

Neighborhood Battle Of Four Cities

Vijay Oommen

1. Introduction

Human migration is the movement by people from one place to another with the intentions of settling, permanently or temporarily in a new location. Early human migrants are usually result of climate change. Though the climate change is no longer severer in modern societies, there are still a number of causes impel migrants to move to another country. Globalization is one of the reasons people move to new places for better job opportunities

Migration is a huge event for most migrants. In addition to searching for new accommodation and new job, migrants also need to overcome culture shocks during settle down. In order to have a smooth transfer in the new city, it is very desirable to move to a similar neighborhood as where migrants live before. For example, a coffee lover who usually needs a cup of coffee every morning from Starbucks will prefer a place with a cafe nearby. I would like to use the Foursquare location data to analyze the neighborhoods of four different cities New York, Toronto, Beijing and Shanghai. These cities are the economic or political center of the countries. I wish to find similar neighborhoods among these cities and provide some useful information for people who are considering moving among Canada, USA and China.

2. Data

The neighborhood data of New York is provided here: https://geo.nyu.edu/catalog/nyu_2451_34572

The neighborhoods of Toronto are extracted from here: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The neighborhoods of Beijing are extracted from Wikipedia: https://en.wikipedia.org/wiki/List_of_township-level_divisions_of_Beijing

The neighborhoods of Shanghai are from: https://en.wikipedia.org/wiki/List_of_township-level_divisions_of_Shanghai

Please note all my data and analysis are available in my GitHub page:

https://github.com/vjoommen/Coursera_Capstone/blob/master/Capstone_Projekt.ipynb

3. Methodology

3.1 Data preparation

The New York data is available online with the neighborhood name and location information (latitude and longitude). For Toronto, Beijing and Shanghai, only neighborhood name is provided, I used the Nominatim library to obtain the location information by searching each neighborhood name. A number of neighborhoods are not found in Nominatim, I didn't include these neighborhoods in my analysis.

For those neighborhoods within the same borough, they were shown on map with the same label color. By doing so, it is easy to find that some neighborhoods jump into different color regimes, indicating wrong location information. These neighborhoods are further deleted.

