Data Design &

Representation

Final Project Group Report

Yifan(Crystal)Cai

Siwei(David)Ran

Sanjana Kallol

Table of Contents

Executive Summary

The popularity of Korean cuisine has increased significantly in recent years, making it a popular food option in many parts of the world, including the United States. In this project, we aimed to create a dataset of the top Korean restaurants in ten major cities in the United States. We gathered information from Yelp, including the name, rating, price range, and number of reviews of each restaurant, to create a comprehensive dataset that provides valuable insights into the Korean restaurant industry in the United States.

We used Python to scrape data from Yelp for each of the selected cities, which include New York, Los Angeles, Chicago, Oakland, Houston, Philadelphia, Seattle, Atlanta, Dallas, and San Jose. Our analysis of the data allowed us to identify the most popular Korean restaurants in each city and provide a list of recommendations for anyone interested in trying Korean cuisine.

Our findings suggest that Korean cuisine is highly popular in some of the selected cities, such as Los Angeles and New York, while still growing in others, such as Atlanta and Oakland. We also found that Korean BBQ is a highly popular dish in most of the selected cities.

Our dataset and analysis can be used by Korean restaurant owners and consumers alike to better understand the Korean restaurant industry in the United States. By providing information on the most popular Korean restaurants, as well as insights into consumer preferences, this project can help restaurant owners identify areas for growth and improvement, and consumers make informed decisions about where to dine. Overall, this project serves as a valuable resource for anyone interested in Korean cuisine or the Korean restaurant industry in the United States.

Introduction

In recent decades, Korean cuisine has gained significant popularity around the world, including in the United States. Korean restaurants in the United States offer a wide range of food experiences, from traditional Korean dishes to creative fusion concepts that combine Korean flavors with elements from other cuisines. As Korean cuisine continues to grow in popularity across the United States, our team aims to create a dataset by web-scraping basic information on Korean food from Yelp. Our team is passionate about Korean food, including dishes such as bibimbap, bulgogi, and Korean BBQ. To gather information, we researched a complete list of cities in the United States that have a designated Koreatown or a strong Korean food scene. Based on this research, we have selected the following cities to investigate on Yelp: New York, Los Angeles, Chicago, Oakland, Houston, Philadelphia, Seattle, Atlanta, Dallas, and San Jose. Our goal is to compile comprehensive information on Korean restaurants in these cities.

Our project aims to create a comprehensive dataset of the top Korean restaurants in each of the ten cities we have selected. By analyzing the information gathered from Yelp, we will be able to determine the most popular Korean restaurants in each city, and provide a list of recommendations for people who are interested in trying Korean cuisine.

In addition to providing a list of recommended Korean restaurants, our project will also gather and analyze data on the average rating, price range, and number of reviews of each restaurant. This information will help us to determine the overall popularity of Korean food in each city and provide insights into the consumer preferences in different regions.

Our analysis of the data gathered will provide valuable insights for both Korean restaurant owners and consumers. Restaurant owners will be able to identify areas for improvement and growth, and tailor their menus and services to better suit the preferences of their local customer base. Consumers, on the other hand, will have access to reliable information on the best Korean restaurants in each city, making it easier for them to make informed decisions and discover new and exciting dining experiences.

Overall, our project aims to provide a comprehensive and valuable resource for anyone interested in Korean cuisine or the Korean restaurant industry in the United States.

Data Source

Source:

Our data source for this project is Yelp, a popular online directory for exploring local businesses such as restaurants, bars, and other establishments. Yelp publishes crowd-sourced reviews and detailed information about each business, including its rank, name, location, rating, and more. Customers can share their reviews and feedback through this community, making Yelp a valuable resource for gathering information about businesses.

To begin our data collection process, we entered specific keywords such as "Korean" and the name of the city we were researching into the Yelp search bar. This allowed us to access the search results page for Korean restaurants in each city, which contained detailed information about each establishment. We did not use any filters on the left-hand side of the page, as our goal was to gather broad results for Korean restaurants in each city.

By gathering and analyzing data from Yelp, we aim to provide a comprehensive and accurate dataset of the top Korean restaurants in each of the ten cities we have selected, along with valuable insights into consumer preferences and trends in the Korean restaurant industry.

