The Battle of Neighborhoods

Applied Data Science Capstone Project

Wei Wang

1. Introduction

1.1 Background

According to UN Department of Economic and Social Affairs (DESA), there are nearly 272 million international migrants worldwide in 2019. Toronto and Shanghai, two major well known cities in Canada and China, are the economic centers and financial capitals of their countries. In 2019, Toronto welcomed 117,000 immigrates while Shanghai's population is growing at the rate of 700,000 people a year. Despite the fact that Toronto and Shanghai are far away from each other, it is still possible to find the similarities between these two cities.

1.2 Business problem

Relocation can be considered a big decision for a person. In addition to job opportunities, environment and culture shock are also big concerns for migrants when moving to another city. Therefore, it is advantageous for them to find a similar neighborhood in the new city as the one they live in before. In this project, we will adopt machine learning tools to cluster Toronto and Shanghai neighborhoods in order to recommend the neighborhoods which are the best choices for migrants based on surrounded essential facilities such as school, hospital, and stores etc.

As a result, the audiences of this report will be anyone who is planning to move from one city to another. This recommendation system can also benefit stakeholders who are interested in citing a business in a new city.

2. Data

We will use the following datasets for this project:

- Toronto.csv that consists of Toronto's postcodes boroughs, neighborhoods.
  Data source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Shanghai.csv that consists of Shanghai's city name, districts and subdistrict.
  Data source: https://en.wikipedia.org/wiki/List_of_township-level_divisions_of_Shanghai

PostalCode	Borough	Neighbourhood	City	Borough	Neighbourhood
M3A	North York	Parkwoods	Shanghai	Baoshan	Wusong
M4A	North York	Victoria Village	Shanghai	Baoshan	Youyi
M5A	Downtown Toronto	Regent Park / Harbourfront	Shanghai	Baoshan	Zhangmiao
M6A	North York	Lawrence Manor / Lawrence Heights	Shanghai	Baoshan	Dachang
M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government	Shanghai	Baoshan	Gaojing

Table 1: Toronto and Shanghai Neighbourhoods

- A csv file that has the geographical coordinates of each postal code for neighbourhoods in Toronto is provided.
  Data source: http://cocl.us/Geospatial_data
- Foursquare API search feature will be used to collect neighborhood venues information, which will be used to explore and compare geographical locations of Toronto and Shanghai.

3. Methodology

3.1 Obtain Coordinates

Next step is to obtain latitude and longitude coordinates of each neighbourhood in Toronto and Shanghai. Given that Geocoder package can be very unreliable, we will use a csv file that has the geographical coordinates of each postal code for neighbourhoods in Toronto. The geographical coordinates for neighborhoods in Shanghai are not provided; therefore, we will use Geopy library and Nominatim API.

City	Borough	Neighbourhood	Latitude	Longitude
Toronto	North York	Parkwoods	43.753259	-79.329656
Toronto	North York	Victoria Village	43.725882	-79.315572
Toronto	Downtown Toronto	Regent Park / Harbourfront	43.654260	-79.360636
Toronto	North York	Lawrence Manor / Lawrence Heights	43.718518	-79.464763
Toronto	Downtown Toronto	Queen's Park / Ontario Provincial Government	43.662301	-79.389494

Table 2: Geometric Coordinates of Neighbourhoods in Toronto

3.2 Data Visualization

Now we have obtained the lists of boroughs, neighbourhoods and their respective geometric coordinates for Toronto and Shanghai. Let's use Folium library to plot a map of all the neighbourhoods in each city. The neighbourhoods belong to the same borough are plotted with the same color. If we look at the maps carefully, we will notice that there are some outliers do not lie into their boroughs due to wrong location data. We will simply clean it out for better analysis.