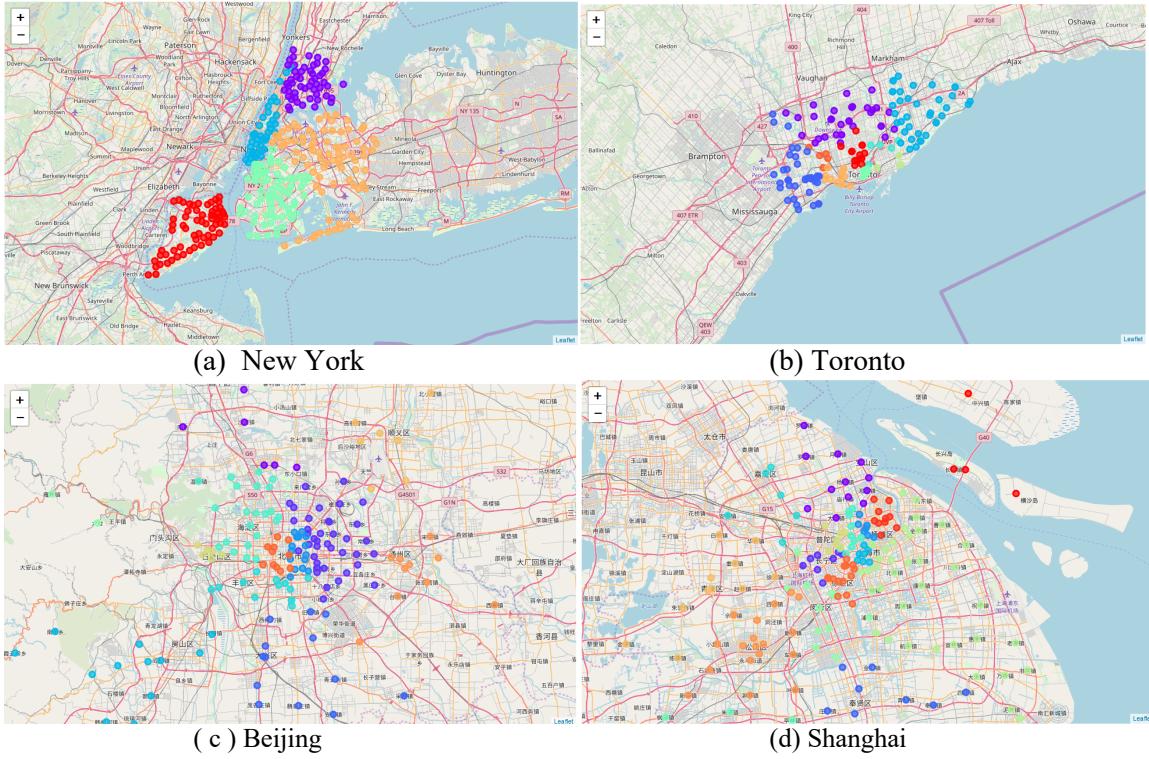


FIG. 1. Map of (a) New York, (b) Toronto, (c) Beijing and (d) Shanghai. The markers are the neighborhoods in my analysis. Different marker colors indicate different borough regimes.

3.2 Data acquisition

The Foursquare API was used to search for nearby venues of each neighborhood in radius of 1000 meters. Only the venues name and venues category (i.e. café, restaurant, school, etc.) are extracted.

After obtaining all the venues, the total number of venues in each category is counted for each neighborhood. To clarify, I also build a table with top 10 most frequent venue categories for each neighborhood.

3.3 Data Analysis

According to the total number of venues in each category, I applied hierarchical agglomerative clustering method to compare neighborhoods among cities. Unlike kmeans or other machine learning clustering methods, hierarchical agglomerative clustering doesn't require the number of clusters at the beginning. Moreover, it could also tell us whether the dataset is good for clustering at first glance.

To do so, firstly, I build a hierarchical clustering dendrogram (FIG. 2) with `scipy` library. According to FIG. 2, it is natural to separate the neighborhoods into 2 or 3 clusters. However, 2 or 3 clusters will just separate the neighborhoods into city center and suburb area, which doesn't satisfy the goal of my study. To gain deeper insight from the dataset I decide to separate the neighborhoods into 9 clusters by cutting at the distance of 40.

Secondly, I applied hierarchical agglomerative clustering of the `sklearn.cluster` library to cluster these neighborhoods.

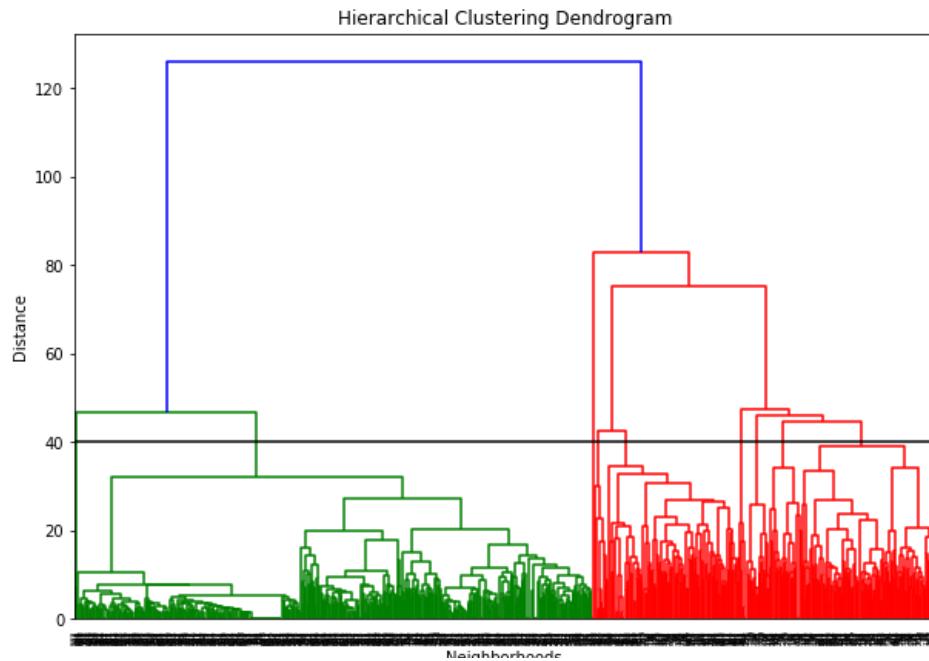


FIG. 2. Hierarchical clustering dendrogram of all the neighborhoods in New York, Toronto, Beijing and Shanghai. Horizontal black line is the cutoff at distance 40 to separate the neighborhoods into 9 clusters.

