Project Name: Taste of Korea in LA: Analyzing LA's Korean Restaurants through Yelp
Name: Xinyao Fu

Introduction

Have you ever wondered what traits make a restaurant with a good reputation stand out? Since arriving in Los Angeles, I've dined at many Korean restaurants. There are numerous Korean restaurants in Los Angeles, and I often hear about new ones that my friends recommend. However, I sometimes hear negative reviews about some restaurants too. This made me ponder what customers value more in these Korean restaurants. What factors make them think a restaurant is worth trying? What factors lead to negative impressions? This research aims to answer these questions by analyzing restaurant data and review data of Korean restaurants in Los Angeles on Yelp. This research may also offer insights to new Korean restaurants looking to enter the Los Angeles market.

Data Collection

Yelp provides a public, free API that allows access to a wealth of data, including data on restaurants and reviews. Therefore, I decided to use this API to begin gathering my raw data. I registered for an API key and followed Yelp's instructions to collect my data, which includes each restaurant's ID, name, location, rating, review count, price, readable ID, latitude, and longitude. It also includes each review's ID, corresponding restaurant ID, review text, rating, and time created. I used a SQLite database to store this data, with restaurant data in one table and review data in another. These two tables are linked through the restaurant's ID.

However, I encountered several problems while using the API to collect data. A major issue I discovered was that Yelp's API only supports collecting the latest three reviews per restaurant, and these reviews display only the first 160 characters, meaning many are incomplete. I explored the possibility of solving this through web scraping or third-party platforms. However, this is difficult due to Yelp's anti-scraping technology, and their official website states that scraping their site or using any third-party software violates their Terms of Service. Therefore, I decided to analyze the existing data to see if I could get any useful insights. Additionally, the Yelp API has a daily limit on API calls, so I couldn't collect all the reviews at once. To address this, I created a new Python file to collect the reviews that were missed due to these limitations. In total, I gathered data on 1000 restaurants and 2978 reviews.

Data Cleaning

During the data cleaning step, I first ensured that all columns in the SQLite database that should be numerical were integers or decimals. Then, I checked for empty rows in both tables and found that the price and location columns in the restaurant table had some missing data. I then marked them as 'unknown'. I also noticed that there were 9 non-English texts in the review data, so I used the deep_translator tool to translate these texts into English.

I further attempted to correct misspelled words in all texts using SpellChecker, but the results were not entirely accurate. SpellChecker mostly corrected interjections, like changing 'yooo' to 'yoon' or 'hmmm' to 'ummm', but it also made some errors that altered the meaning

and could impact subsequent sentiment analysis, like changing ‘yooooo!’ to ‘nooooo!’ and ‘mmm’ to ‘mom’. It also incorrectly altered restaurant and dish names, like changing ‘Jumak’ to ‘juma’ and ‘Magal Korean BBQ’ to ‘maga Korean be’. Since the actual misspellings corrected by SpellChecker were few and most changes were these types of words, I decided to use the data without SpellChecker processing for greater accuracy.

Sentiment Analysis and Outliers Analysis

I used the SentimentIntensityAnalyzer from NLTK for sentiment analysis and stored the results along with other data in a new database file for analysis. While reviewing the data, I noticed some reviews with low sentiment scores have high ratings of 4 or above. I conducted an outlier analysis on these data points. I converted the sentiment scores into a 1-5 scale and identified the number of reviews where the sentiment score differed from the rating by more than 3 points, finding 99 reviews identified as outliers. Upon examining some of these outliers, I found that their occurrence might be due to the analyzer could produce some inaccuracy and some of the review contents are incomplete.

