IBM Applied Data Science

Capstone Project:
The Battle of Neighbourhoods in Shanghai

Xiaoyue Zou

July 2020

Table of Contents

Introduction

Background

Shanghai, one of the most prosperous cities in China, is the economic, financial, trade and shipping center.
In Shanghai, you can not only meet the trendy frontier, but also explore the history. More than a hundred years ago, after setting Shanghai as a port for the convenience of transportation, thousands of merchants gathered, the industry was prosperous, and the inheritance of culture was orderly.

The metropolitan city, known as 'the city that never sleeps', glows with a charming demeanor from beginning to end. It is both nostalgic and modern, rich in oriental charm and western flavor, which is fascinating.

Description of the business problem

In this project, Foursquare data along with Shanghai district information will be used by machine learning methods (K- Means) collaboratively to cluster districts based on different categories' ratings/frequencies. Not only benefiting the travelers who are interested in some Chinese tastes, but also locals who are exploring the city.

Data

The information of the administrative division of Shanghai was retrieved from Wikipedia. The table contains district names (including Chinese and Pinyin), Division code, Area, Population and Density. For further analysis, irrelevant information such as the Pinyin column was deleted.
The geographical coordinates of each districts were generated by using geocoder class of Geopy client. They were added in the district information table.
Using Foursquare APIs, information on nearby venues in each shanghai districts was collected including different categories such as restaurants, gym, café, etc. Later on, it was possible to filter these data to get specific categories.

Methodology

In the notebook, I present the procedure of clustering the Shanghai districts (Neighborhoods) according to the most frequent venues retrieved by Foursquare API. Data preprocessing and cleaning steps were conducted on the district information as well as adding spatial location data. One hot encoding was used for the venue categories in order to create model for machine learning. Here, K-Means clustering was selected for its convenience and high quality in terms of accuracy. Finally, five clusters of Shanghai districts were created offering city insights for both visitors and locals.

Data Wrangling

Scraping Shanghai districts from Wikipedia

In order to get a dataframe, Pandas package was used to transform and rename the columns shown as below (first five rows).

	Neighborhood	Chinese	Area	Population	Density
0	Huangpu District[4](City seat)	黄浦区	20.46	653800	31955
1	Xuhui District	徐汇区	54.76	1084400	19803
2	Changning District	长宁区	38.30	694000	18120
3	Jing'an District	静安区	36.88	1062800	28818
4	Putuo District	普陀区	54.83	1281900	23380

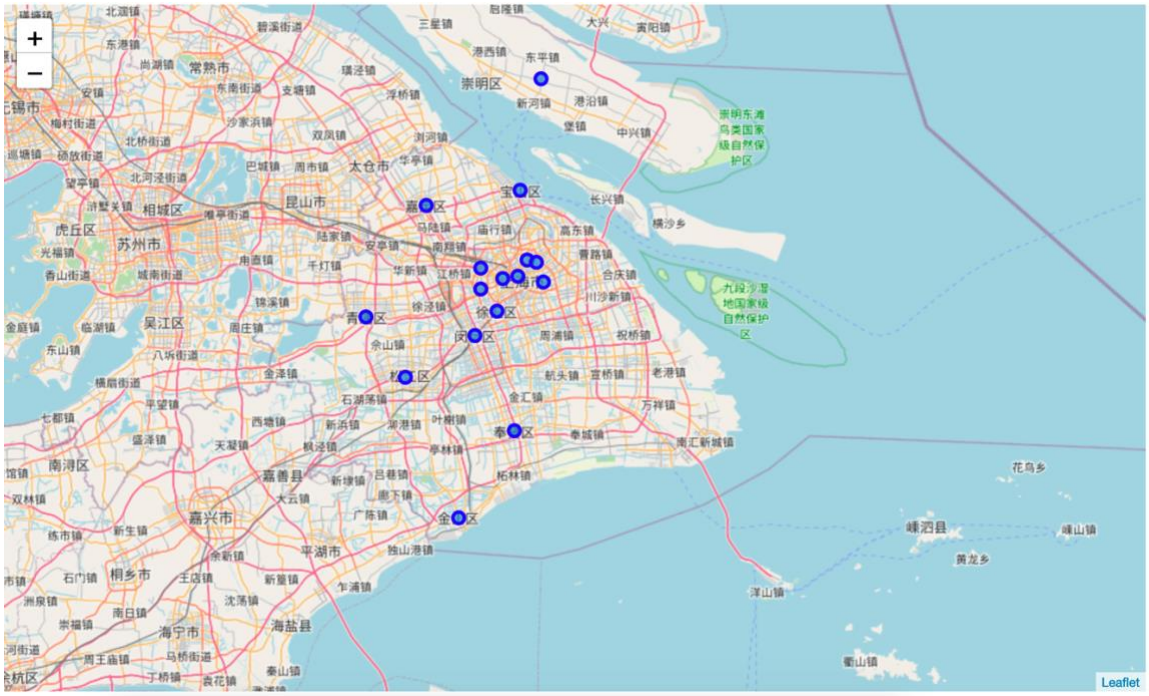
In this stage, values were modified, and irrelevant columns were deleted.