4. Results

4.1 Total number of neighborhoods in each cluster

Let's look at the total number of neighborhoods in each cluster to make sure they are properly clustered.

Cluster	0	1	2	3	4	5	6	7	8
Total number of neighborhoods	122	121	21	31	8	2	2	478	13

Table.1 Total number of neighborhoods in each cluster.

4.2 Outliers: Cluster 5 and cluster 6

We note that Cluster 5 and 7 has only 2. These two clusters probably just contain the outliers. Cluster 5 is the same neighborhood throughout two boroughs, which is also unique in New York. Cluster 6 is the historical area of Beijing; thus, no similar area was found in other cities..

	City	Borough	NeighborhoodName	Latitude	Longitude	Total Number of Venues	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	
115	New York	Manhattan	Murray Hill	40.7483	-73.9783	155	Korean Restaurant	Pizza Place	Bakery	Sandwich Place	Grocery Store	Deli / Bodega	Karaoke Bar	Gourmet Shop	Sup
180	New York	Queens	Murray Hill	40.7641	-73.8128	155	Korean Restaurant	Pizza Place	Bakery	Sandwich Place	Grocery Store	Deli / Bodega	Karaoke Bar	Gourmet Shop	Sup

Table.2 Neighborhoods in cluster 5.

	City	Borough	NeighborhoodName	Latitude	Longitude	Total Number of Venues	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	
532	Beijing	Dongcheng	Jingshan	39.9221	116.396	39	Historic Site	French Restaurant	Scenic Lookout	Park	Vietnamese Restaurant	Lounge	Noodle House	Garden	
533	Beijing	Dongcheng	Donghuamen	39.9138	116.395	58	Historic Site	Park	Pizza Place	Clothing Store	Scenic Lookout	Dessert Shop	Lounge	Concert Hall	

Table.3 Neighborhoods in cluster 6.

4.3 Cluster 0

Cluster 0 contains the neighborhoods in New York and Toronto. The New York region consists a wide range of local business like restaurants and pizza shops. As for Toronto, the neighborhood primarily constitutes shopping centers.

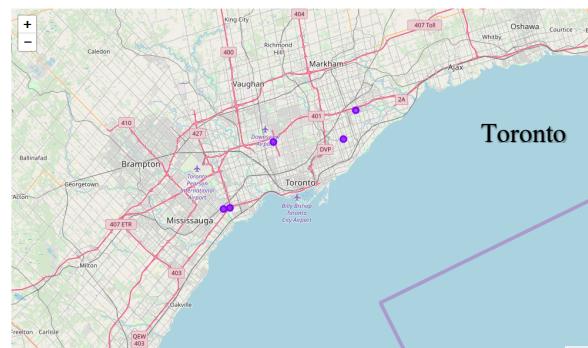
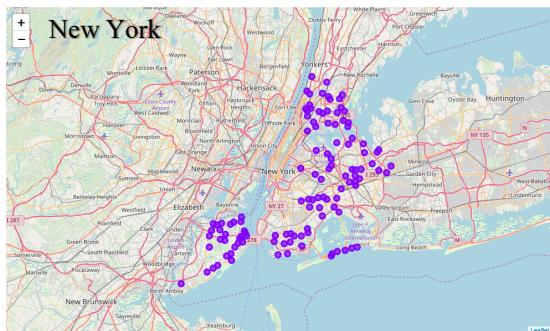


FIG. 3 Neighborhoods in Cluster 0.

4.4 Cluster 1

Cluster 1 is basically the city center where surrounded by a variety of venues, including theaters, parks, restaurants and bars.