Text Analysis

Considering the potential inaccuracies in sentiment analysis results, I decided to focus the text analysis on ratings. I first categorized reviews into four groups: those with a sentiment score greater than or equal to 0, those with a sentiment score less than 0, those with a rating greater than or equal to 3, and those with a rating less than 3. I then created word clouds for each of these groups. These Figures can be found through this link: https://github.com/xinyaofu/Yelp_LA_Res/tree/cdbce3d7c77aaf710198ea77ad5d3c28d71663ab/results/Visualizations/Text%20Analysis%20Word%20Clouds. The word clouds generated from reviews with high sentiment scores and high ratings were very similar, frequently featuring

Figure 1: Word Cloud - Low Rating Reviews

Figure 2: Word Cloud - Low Sentiment Score Reviews

words like ‘food’, ‘service’, ‘good’, ‘amazing’, and ‘delicious’. However, there were some differences in the word clouds generated by the low sentiment score group and the low rating group. The following two images show the word clouds produced by reviews with low sentiment scores and low ratings.

However, it's still noticeable that in both groups, words like 'good' or 'great' are significantly less prominent compared to the other two groups, and despite slight differences in specific frequencies, the most frequently occurring words in both figures include 'food', 'time', 'service', 'restaurant', and 'chicken'. To highlight the specific frequencies of each word, I used horizontal bar charts in Figures 3 and 4 to show common word usage in high and low rating reviews. This analysis clearly showed that 'food' is the most frequently occurring word in both groups, significantly more than any other word, and 'service' also appears with high frequency in both charts. However, high rating reviews feature positive words like 'good' and 'delicious' more frequently. Additionally, we can infer from both charts that the number of high rating reviews is much greater than that of low rating reviews.

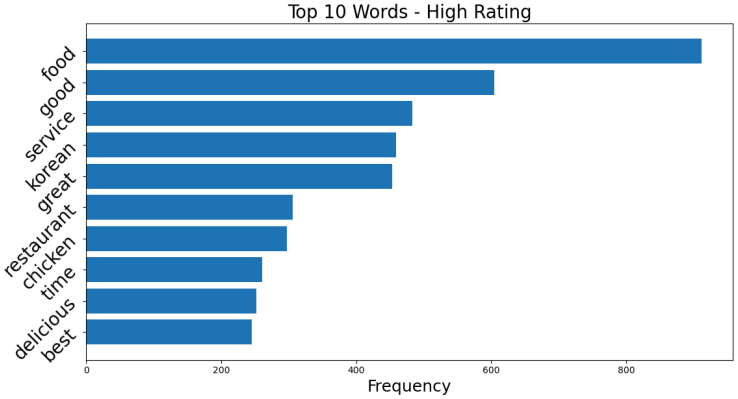


Figure 3: Top 10 Words - High Rating

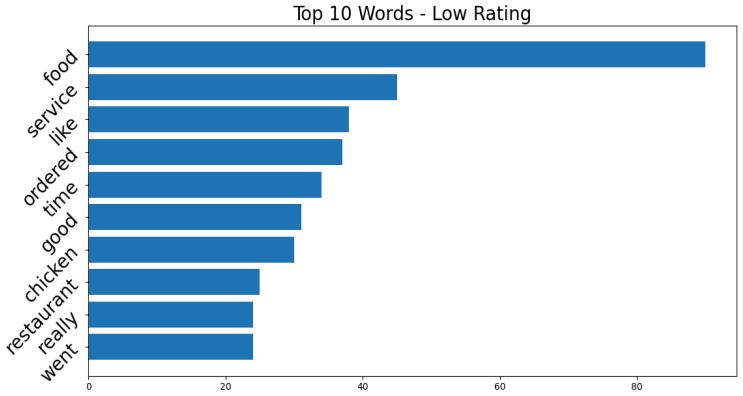
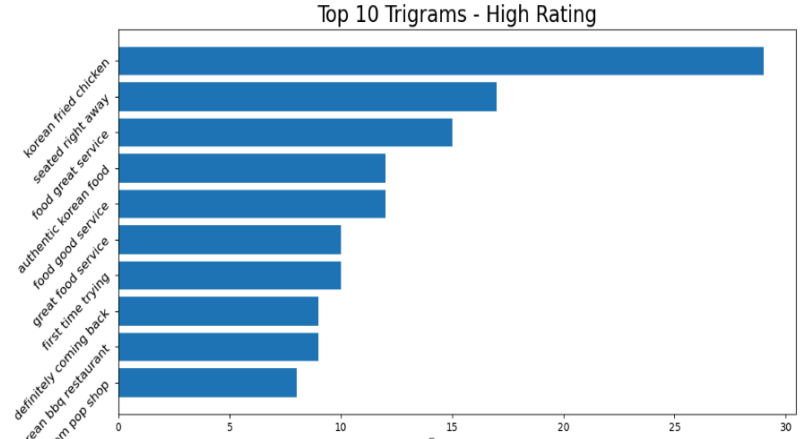
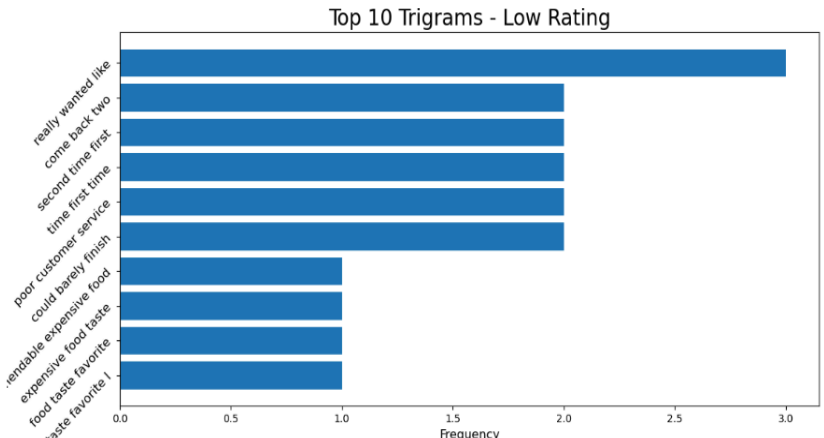


