Analyses of Top US Restaurants on Yelp.com

MSBA 6330 Harvesting Big Data

Group 7: Shuo, Kaushik, Rohan, Wenqiuli

11/02/2015

1. INTRODUCTION

Yelp.com is a crowd-sourced local business review and social networking site. Customers can access businesses based on their locations, and can also provide reviews and ratings (1-5 scale) for a business. Yelp ratings are often considered a good metric to identify the quality of a business.

We are motivated to analyze US restaurants on Yelp.com because it is a popular site and our analyses help customers to choose the best restaurant. Also we identified suggested reviews to help restaurants improve their service quality. This project will achieve this goal through analyses of the following questions.

1. Part A: Average ratings of restaurants across states in US;
   Part B: Number of restaurants across cities in US
2. Number of popular restaurants by states/cities/categories in US
3. Identify the top 20 Chinese restaurants in Minneapolis by average rating, and help them find helpful reviews that are suggestions rather than complaints from all low-rating reviews.

2. METHODOLOGY

APIs Used	Software/Techniques Used
Yelp API	Python (Web scrapping)
Google Geocoding API V3	R
	Tableau

2.1 Data Retrieval: Fetch US restaurants data across cities

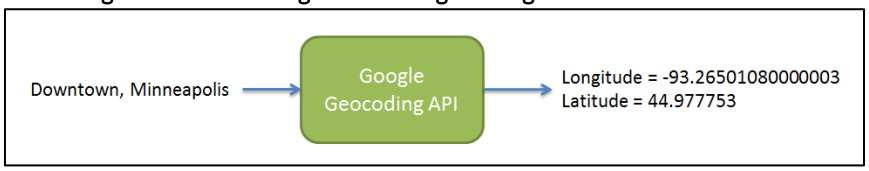
Yelp API provides a rich set of information about businesses (restaurants, bars, etc.) including their name, location, ratings, review count, contact number, menu etc. The **features** of Yelp API are: (1) provides detailed information of top 40 restaurants in a particular region; (2) restaurant preferences can be sorted by average rating, distance, or best match.

However, it also has some **limitations**. First, inefficient search to get restaurants in some locations. For example, if the user wants to search for restaurants in Madison, the API will get confused to choose between Madison, WI and Madison, IL. Second, number of reviews of a restaurant. Currently Yelp API provides only one review of any restaurant.

To find some patterns of US restaurants, this project needs a large dataset to operate on. However, Yelp API returns only 40 restaurants in a particular region, which means that the size of dataset generated would be considerably small if we fetch data based on city level, and this may influence our analyses results. Therefore, instead of finding restaurants across cities, we narrowed down to restaurants of each neighborhood in each city. Suppose we have 19 neighborhoods in Minneapolis and each neighborhood returns 40 top restaurants in it, we will get data of $19 \times 40 = 760$ restaurants in Minneapolis. This number is significantly higher than if we base on city. In later analyses we aggregate data when our target are across cities or states.

Another major challenge is to retrieve restaurant data for correct location. As mentioned in limitations, Yelp API's search by location is not efficient when it comes with places of same name. To overcome this, we use Google Geocoding API V3, which provides us with co-ordinates for a particular location given its name. Below figure explains high level functionality of use of Google Geocoding API in our project.

Figure 1: Use of Google Geocoding API to get location co-ordinates



The data scraping part can be divided into two parts:

2.1.1. Preprocessing of Input Data

As said, we need a list of neighborhoods and their respective cities to fetch the top 40 restaurants in each neighborhood. We used data from Yelp's neighborhood list and transform it into pipe delimited format – ***CityName | Neighborhood***

2.1.2. Data Scraping:

First, read the input file and create a dictionary with CityName as key and Neighborhood Names list as its value.

{City1 : [Neighbourhood1, Neighbourhood2,....],
City2 : [Neighbourhood1, Neighbourhood2,....],
...
}