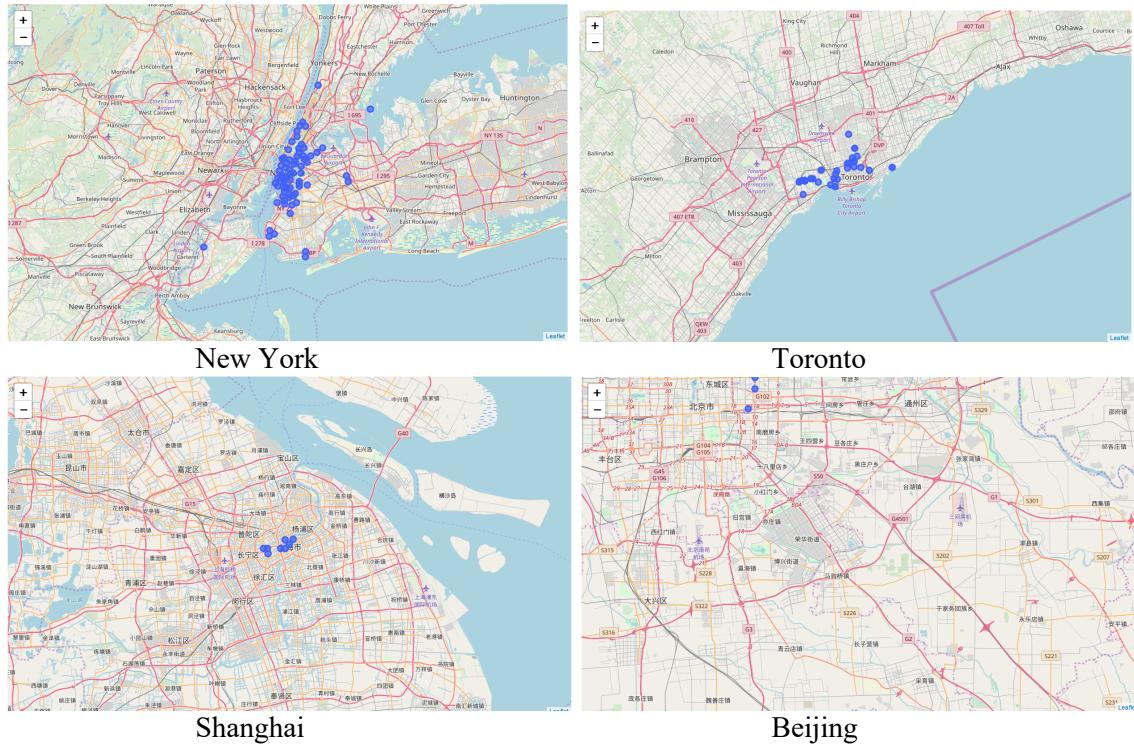


FIG. 4 Neighborhoods in Cluster 1.

4.5 Cluster 2

Cluster 2: Neighborhoods in New York with a lot of nearby pizza places, fast food restaurants and Caribbean restaurants.