Web-Scraping Routine:

1st Step: Verification of Yelp web scraping

Before beginning web scraping on Yelp, we needed to ensure that Beautiful Soup was capable of scraping Yelp. To do this, we attempted to retrieve information from Yelp by downloading pages from the website. After verifying that this was possible, we proceeded to the next step.

2nd Step: Accessing and downloading Yelp pages

We created a specific URL for each Yelp page by looping through a list of cities, and then used a get request command to download three pages for each city. These pages contained information on the top 30 Korean restaurants. We also wrote code to save each page with a specific name for future reference (Appendix 3), and confirmed that the contents were available in the HTM file by opening several of the downloaded pages.

3rd Step: Extracting information from downloaded files

We looped through each downloaded file and identified the selector for the detailed information of each restaurant, which was found within the "#main-contents" section of the page (Appendix 4). From this section, we were able to extract information such as the restaurant's rank, name, cost, and rating.

4th Step: Extract information from Yelp Restaurant webpage

We also downloaded the restaurant yelp webpage files by loop through the list of link of each restaurant. So we could scrape more detailed information such as address, phone number and popular dishes. Additionally, we uses the positionstack API calls to get the geolocation of each restaurant by its address(Appendix 5).

5th Step: Transforming information

To facilitate future use, we decided to transform some variables from strings to floats or integers. For example, we converted the dollar sign range from "$$$$" to "$" into an integer range from 4 to 1 (Appendix 5). We also removed the string "rating" from each rating and converted the remaining portion into a float (Appendix 6). These transformations allowed us to perform calculations on the average rating and cost of the top 30 Korean restaurants in each city.

6th Step: Storing data in MongoDB

The final step was to store the data we extracted from the HTM files in MongoDB (Appendix 7). We first connected to MongoDB, then created a database called "ucdavis" and a collection called "restaurant". Finally, we inserted the extracted data into this collection for future use.

Data Design Choice:

Variable	Type	Description
City	String	The city name of the restaurant

Name	String	The name of the restaurant
Rank	String	The rank of the restaurant in each city
Link	String	The Yelp page link of the restaurant
Rating	Float	The rating of the restaurant on Yelp
Image	String	The photo of the restaurant on Yelp
Price Indicator	String	The average price of the restaurant $: Inexpensive (typically less than $10 per entree) $$: Moderately expensive (typically $10-$20 per entree) $$$: Expensive (typically $20-$30 per entree) $$$$: Very expensive (typically $30 or more per entree)

Price Indicator Number	Int	The average price of the restaurant
Location	String	The area of the restaurant in its city
Review Count	String	The number of reviews of restaurant on Yelp
Address	String	The exact address of the restaurant
Golocation	String	Latitudes and longitudes of the restaurant location
Phone Number	String	10 digit number
Popular Dishes	String	List of popular dishes

Discussion on Business Impact

Our web scraping project has successfully gathered data on Korean restaurants in ten major cities across the United States from Yelp. By analyzing the data we have collected, we can provide insights into the popularity and preferences of Korean cuisine in different regions.

One interesting finding from our analysis is that Korean cuisine is most reviewed in San Jose, followed by Los Angeles and Chicago. The average price range for Korean restaurants is around $11-$30 per person, with some restaurants offering more expensive options. The most popular dishes in Korean restaurants include Korean BBQ, bibimbap, and bulgogi.

Our data can also be used to identify the most popular Korean restaurants in each city, based on the number of reviews and average rating. For example, the most popular Korean restaurant in Los Angeles is BROKEN MOUTH | Lee's Homestyle, with over 1400 reviews and an average rating of 5 stars. This information can be useful for both restaurant owners and consumers, as it allows them to identify popular and well-reviewed restaurants.

Summary and Conclusion

In conclusion, our web scraping project has provided valuable insights into the Korean restaurant industry in the United States. By analyzing data from Yelp, we have been able to identify popular Korean restaurants in ten major cities, as well as determine the average price range and popular dishes. Our project has created a comprehensive dataset that can be used by anyone interested in Korean cuisine or the Korean restaurant industry.

The data we have collected can also be used for further analysis, such as identifying the most popular Korean dishes in each city or comparing the popularity of Korean cuisine to other types of cuisine. The metrics we have gathered, such as the number of reviews and average rating, can also be used to analyze consumer preferences and identify areas for improvement and growth for restaurant owners.