Figure 4: Top 10 Words - Low Rating

From the above analysis, we can deduce that customers place significant importance on a restaurant's food and service. However, the information that can be gleaned from individual words is somewhat limited; for instance, I don't know the nature of people's comments about the restaurant's food and service. To gain deeper insights, I analyzed the top 15 most frequent trigrams - sequences of three consecutive words - in both high and low rating reviews. Figures 5 and 6 display the results I obtained. From this analysis, I learned that in terms of food, Korean fried chicken is particularly popular among customers, and they care about the authenticity of the Korean food. From the low rating reviews, it's evident that expensive food can lead to customer dissatisfaction. Regarding service, customers prefer to be seated right away and place a high value on service quality; in fact, three of the top ten trigrams in high rating reviews are related to service. However, one limitation of the trigram analysis is the small volume of low rating reviews, which could introduce some bias into the results.





Additionally, using the Word2vec algorithm, I predicted the words commonly associated with 'service' in high and low rating reviews. The bar chart results are available here: https://github.com/xinyaofu/Yelp_LA_Res/tree/cdbce3d7c77aaf710198ea77ad5d3c28d71663ab/results/Visualizations/Text%20Analysis%20Bar%20Charts. However, the results did not provide many useful insights. For high rating reviews, the most probable words included ‘food’, ‘great’, ‘super’, ‘staff’, and ‘amazing’, while for low rating reviews, the top words were ‘packaging’, ‘overnight’, ‘dry’, ‘two’ and ‘variety’. This might suggest that good service can significantly improve a customer’s overall impression of a restaurant, but beyond that, there wasn't much valuable information gleaned.

Price Analysis

In addition to text analysis of reviews, I was also curious if a restaurant's price level influences its overall customer ratings and sentiment scores. Therefore, I conducted a price analysis on these 1000 restaurants. I categorized the restaurants into four price levels and calculated the mean and standard deviation of ratings and sentiment scores for each price level. Figure 7 shows the results I obtained. Surprisingly, I found that except for price levels 2 and 3, the means generally increased with the price level, while the standard deviations decreased. Although this relationship is not particularly strong, this reflected that lower-priced restaurants are more likely to receive lower ratings, and the ratings they receive tend to be more variable.

	sentiment_score		review_rating	
	mean	std	mean	std
price_level				
1.0	0.421	0.469	4.161	1.183
2.0	0.521	0.420	4.328	1.082
3.0	0.559	0.407	4.290	1.151
4.0	0.570	0.391	4.533	0.875

