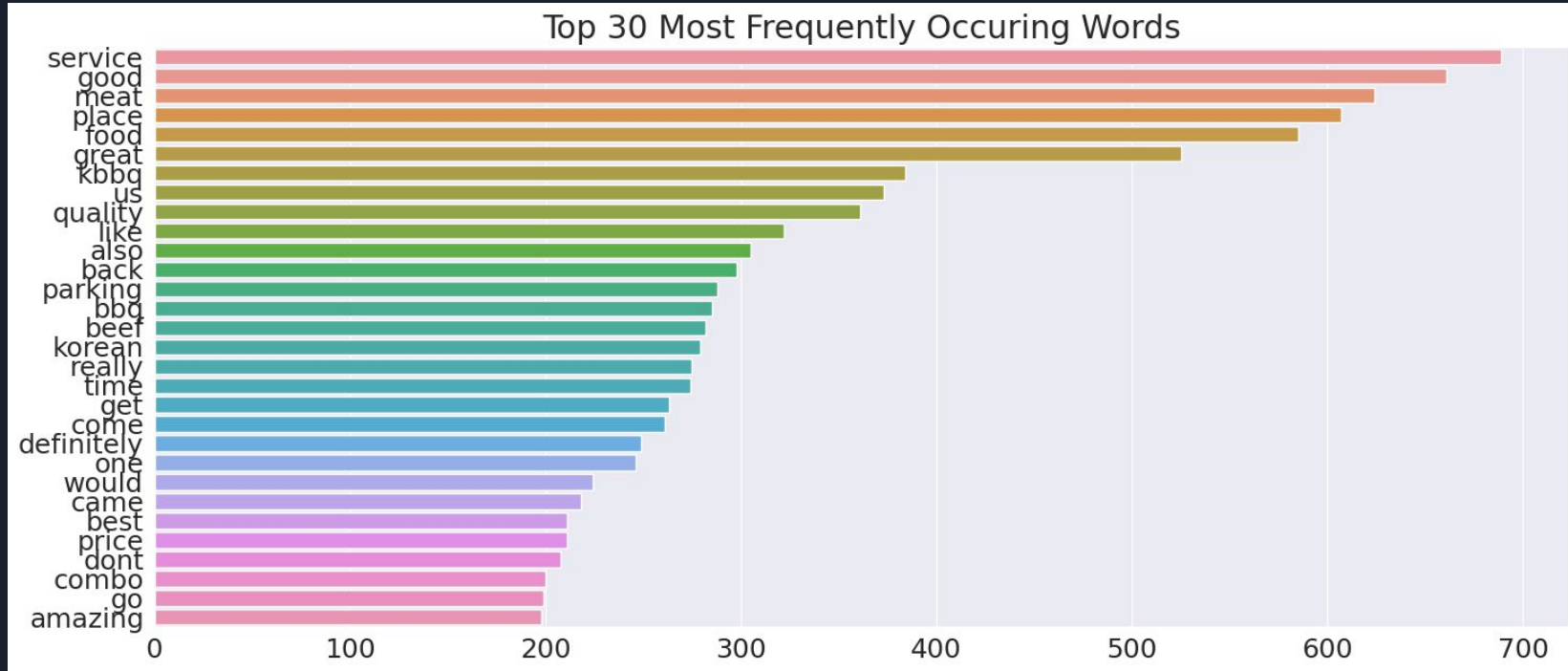
- After conversion it was necessary to preprocess all the reviews with tokenization.
 - Punctuation removal
 - Upper Case to Lower Case
 - Stop Words removal
- All terms now become comparable tokens ready for analysis