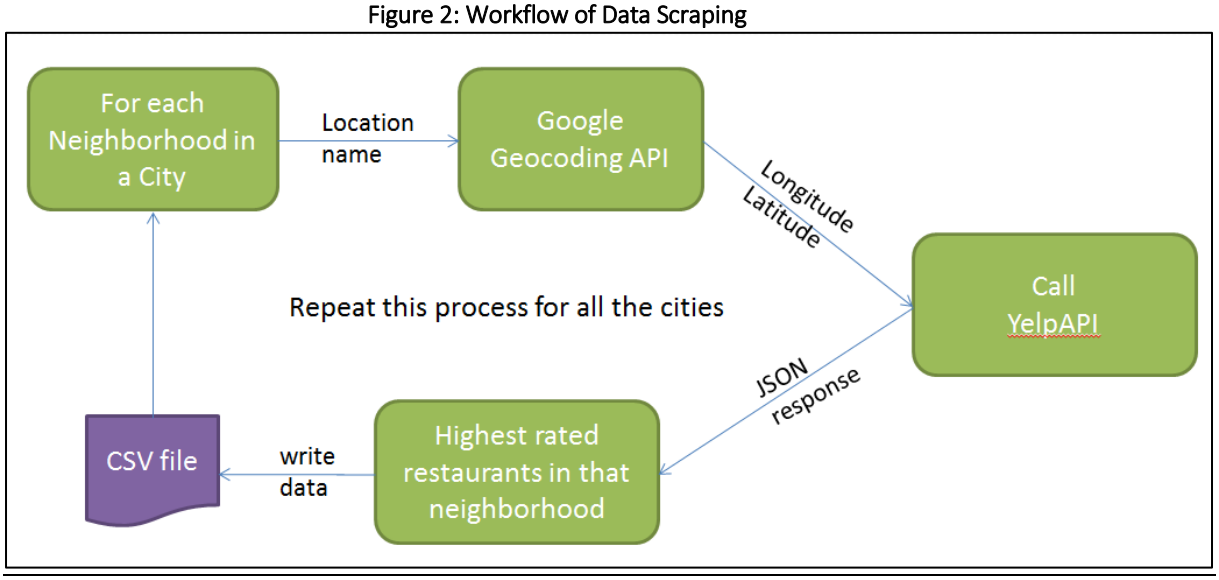
Second, for each neighborhood of city:

1. Get neighborhood coordinates from Google Maps Geocoding API
2. Retrieve restaurant data from Yelp for the neighborhood, with the following parameters specified:

term = '***restaurants***'	sort_filter = 'Highest Rated'
Neighborhood location Latitude and longitude (received from Google Geocoding API)	OFFSET (We can only get 20 restaurants information for each call through API, so we run twice for each neighborhood to get all 40 records)

3. Scrape the information for each restaurant: *Id; Name; Review_Count; Rating; Longitude; Latitude; City; State; Zip; Country; Categories*.
4. Write scrapped data into a pipe delimited file for each neighborhood;
5. Output File: The output file will be generated as '***/output/restaurants.csv***' folder.

