APPLIED DATA SCIENCE CAPSTONE

Project: The Battle of Neighborhoods

Opening a new shopping mall in Beijing

Yvette Lu

2020/10/03

Introduction

A large shopping mall is undoubtedly a good place for entertainment. In the mall, people can relax and enjoy themselves. Within a short walk, they can reach a variety of shops, such as grocery stores, restaurants, cinemas, supermarkets, fruit shops and even indoor amusement parks. Shopping centers can meet the needs of many types of consumers. However, the location of a shopping center is very important for the profit of a store. It should be located in a prominent and convenient place where there is a lot of people. Although there are many concentrated and huge shopping centers in the center of Beijing, there are still many places where residents need shopping centers closer to them. Besides, the location of shopping malls plays an important role in the profits of dealers and real estate developers.

● **Business Problem**

The goal of the project is to use a data-science approach, using various algorithms, such as clustering, to select a location in Beijing for real estate developers or overseas investors to build a new shopping mall.

● **Who would be interested**

In recent years, shopping malls in Beijing have been mushrooming, and their development has been stable and orderly. This is due to the strong strength and advanced concepts of developers, as well as the superior location of site selection, and the continuous optimization and adjustment of developers by relying on large-scale residential communities. In the words of Jianhui Han, the vice President of China Urban Commercial Network Construction management Federation, Beijing's commercial project status could be characterized by 'rapid development, fierce competition, survival of the fittest and wide traps' [1]. Some industry watchers said, Although commercial properties are relatively concentrated in the urban center of Beijing, the economic development is stable and mature, old projects are being renovated and upgraded, and the sub-center of the city is becoming more and more perfect, with the feature of multi-center becoming increasingly obvious.

Commercial property still has a broad space for development and is still the focus of many large commercial real estate developers and brand retailers[2].

Data

● **Description of data**

We might need the following data to solve the problem above:
1. A list of neighborhoods in Beijing, especially the prominent neighbors (In this project, Ethnic enclaves might not be included).
2. The latitude and longitude coordinates of these neighborhoods.
3. The site data necessary related to the mall.

- **How to use the data**

  A list of major neighbourhoods in Beijing, including neighbors like Qianmen and Sanlitun, etc. could be crawled through the Wikipedia page [3]. There are 41 communities in total.

  By using the beautifulsoap package, we can crawl data from the Wikipedia and use the Python Geocoder package to get the geographical coordinates of our neighbours, which can give us the latitude and longitude coordinates of our neighbours.

  After that, we'll use Foursquare APl to get the location data for these communities. By using this API, we can get the data we want, especially the mall data, then data cleansing and machine learning techniques would be applied, then it's time to visualize the data using Folium.

Methodology

First, we need to get the community list of Beijing, this could be done through Wikipedia page(https://en.wikipedia.org/wiki/Neighborhoods_in_Beijing) crawling Beijing Prominent neighbors. Although there are many ethnic enclaves[3] in Beijing, because of the geographical position, the project does not include the situation of the ethnic enclaves. Second, we'll do web fetching using a Python request and beautiful Soup package, extracting all the 'li' section in a 'div' section to get the list of names. However, there are some Chinese characters in the name list, and we need to replace all characters except alphabets, numbers and spaces with nothing. We also need to use the geocoder package to get the geographic coordinates in the form of latitude and longitude so that we can use the Foursquare API. The geographical dataset was merged into the same dataframe with the original data and was visualized using the Folium package.

Next, we used the Foursquare API to get the top 100 places within a radius of 2,000 meters. In this step, we make an API call to Foursquare, passing the community's geographic coordinates through a Python loop. Foursquare will return venue data in JSON format, and we will extract venue names, venue categories, venue latitude and longitude, check how many venues are returned from each community, and how many unique categories can be selected from all returned venues. Then, we analyze each neighborhood separately, grouped by neighborhood, and calculate the average frequency for each site category. By doing so, the data for subsequent clustering is prepared, and the "shopping mall" is selected for filtering as the location category of adjacent areas.

Finally, we will use k-means clustering to cluster the data. The K-means clustering algorithm is one of the simplest and most popular unsupervised machine learning algorithms. We will divide the neighbourhoods into three clusters according to the frequency of the "shopping mall". The results will enable us to determine which areas have a higher concentration of shopping centers, and which communities have a smaller number of shopping centers. Through the emergence of shopping centers in different communities, it will help us answer which areas are most suitable for

opening up new shopping centers.