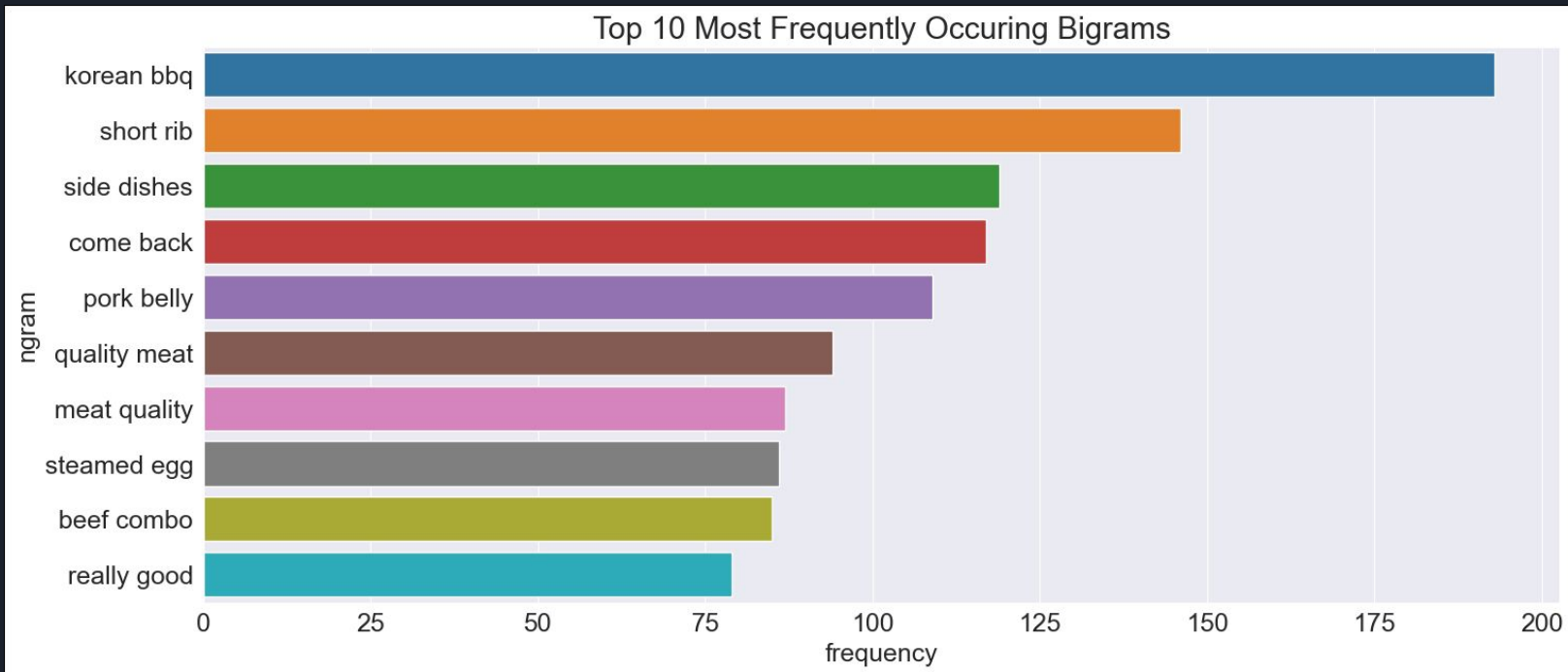
Exploratory Data Analysis (EDA)

- Term Frequency
 - We examine the most frequently occurring words (tokens) within the dataset
- Frequently occurring N-grams
 - An n-gram is sequence of n words in a text, we want to examine the most frequently occurring
 - Bi-gram : two words in a sequence (ex: 'very good')
 - Tri-gram: three words in a sequence (ex: 'good service today')