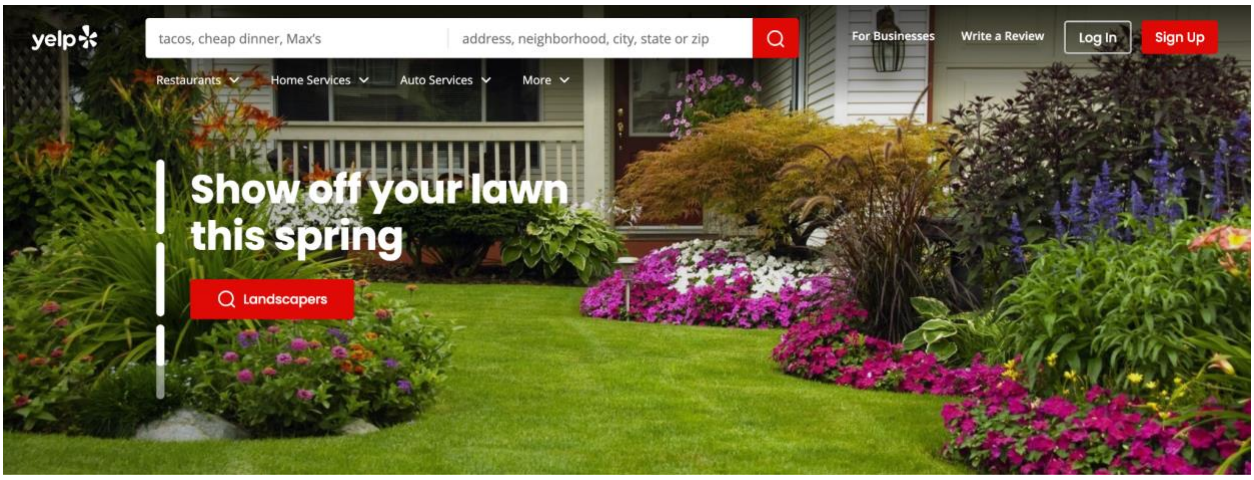
Overall, our web scraping project has provided a valuable resource for anyone interested in Korean cuisine or the Korean restaurant industry in the United States. By providing insights into the popularity and preferences of Korean cuisine in different regions, our project has the potential to contribute to the growth and success of the Korean restaurant industry.

Reference

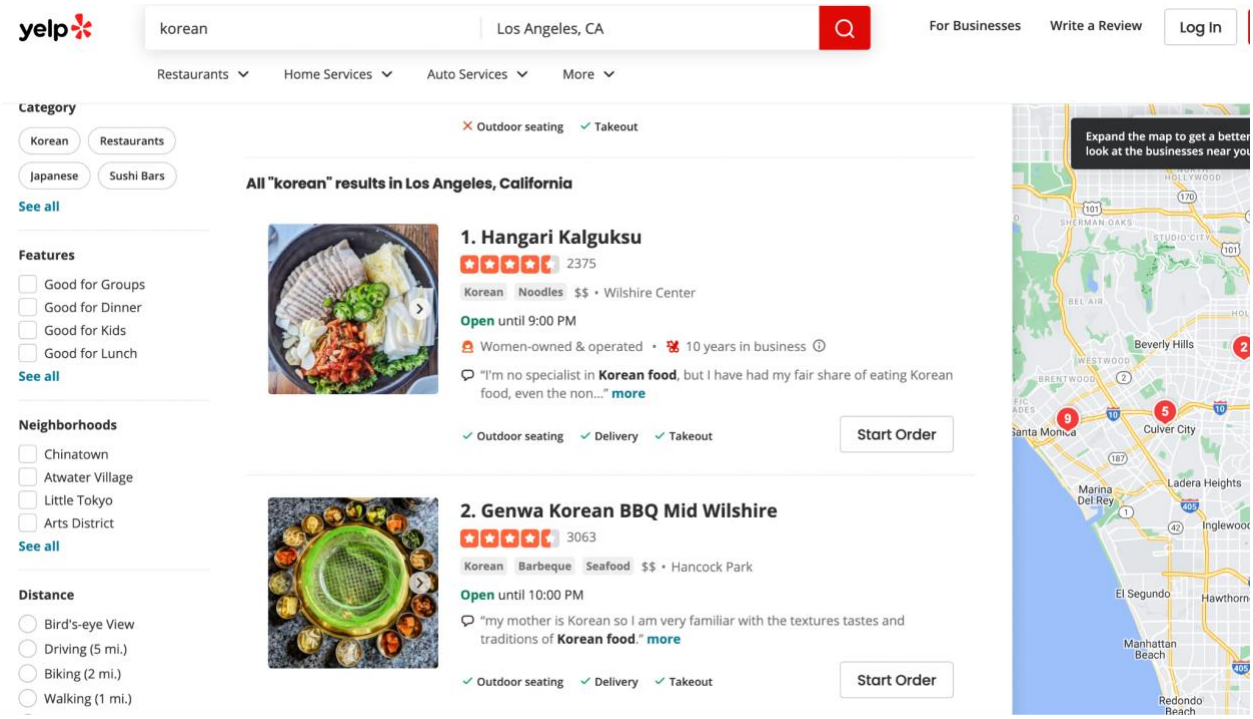
R. (2022, October 3). *The Complete List of Koreatowns in North America*. Lingua Asia. https://linguasia.com/koreatown-usa