Adding spatial data using geocoders from Geopy Package

A table with 16 rows and 7 columns was generated finally.

	Neighborhood	Chinese	Area	Population	Density	Latitude	Longitude
0	Huangpu District	黄浦区	20.46	653800	31955	31.233593	121.479864
1	Xuhui District	徐汇区	54.76	1084400	19803	31.163698	121.427994
2	Changning District	长宁区	38.30	694000	18120	31.209276	121.389986
3	Jing'an District	静安区	36.88	1062800	28818	31.229776	121.443060
4	Putuo District	普陀区	54.83	1281900	23380	31.251326	121.391229
5	Hongkou District	虹口区	23.48	797000	33944	31.266703	121.501751
6	Yangpu District	杨浦区	60.73	1312700	21615	31.262011	121.521430
7	Pudong New Area	浦东新区	1210.41	5550200	4585	31.221783	121.538740
8	Minhang District	闵行区	370.75	2543500	6860	31.114767	121.376943
9	Baoshan District	宝山区	270.99	2042300	7536	31.406634	121.485158
10	Jiading District	嘉定区	464.20	1588900	3423	31.377756	121.260612
11	Jinshan District	金山区	586.05	805000	1374	30.744817	121.337257
12	Songjiang District	松江区	605.64	1762200	2910	31.029593	121.210838
13	Qingpu District	青浦区	670.14	1219100	1819	31.152164	121.119552
14	Fengxian District	奉贤区	687.39	1152000	1676	30.920449	121.469383
15	Chongming District	崇明区	1185.49	688100	580	31.631339	121.533777

A map of Shanghai was drawn using Folium Package to visualize the locations of each district.



Accessing Foursquare data via API

Venues within different categories in the nearby neighborhood can be retrieved via Foursquare API. A new data table was made in order to record no more than 100 nearby venues within 500m radius for each district.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Huangpu District	31.233593	121.479864	Campanile Hotel and Restaurant	31.232123	121.479144	Hotel
1	Huangpu District	31.233593	121.479864	M1NT Restaurant & Grill	31.236920	121.479641	Restaurant
2	Huangpu District	31.233593	121.479864	The Westin Bund Center (上海威斯汀大饭店)	31.233935	121.482653	Hotel
3	Huangpu District	31.233593	121.479864	Old Beijing Qianmen Roast Duck (老北京前门烤鸭)	31.232480	121.482457	Peking Duck Restaurant
4	Huangpu District	31.233593	121.479864	Épices & Foie-gras	31.237557	121.479580	French Restaurant
5	Huangpu District	31.233593	121.479864	台北纯K	31.230826	121.477331	Karaoke Bar
6	Huangpu District	31.233593	121.479864	M1NT	31.236609	121.479798	Nightclub
7	Huangpu District	31.233593	121.479864	Foreign Languages Bookstore (外文书店)	31.235917	121.478498	Bookstore
8	Huangpu District	31.233593	121.479864	东菜 海上	31.234365	121.477760	Seafood Restaurant
9	Huangpu District	31.233593	121.479864	Tock's	31.236397	121.481310	Deli / Bodega