Figure 2: Workflow of Data Scraping



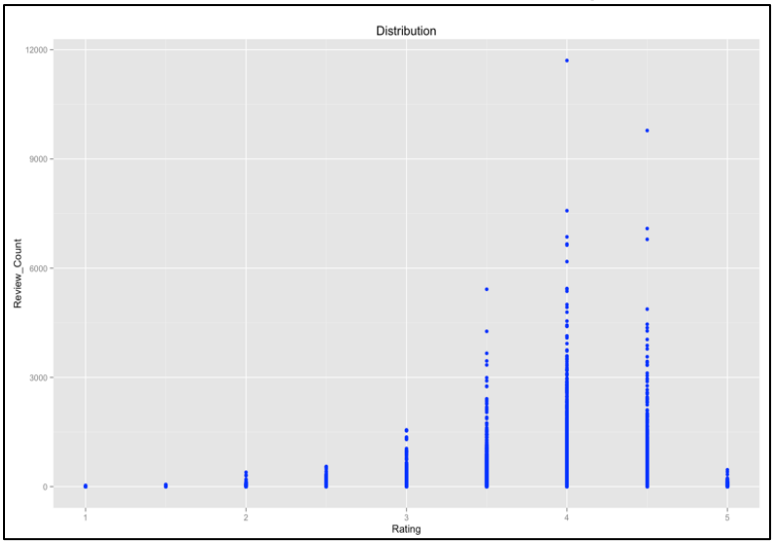
2.2 Overview of the dataset

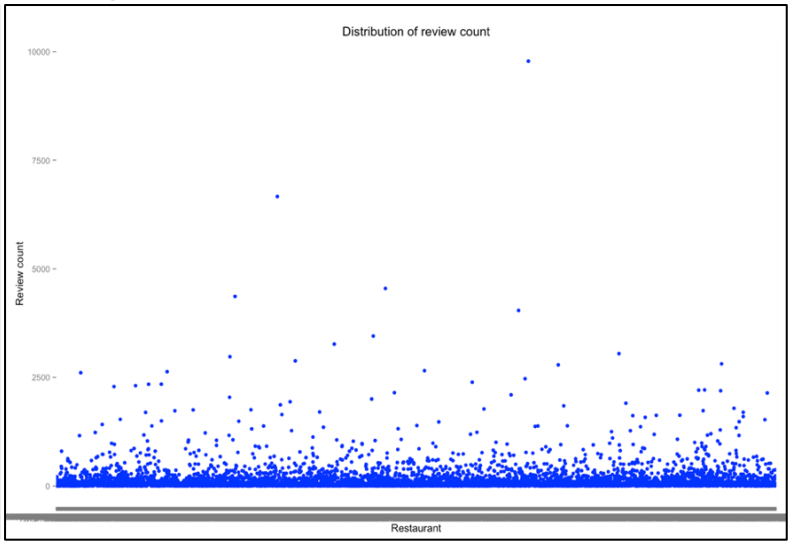
Through our API query, there are 134,893 restaurant records retrieved. After removing all duplicated and noise restaurant data (no zip code, no location co-ordinates, etc.), we have 60,395 restaurant records. The dataset has 11 variables: restaurant ID, restaurant name, review count, rating, longitude, latitude, city, state, zip code, country, and category. Among these attributes, review count and rating can reflect whether customers favor to visit a restaurant or not.

The left figure below shows that most restaurants with rating smaller than 3 have low review counts, but there is no clear relationship between ratings and review counts for higher rated

restaurants. In the right figure, we randomly sample 10% of the restaurants and plot their review counts. It shows that most restaurants have review counts under 400.

Figure 3: Distribution of Ratings and Review count





2.3 Analysis Approach

2.3.1 Average ratings of restaurants across states/cities in US

We keep only cities with more than 100 restaurants to avoid extreme cases and better present the pattern across cities and across states. Then the data size reduces into 47,174 records. We also create a unique key concatenating State and City name (State_City) since there are restaurants in cities with the same name, yet in a different state (e.g., city called Madison is in 9 different states).

2.3.2. Number of popular restaurants across states/cities/categories in US

From all the restaurants we want to explore more about popular restaurants which customers may care about more. We use rating and review count as our KPIs to identify popular businesses. That is, we keep only restaurants with average rating higher than 4 (since the median rating is 4), and review counts more than 100 (since it is at 70th percentile of data). Then we have 2,692 popular restaurant records left in our data.

2.3.3. Identifying suggested reviews

Yelp API does not provide all reviews of a restaurant through its API, so we use Yelp API to identify the Top 20 Chinese restaurants in Minneapolis by their average ratings, then use Web Scraping techniques to obtain all these restaurants' reviews. To be more concrete, we use "Search" method in Yelp API with corresponding parameters to fetch the Top 20 Chinese restaurants' ID in Minneapolis, then use Web Scraping techniques to obtain their reviews page by page according to the ID. To help restaurants identify suggested reviews, we chose a subsample of reviews whose rating are 1.0, 2.0, or 3.0, because these low-rating reviews are either pure complaints or with suggestions. Then we set up a dictionary ('suggestion', 'suggest', 'should', 'improve', 'improvement', 'future', 'hope', 'wish') assuming that reviews with these keywords are helpful. By matching this dictionary with all reviews, we are able to help restaurants find out suggested reviews that are valuable to them.