Appendix:

Appendix 1:



Appendix 2:



Appendix 3:

```
Download Yelp Korean food in New York page_1.htm successfully
Download Yelp Korean food in New York page_2.htm successfully
Download Yelp Korean food in New York page_3.htm successfully
Download Yelp Korean food in Los Angeles page_1.htm successfully
Download Yelp Korean food in Los Angeles page_2.htm successfully
Download Yelp Korean food in Los Angeles page_3.htm successfully
Download Yelp Korean food in Chicago page_1.htm successfully
Download Yelp Korean food in Chicago page_2.htm successfully
Download Yelp Korean food in Chicago page_3.htm successfully
Download Yelp Korean food in Oakland,CA page_1.htm successfully
Download Yelp Korean food in Oakland,CA page_2.htm successfully
Download Yelp Korean food in Oakland,CA page_3.htm successfully
```

Appendix 4:

```
    for i in range(len(list_city)): #len(list_city)
        for n in range(1,4):
            with open(f'Yelp Korean food in {list_city[i]} page_{n}.htm') as f:
                soup = BeautifulSoup(f, 'lxml')
                for k in range(3,19):
                    info = soup.select(f'#main-content > div > ul > li:nth-child({k}) > div
                    for restaurant in info:
```

Appendix 5:

```
for i in range(len(df)):
    with open(f'{name_list[i]}.htm') as f:
        soup = BeautifulSoup(f, 'html.parser')
        popular_dishes_list = []

        city.append(df['City'].loc[i])
        restaurant.append(df['Name'].loc[i])
        #print(city[i],restaurant[i], address)
        # Restaurant Address

        try:
            address_soup = soup.select('div.arrange-unit__09f24__rqHTg.arrange-unit-fill__09f24__CUubG.border-color--default__09f24__NPAKY > p.css-qyp8bo')
            #address = soup.find('p', text='Phone number')
            address_tag = soup.find('p', text='Get Directions')

        # Find the next <p> tag that contains the phone number
            address = address_tag.find_next('p').text
            access_key = 'f76819fd4043da5b70bb1582a2b7523c'#This is my access key for the free account I signed up for

            api ='http://api.positionstack.com/v1/forward?access_key='+access_key+'&query='+address
            response = requests.get(api)
            geo_info = response.json()
```

Appendix 6:

```
## Convert Price to int for further calculation
if price is None:
    price_list.append(None)
elif price.text == "$":
    price_list.append(1)
elif price.text == "$$":
    price_list.append(2)
elif price.text == "$$$":
    price_list.append(3)
else:
    price_list.append(4)
```

Appendix 7:

```
## Convert string to int for further calculation
rating = re.split(" ", rating)
rating = rating[0]
rating = float(rating)
rating_list.append(rating)
ept:
rating_list.append(None)
```

Appendix 8:

```
# Connect to MongoDB
client = pymongo.MongoClient("mongodb://localhost:27017/")
db = client["ucdavis"]


#Here, if a collection (table) called bayc doesn't already

if 'restaurant' not in db.list_collection_names():
    db.create_collection('restaurant', capped=False)

restaurant = db['restaurant']
```