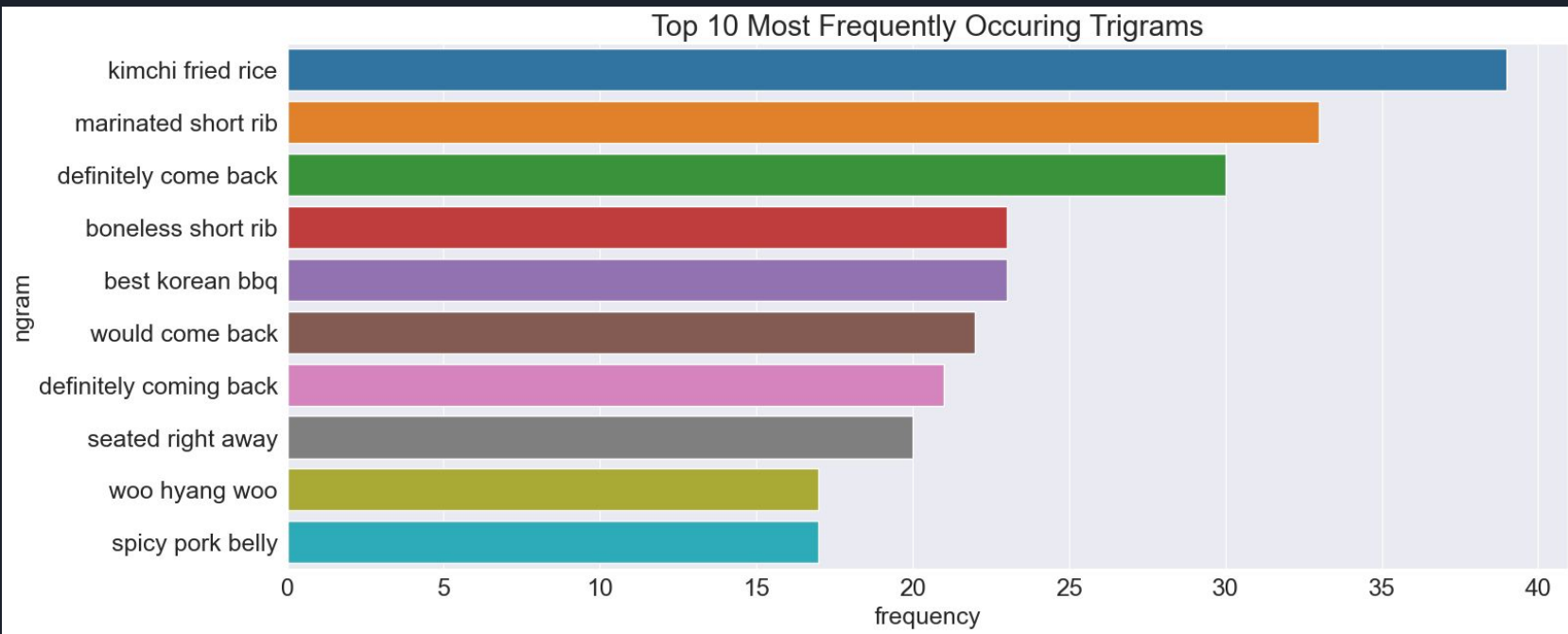
Term Frequency




Frequently occurring bi-grams



Frequently occurring tri-grams





TF-IDF (term frequency-inverse document frequency)

- Measure used to assess the importance of a word relative to the document as a whole and the corpus.
- Importance is determined by frequency within reviews but can be offset by word frequency within the whole corpus.
- Will be conducted for individual restaurants to determine which words possess the most importance in regards to both the review and the restaurant as a whole.



TF-IDF

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

- Important words showing up in both higher and lower rated restaurants
 - Both reflect reviewers' focus on service and overall quality of food served



TF-IDF - Sentiment Analysis

```
# contractions
def expand_contractions(s):
    s = re.sub(r"won't", "will not", s)
    s = re.sub(r"wouldn't", "would not", s)
    s = re.sub(r"couldn't", "could not", s)
    s = re.sub(r"can't", "can not", s)
    s = re.sub(r"n't", " not", s)
    return s
```

```
# group similar words
lemmatizer = WordNetLemmatizer()
```

```
sia = SentimentIntensityAnalyzer()
```

- Clean text data
 - Expand contractions
 - Change to all lowercase
 - Remove any reviews that may not be in English
- WordNet database
 - English language
 - Based on Natural Language Toolkit
- Use of pre-trained model
 - For sentiment analysis
 - Assigns sentiment score to text



TF-IDF - Sentiment Analysis

```
print (data['Sentiment'].mean())
```

- Get mean Sentiment score for each restaurant
 - Closer to 1 = positive sentiment
 - Closer to -1 = negative sentiment
- All restaurants had Sentiment score above 0.5
 - Mostly positive sentiment from all reviewers
- Higher rated restaurants had higher Sentiment score
 - Restaurants that included important points valued by reviewers



Conclusions

- The most important terms from our analysis involved food quality and variety.
- Term Frequency provided more context for the words that users are using to describe the restaurants.
- TF-IDF ran on the individual restaurants shows the words most unique to that restaurant and the most prominent TF-IDF scores were words focused on service and quality of food.
- Our project could be improved by having a larger dataset including more restaurants and in different areas for more variance.



Questions

1. What prompts had to be cleaned after our data was concat?
2. What is tokenization and why is it necessary?
3. What can you infer about the differences in use cases between bi-grams and tri-grams?