3. FINDINGS

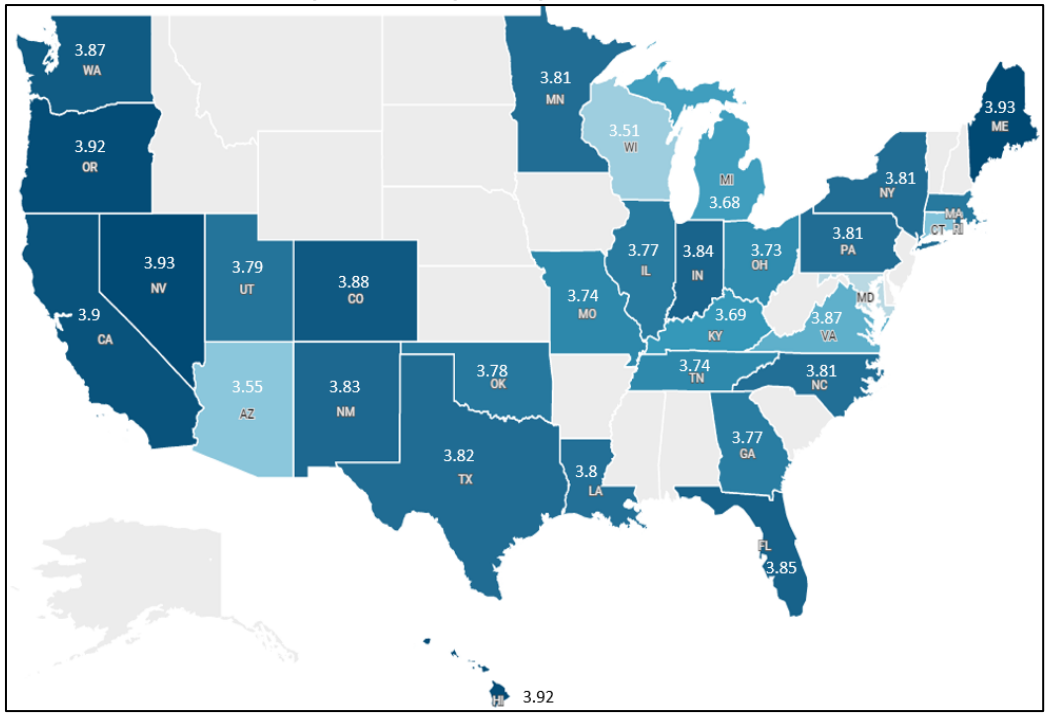
3.1. Average rating across states in US

Fig. 3 shows the. From the heat map (Fig. 3), Maine (ME) and Nevada (NV) have the highest rating while Maryland (MD) and Wisconsin (WI) have the lowest average rating.

3.2. Number of restaurants across cities in US

For this analysis we grouped the data by State_city (unique key). Chicago, Los Angeles and Houston are the top 3 cities with the most number of restaurants in US.