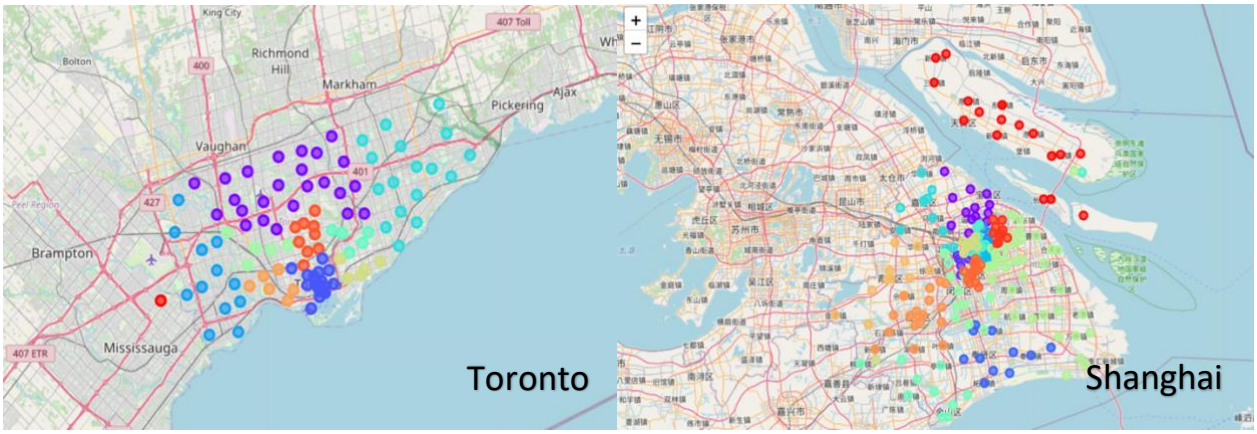


Figure 1: Maps of Toronto and Shanghai Neighbourhoods

3.3 Foursquare API Search Feature

Foursquare API provides access to massive datasets of location data and venues information including address, images, tips, ratings and comments. In this project, we will use Foursquare API and Geopy data to locate nearby venues within 500 meters of each neighbourhood in Toronto and Shanghai. After defining Foursquare credentials and creating the API request URL, we will send HTTP request and receive venues information in json file. Only venue names and categories will be extracted from the results.

City	Borough	Neighbourhood	Latitude	Longitude	Venue	Venue Category
Toronto	North York	Parkwoods	43.753259	-79.329656	Brookbanks Park	Park
Toronto	North York	Parkwoods	43.753259	-79.329656	Variety Store	Food & Drink Shop
Toronto	North York	Parkwoods	43.753259	-79.329656	Corrosion Service Company Limited	Construction & Landscaping
Toronto	North York	Victoria Village	43.725882	-79.315572	Victoria Village Arena	Hockey Arena
Toronto	North York	Victoria Village	43.725882	-79.315572	Tim Hortons	Coffee Shop

Table 3：Venue Names and Categories in Each Neighbourhoods in Toronto

After merging Toronto and Shanghai venues dataset into one single dataframe, we will change all types of restaurants such as 'French Restaurant', 'India Restaurant' and 'Sichuan Restaurant' into 'Restaurant' since almost all the restaurants in Shanghai are Chinese restaurants. We will then generate one-hot vectors for venue categories and group the dataset by neighbourhood and by taking the mean of the frequency of occurrence of each category. By looking at the grouped venues, we find that there are similar venues names; for example, ‘Bus Line' and 'Bus Stop' are basically in the same category ‘Bus Station’. We will clean up these columns for better analysis. Finally, we create a new dataframe and display the top 10 venues for each neighbourhood.

3.4 Machine Learning Models

K-Means clustering is a machine leaning algorithm that group unsupervised data based on the similarity. In this project, we will apply K-Means model to segment and cluster all the neighborhoods in Toronto and Shanghai based on the similarity of the venue types. Here we will use Elbow method to determine the right value of K. To do this, we will run the clustering with different values of K and calculate the sum of squared errors (SSE). The elbow point of the line chart is determined as the right K for clustering. Here K equals to 4.

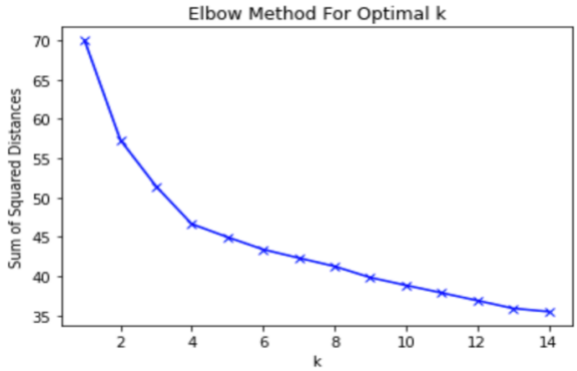


Figure 2: Elbow Method For Optimal K

4. Results

After all the process, we have got the summarization bellow:

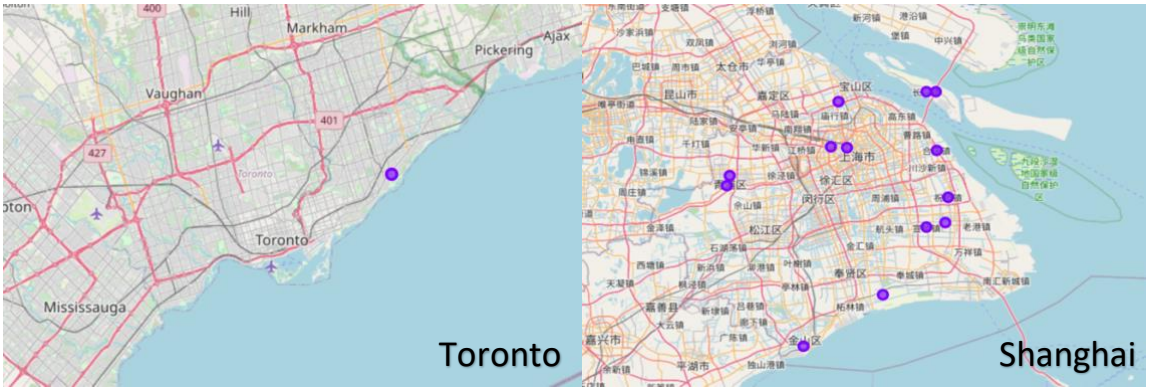
```
for cluster_num in range(4):
    num_of_nbh = toronto_sh_merged[toronto_sh_merged['Cluster_Labels'] == cluster_num].shape[0]
    print('The number of neighbourhoods in cluster {} is {}'.format(cluster_num+1, num_of_nbh))
```

```
The number of neighbourhoods in cluster 1 is 14
The number of neighbourhoods in cluster 2 is 85
The number of neighbourhoods in cluster 3 is 27
The number of neighbourhoods in cluster 4 is 130
```

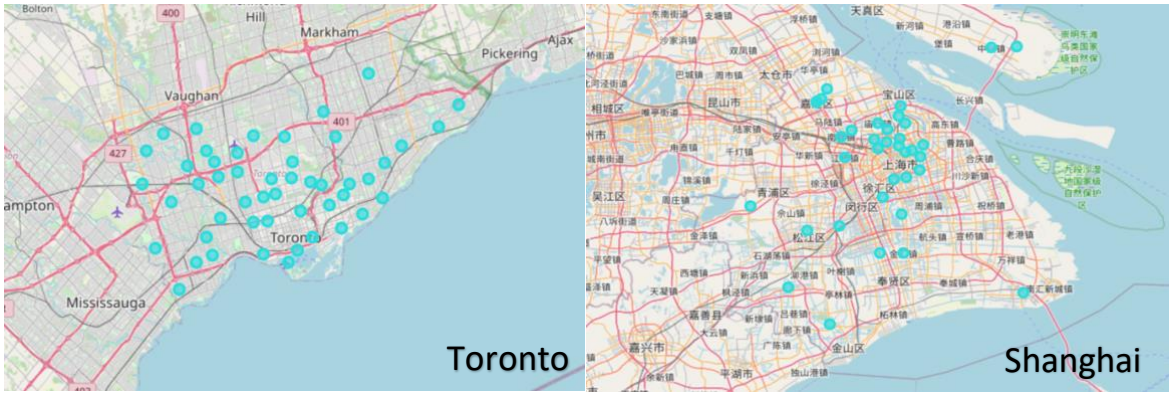
Figure 3: Number of Neighbourhoods in each Clusters

- Cluster 1 are mostly suburb areas where have access to hotels, airports and football stadium.
- Cluster 2 are basically residential areas with parks, grocery stores, pharmacy and restaurants.
- Cluster 3 includes neighbourhoods with restaurants and distribution centers.
- Cluster 4 are mostly downtown areas where surrounded by lots of restaurants, cafeteria, bars, convenience stores and different kinds of shops.