Result

The 3 clusters I got are as follows:

Cluster 1:

```
k1_merged.loc[k1_merged['Cluster Labels'] == 0]
```

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
18	Jianguomen	0.030000	0	39.91460	116.41671
33	Yayuncun	0.021739	0	40.01388	116.39644
31	Xizhimen	0.016393	0	39.93889	116.35028
30	Xidan	0.020000	0	39.90750	116.39723
27	Wangjing	0.030000	0	39.99330	116.47284
25	Shifoying	0.020000	0	39.90750	116.39723
24	Sanlitun	0.020000	0	39.93609	116.44375
22	Pinganli	0.020000	0	39.90750	116.39723
20	Kuanjie	0.020000	0	39.90750	116.39723
36	Zhongguancun	0.030000	0	39.98111	116.30889
17	Huashi	0.020000	0	39.90750	116.39723
16	Hepingmen	0.035294	0	39.89996	116.37435
14	Guomao	0.020000	0	39.90750	116.39723
37	Zuoanmen	0.023810	0	39.86929	116.43320
11	Gongzhufen	0.020000	0	39.90750	116.39723
1	Beijing CBD	0.020000	0	39.90750	116.39723
2	Chaoyangmen	0.030000	0	39.91460	116.41671
3	Chongwenmen	0.020000	0	39.89972	116.41222
8	Dongzhimen	0.030000	0	39.93596	116.43027
7	Dongdan	0.030000	0	39.91460	116.41671
4	Dengshikou	0.020000	0	39.90750	116.39723

Cluster 2:

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
12	Guanganmen	0.00	1	39.88806	116.34194
34	Yongdingmen	0.00	1	39.87229	116.10337
28	Wudaokou	0.00	1	39.99257	116.33208
13	Guangqumen	0.01	1	39.89199	116.43661
5	Deshengmen	0.00	1	39.94126	116.37929
19	Jiaodaokou	0.01	1	39.93987	116.40677
15	Hepingli	0.00	1	39.95276	116.41093
6	Dianmen	0.00	1	39.93943	116.39301
0	Andingmen	0.00	1	39.94382	116.39952

Cluster 3:

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
26	Tiananmen	0.040000	2	39.90444	116.39139
23	Qianmen	0.040000	2	39.89290	116.40607
21	Panjiayuan	0.057143	2	39.88392	116.47576
29	Xibianmen	0.053333	2	39.89778	116.35028
9	Fuchengmen	0.049180	2	39.92295	116.34780
32	Xuanwumen	0.040541	2	39.89833	116.36806
10	Fuxingmen	0.043478	2	39.90556	116.35111
35	Youanmen	0.060606	2	39.86554	116.35758

Discussion

Most of the shopping centers are concentrated in the central area of Beijing, with the largest number of shopping centers in Cluster 3 and the moderate number of shopping centers in Cluster 1. Cluster 2 hardly has shopping malls. So, this area has a huge opportunity and high potential for the construction of new shopping malls, as there is little competition for existing shopping centers. At the same time, the malls in Cluster 3 May face intense competition due to the high concentration and oversupply

of shopping malls, so the developer need to avoid opening a new shopping mall there.

Conclusion

In this project, we have obtained the required data through crawlers, extracted and prepared the data, carried out machine learning data, and distributed them to the three clusters based on their similarities through clustering. Finally, we provided Suggestions to stakeholders, namely the best location for the real estate developers and investors for the new shopping city.The answer for this project is that the area adjacent to Cluster 2 is the preferred location for new shopping centers.

The findings of the project will help stakeholders to avoid competitive areas and take advantage of high-potential sites when deciding to build new shopping centres.

References

[1] Chu Qian, From "Wolf" to "cat" -- Ten years of Changes in Beijing shopping Malls [J].China Market, 2008(4):36-36.
[2] Ninghui Chen, List of Shopping Centers to be opened in Beijing 2020: Dongba Wanda Plaza, Joy in the Spring Breeze... Lianshang Retail research Center. *http://www.linkshop.com.cn/web/archives/2020/444113.shtml*
[3] Wikipedia page: *https://en.wikipedia.org/wiki/Neighborhoods_in_Beijing*