One hot encoding was used then, and rows were grouped by neighborhood and by taking the mean of the frequency of occurrence of each category. Five most common venues in each district can be viewed. An example is presented below.

```
----Baoshan District----
                venue  freq
0  Asian Restaurant  0.25
1       Coffee Shop  0.25
2              Park  0.25
3        Hobby Shop  0.25
4       Art Gallery  0.00


----Changning District----
           venue  freq
0      BBQ Joint   0.2
1  Noodle House    0.1
2       Tea Room   0.1
3   Coffee Shop    0.1
4       Gym Pool   0.1
```

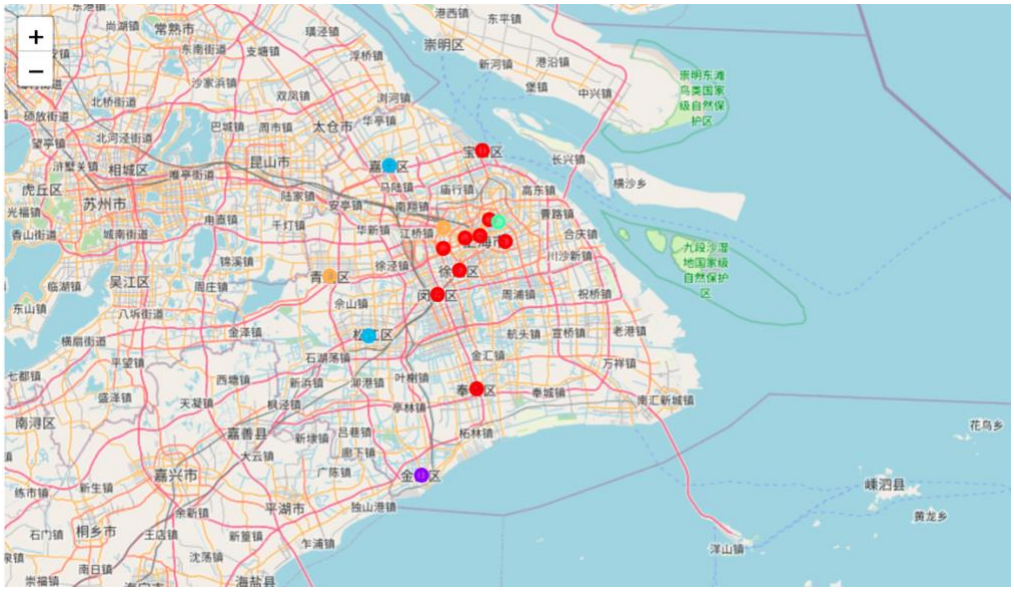
In order to build a model which contains more data for machine learning, 10 most common venues were included for each suburb.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Baoshan District	Asian Restaurant	Coffee Shop	Hobby Shop	Park	Xinjiang Restaurant	Grocery Store	Doner Restaurant	Dumpling Restaurant	Fast Food Restaurant	Flea Market
1	Changning District	BBQ Joint	Gym Pool	Art Museum	Dumpling Restaurant	Fast Food Restaurant	Hotel	Tea Room	Coffee Shop	Noodle House	Gym
2	Fengxian District	Plaza	Grocery Store	Asian Restaurant	Auto Garage	Steakhouse	Shanghai Restaurant	Fruit & Vegetable Store	Dessert Shop	Dim Sum Restaurant	Doner Restaurant
3	Hongkou District	Bakery	Chinese Restaurant	Multiplex	Plaza	Xinjiang Restaurant	Gym	Doner Restaurant	Dumpling Restaurant	Fast Food Restaurant	Flea Market
4	Huangpu District	Hotel	Coffee Shop	Bookstore	Chinese Restaurant	French Restaurant	Bar	Restaurant	Italian Restaurant	Gym	Shanghai Restaurant

Machine learning

K-Means clustering was used with k = 5.