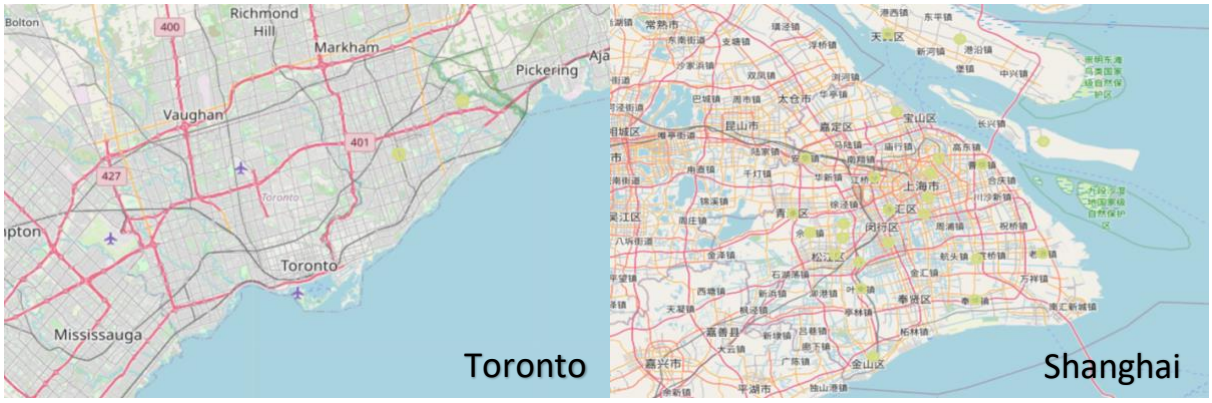
Folium maps of each cluster are created to facilitate the analysis. Similar neighbourhoods in Toronto and Shanghai are plotted with the same color.



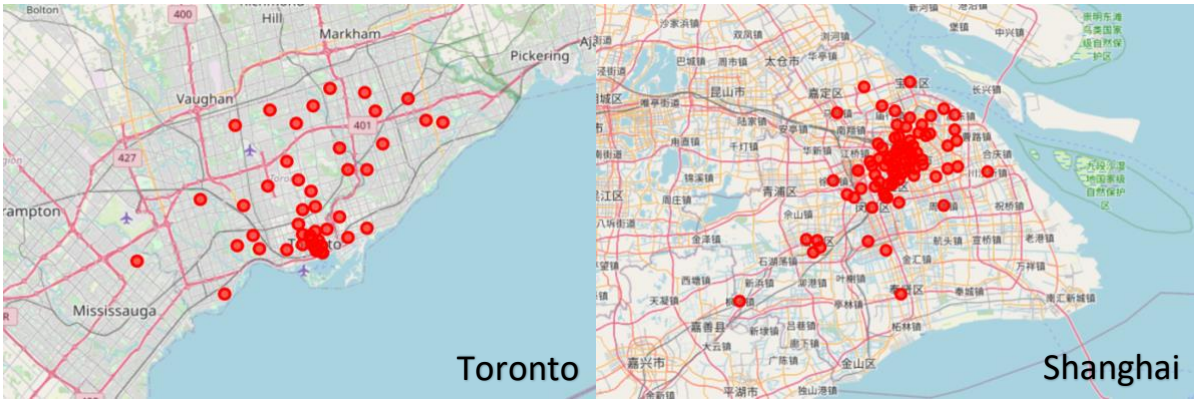
Cluster 1



Cluster 2



Cluster 3



Cluster 4

5. Discussion

From the results, we can conclude that for those who prefer to settle down in a residential area where surrounded by parks, grocery stores, pharmacy and restaurants, cluster 2 would be the best choice. While for those who prefer to live in a more crowded area where have access to a variety of venues, cluster 4 would be the best choice.

However, from the results, we notice that the majority of the neighbourhoods in Toronto lie into cluster 2 and 4. This is due to the limitations this research hold. To improve model performances and result in a better clustering, we will need further data such as more detailed venues information in Shanghai provided by China.

6. Conclusion

In this study, I built clustering model to group the similar neighbourhoods in two big cities, Toronto and Shanghai, based on their nearby venues. This recommendation system can be applied to any other cities rather than Toronto and Shanghai. People in the world are nowadays moving more frequently than before, I hope this analysis will help you to make a decision of choosing the neighbourhood in the destination city that fit for your needs. Furthermore, this research can be useful for real estate business if combined with Price Paid data, or city security solution if combined with crime data, etc.