Figure 4: Average Ratings across States in US



3.3. Number of popular restaurants across states/cities/categories in US

Figure 5: Top 10 cities with highest number of Popular Restaurants

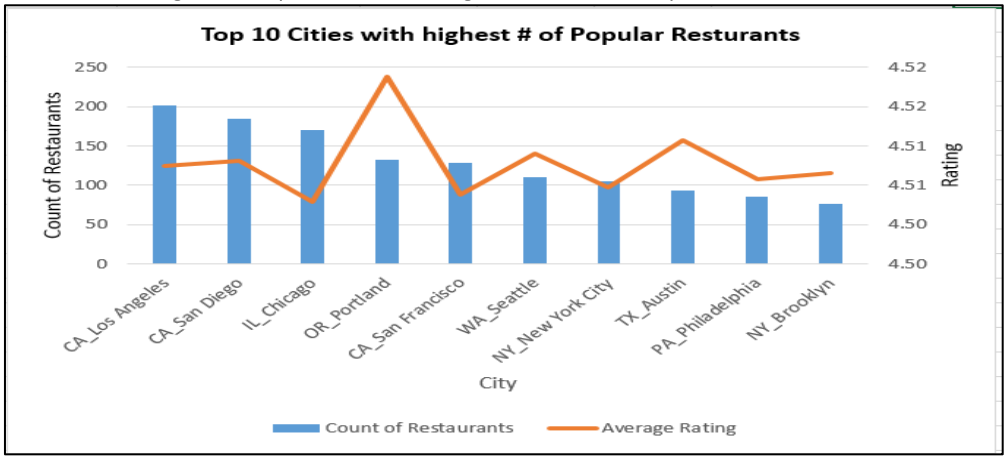


Figure 6: Top 10 Categories Offered by Popular Restaurants

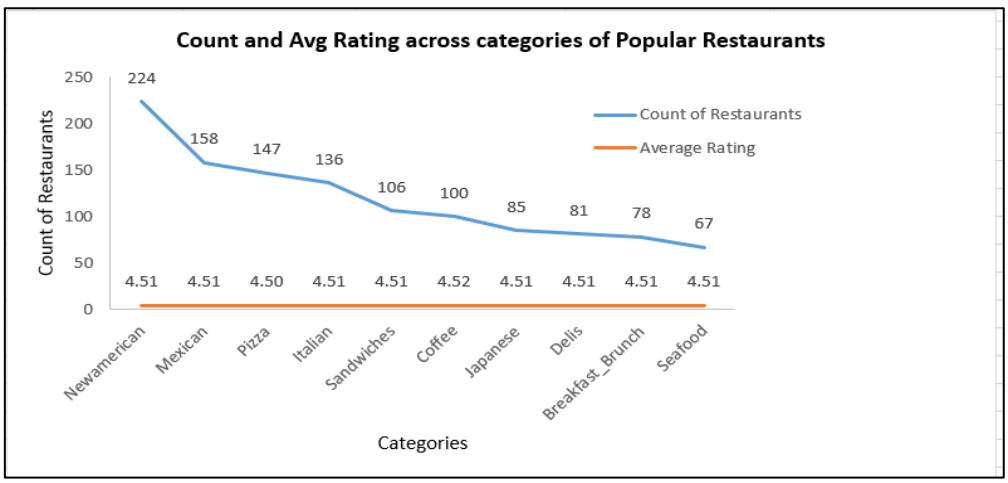
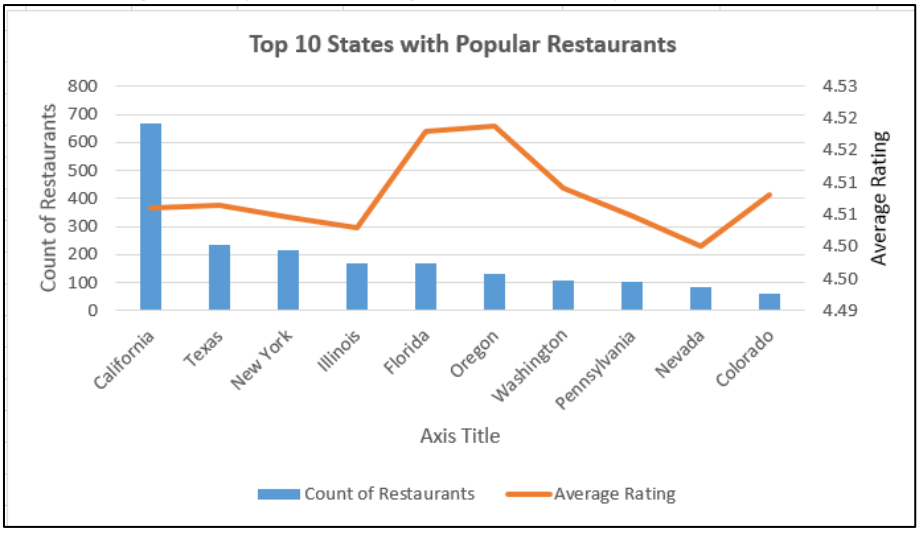


Figure 7: Top States with highest number of Popular Restaurants



3.4. Text Analysis—suggested reviews

Top 20 Chinese restaurants in Minneapolis are: Dragon House, Luis Chow Mein, Hot Wok Delivery, Xin Wong, Fresh Wok, Szechuan Spice, Kowloon, Kinhdo Vietnamese, Evergreen, Shuang Cheng, Chin Dian Cafe, Keefer Court Bakery and Cafe, Que Viet Village House One, Tea House, China Express, Kims Vietnamese, Camdi, Mei Inn, Hong Kong Noodle.

They have 1,389 reviews in total, 469 of them are low-rating reviews. Our analysis shows that 69 of the 469 low-rating reviews are actually suggestions rather than pure complaints, implying that restaurants shall filter their low-rating reviews and focus on these suggested ones to improve their service quality.

4. CONCLUSION

This project uses Yelp API and Web Scraping techniques to get some insights from Top restaurants across US. The main findings are as follows.

1. For all restaurants across the US, Maine and Nevada have the highest average rating of restaurants; Chicago, Los Angeles, and Houston have the highest number of restaurants.
2. For popular restaurants, California has the most number of popular restaurants; Los Angeles and San Diego are the top 2 cities with the highest number of popular restaurants; New American, Mexican, and Pizza are the top 3 cuisines offered by popular restaurants.
3. By text analysis we help restaurant identify suggested reviews that are helpful to them.

This project has one limitation about text analysis that we did not find a more advanced method to filter "suggested reviews" than matching them with the dictionary we set up. Future analysis may work on this (e.g., manually label 10,000 complaints and 10,000 suggestions, find different patterns of each group, then use it to label other reviews).

Another future direction is about the criterion of "popular restaurants". Instead of using review count as one criterion, we may consider the length of opening of restaurants to normalize review count. The metric is then

$$\text{Average Reviews per Day} = \text{Review Count}/(\text{Data Extracted Date(Oct 25th, 2015)} - \text{First Review Date})$$

We created a new variable 'Number of Days' by scrapping the first review date of the restaurant using Restaurant ID. Using this 'Number of Days' Average Reviews per day can be calculated. Further analysis has to be done to understand why some restaurants have higher average review count (location, customer attributes etc.).