Figure 7: Price Analysis - Mean and Standard Deviation

Then, I calculated the correlation and corresponding P-values between price level, sentiment score, review rating, and restaurant rating. Figure 8 displays the results I obtained. These results indicate a very weak positive correlation between price level and sentiment score, price level and review rating, price level and restaurant rating, and sentiment score and restaurant rating. There is a moderate positive correlation between sentiment score and review rating, and a weak positive correlation between review rating and restaurant rating. The relationships between price level and sentiment score, sentiment score and review rating, sentiment score and restaurant rating, and review rating and restaurant rating are statistically significant. However, with P-values greater than 0.05, the relationships between price level and review rating, and price level and restaurant rating, may not be statistically significant. I believe some of the inaccuracies in these findings stem from the fact that each restaurant only had three review data points. For instance, there should be a significant positive correlation between review ratings and restaurant ratings, yet the correlation analysis shows only a weak association.

```
Correlation between price_level and sentiment_score: Pearson corr = 0.071, P-value = 0.000
Correlation between price_level and review_rating: Pearson corr = 0.039, P-value = 0.055
Correlation between price_level and restaurant_rating: Pearson corr = 0.025, P-value = 0.213
Correlation between sentiment_score and review_rating: Pearson corr = 0.411, P-value = 0.000
Correlation between sentiment_score and restaurant_rating: Pearson corr = 0.140, P-value = 0.000
Correlation between review_rating and restaurant_rating: Pearson corr = 0.249, P-value = 0.000
```

Figure 8: Price Analysis - Correlation and P-value

Location Analysis

In addition, I was curious about the importance of restaurant locations and the distribution of Korean restaurants in LA, so I used Folium to visualize their positions based on latitude and longitude. The visualization result can be found through this link: https://github.com/xinyaofu/Yelp_LA_Res/tree/cdbce3d7c77aaf710198ea77ad5d3c28d71663ab/results/Visualizations/Text%20Analysis%20Word%20Clouds. I found that most Korean restaurants are concentrated on certain streets. The area with the highest concentration is Koreatown near downtown LA, with over 300 Korean restaurants. However, I discovered that a restaurant's location does not seem to have a significant correlation with its ratings and the number of reviews, because the ratings and review counts of restaurants in Koreatown vary greatly, and similarly, restaurants in the LA area and surrounding regions also range from having over thousands of reviews to just one, with overall ratings varying widely.

Conclusion

My analysis indicates that the success of a restaurant largely depends on the quality of its food and service, with Korean fried chicken emerging as a particularly popular dish that new Korean restaurants should consider featuring. Additionally, customers highly value quick seating and minimal waiting times, and while excessive pricing can lead to dissatisfaction, according to the price analysis, setting prices too low might show risk of greater variation in reviews received, which might be due to food quality not guaranteed. Although this conclusion might be biased due to the limited number of review data, we can still glean that a balanced approach to pricing and a guarantee in quality is essential for attracting and retaining customers.

Location analysis has shown that most Korean restaurants are concentrated in certain streets or plazas, even if some are just small clusters of three or four restaurants. Notably, Koreatown in Los Angeles has the highest proportion of Korean restaurants. From the visualization data, I observed that there is no significant correlation between the specific location of a restaurant and its ratings and number of reviews. Therefore, I believe that for new Korean restaurants entering the market, the advantage to opening in areas like Koreatown is that this place has a large base of Korean cuisine enthusiasts, but it also has intense competition. Conversely, opening in areas with fewer Korean restaurants presents less competition. A viable strategy might be to open in smaller Korean restaurant clusters outside of Koreatown, benefiting from the existing clientele in these areas while facing reduced competition.

Future Work

My analysis has some inaccuracies and biases due to the limited data available, specifically, only the latest three reviews per restaurant has been collected and there are many incomplete reviews. If I had more time, I would seek solutions to this problem, such as finding ways to access more review data. With a larger dataset for the aforementioned text and price analyses, my results and conclusions would likely be more accurate. Additionally, I hope to conduct a more precise, statistics-based location analysis to determine if there is any correlation between a restaurant's location and its number of reviews and ratings.