Result



Cluster 1

	Chinese	Latitude	Longitude	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	黄浦区	31.233593	121.479864	0	Hotel	Coffee Shop	Bookstore	Chinese Restaurant	French Restaurant	Bar	Restaurant	Italian Restaurant	Gym
1	徐汇区	31.163698	121.427994	0	Italian Restaurant	Supermarket	Shopping Mall	Buffet	Xinjiang Restaurant	Grocery Store	Doner Restaurant	Dumpling Restaurant	Fast Food Restaurant
2	长宁区	31.209276	121.389986	0	BBQ Joint	Gym Pool	Art Museum	Dumpling Restaurant	Fast Food Restaurant	Hotel	Tea Room	Coffee Shop	Noodle House
3	静安区	31.229776	121.443060	0	Coffee Shop	Café	Japanese Restaurant	Hotel Bar	Dumpling Restaurant	Gym / Fitness Center	Pizza Place	Shanghai Restaurant	Hotel
5	虹口区	31.266703	121.501751	0	Bakery	Chinese Restaurant	Multiplex	Plaza	Xinjiang Restaurant	Gym	Doner Restaurant	Dumpling Restaurant	Fast Food Restaurant
7	浦东新区	31.221783	121.538740	0	Fast Food Restaurant	Coffee Shop	Ramen Restaurant	Plaza	Steakhouse	Performing Arts Venue	Convenience Store	Science Museum	Flea Market
8	闵行区	31.114767	121.376943	0	Fast Food Restaurant	Bakery	Metro Station	Café	Plaza	Xinjiang Restaurant	Gym	Doner Restaurant	Dumpling Restaurant
9	宝山区	31.406634	121.485158	0	Asian Restaurant	Coffee Shop	Hobby Shop	Park	Xinjiang Restaurant	Grocery Store	Doner Restaurant	Dumpling Restaurant	Fast Food Restaurant
14	奉贤区	30.920449	121.469383	0	Plaza	Grocery Store	Asian Restaurant	Auto Garage	Steakhouse	Shanghai Restaurant	Fruit & Vegetable Store	Dessert Shop	Dim Sum Restaurant

Cluster 2

	Chinese	Latitude	Longitude	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Com Venu
11	金山区	30.744817	121.337257	1	Fast Food Restaurant	Supermarket	Xinjiang Restaurant	Gym	Dim Sum Restaurant	Doner Restaurant	Dumpling Restaurant	Flea Market	French Restaurant	Fr Vege

Cluster 3

	Chinese	Latitude	Longitude	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Cor Ven
10	嘉定区	31.377756	121.260612	2	Hotel	Chinese Restaurant	Xinjiang Restaurant	Gym	Doner Restaurant	Dumpling Restaurant	Fast Food Restaurant	Flea Market	French Restaurant	Veg
12	松江区	31.029593	121.210838	2	Hotel	Convenience Store	Chinese Restaurant	Xinjiang Restaurant	Gym	Doner Restaurant	Dumpling Restaurant	Fast Food Restaurant	Flea Market	Res

Cluster 4

	Chinese	Latitude	Longitude	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
6	杨浦区	31.262011	121.52143	3	Coffee Shop	Fast Food Restaurant	Museum	Xinjiang Restaurant	Gym	Doner Restaurant	Dumpling Restaurant	Flea Market	French Restaurant	Fruit & Vegetable Store

Cluster 5

	Chinese	Latitude	Longitude	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Comm Venue
4	普陀区	31.251326	121.391229	4	Stadium	Hotel	Fast Food Restaurant	Szechuan Restaurant	Motel	Gym	Dim Sum Restaurant	Doner Restaurant	Dumpling Restaurant	Ma
13	青浦区	31.152164	121.119552	4	Fast Food Restaurant	Asian Restaurant	Hotel	Xinjiang Restaurant	Gym / Fitness Center	Doner Restaurant	Dumpling Restaurant	Flea Market	French Restaurant	Fru Vegeta S

Conclusion

Through the list of these five clusters, we have no way to clearly know the difference between each neighborhood. However, it can be seen from the map that the city center of Shanghai is more concentrated. There are two possible reasons for this result.

First, Shanghai is highly urbanized, and the infrastructure in each district is relatively even, resulting in an even distribution of the number of complexes such as restaurants and shops. The second is that the data obtained from the Foursquare website may not be complete, as a lot of data are related to restaurants and hotels, while city parks and other types of attractions are lack of reviews and ratings. This could also be related to the fact that the website is not a mainstream social network in China, so that the data used here lacks credibility.

Reference

https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Shanghai

https://medium.com/@radialee/capstone-project-the-battle-of-neighborhoods-in-tokyo-restaurants-45a503e65ff