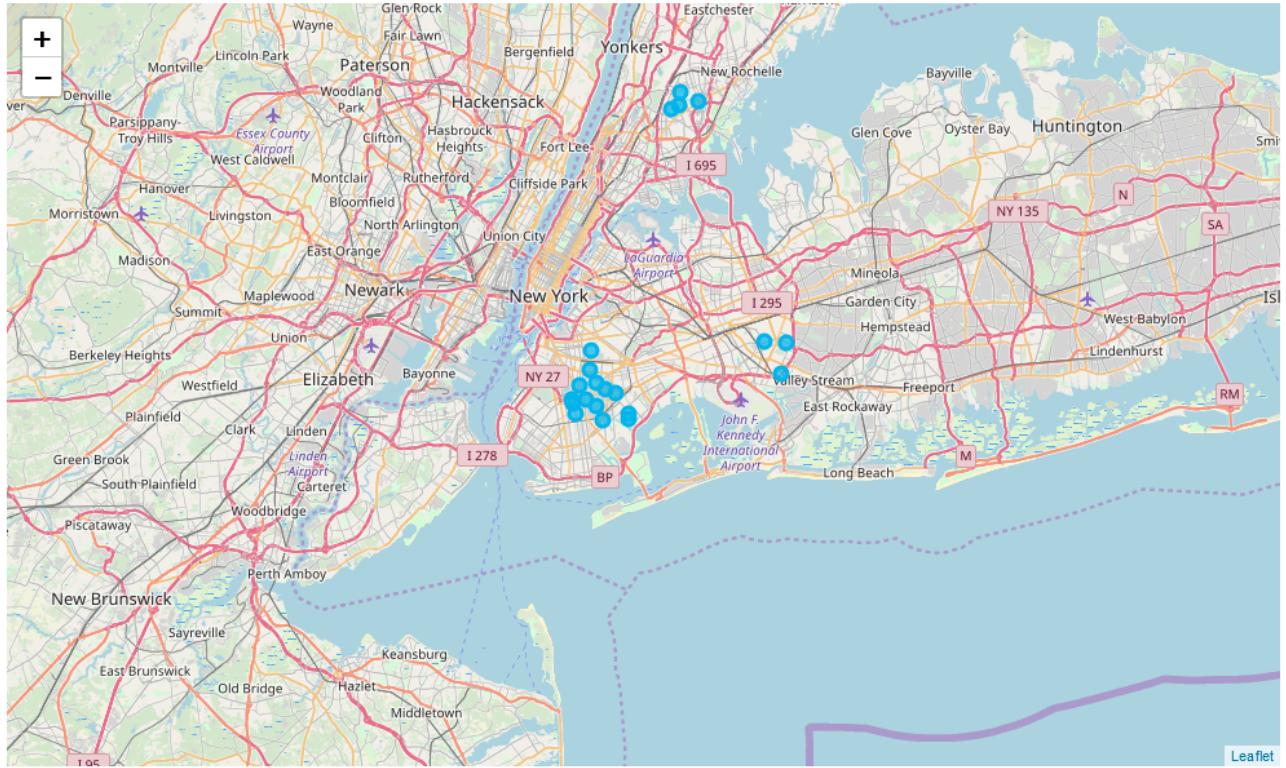


FIG. 5 Neighborhoods in Cluster 2.

4.6 Cluster 3

Cluster 3: Most are New York and Toronto neighborhoods with a lot of Italian and pizza places



FIG. 6 Neighborhoods in Cluster 3.

4.7 Cluster 4

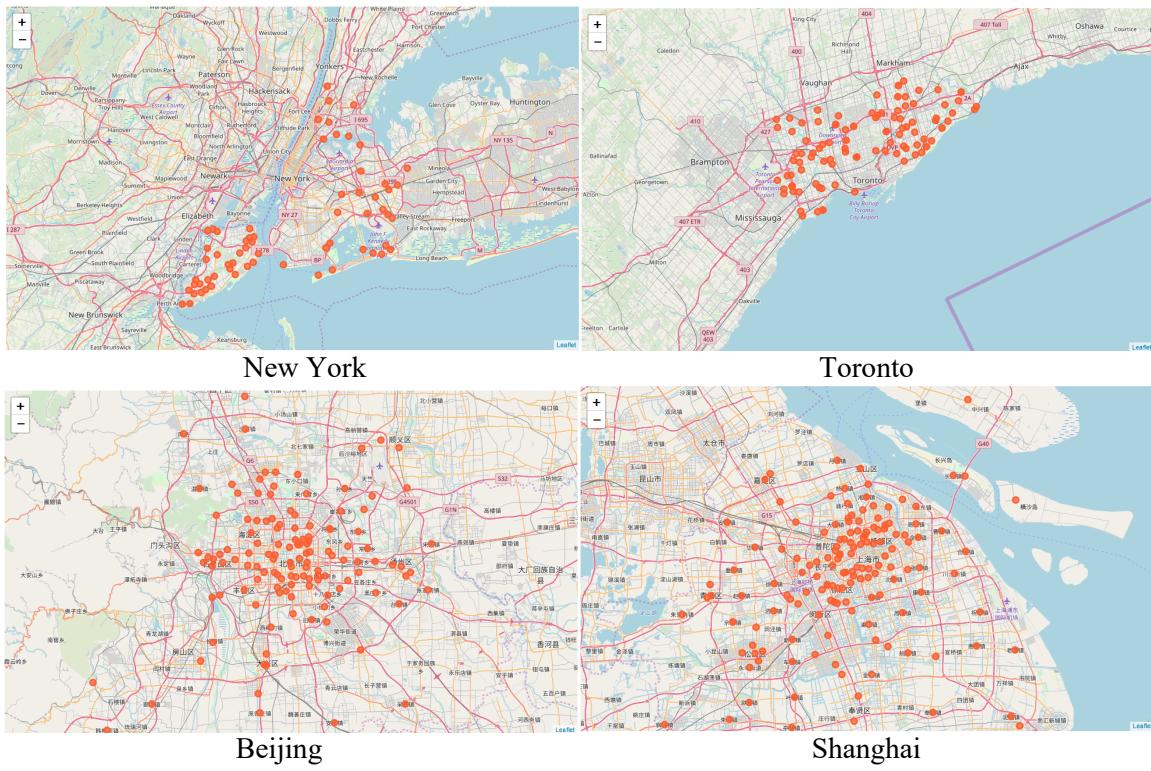
Cluster 4 includes the neighborhoods in New York, Toronto and Beijing. These neighborhoods have a lot of dining place with Korean cuisine.



FIG. 7 Neighborhoods in Cluster 4.

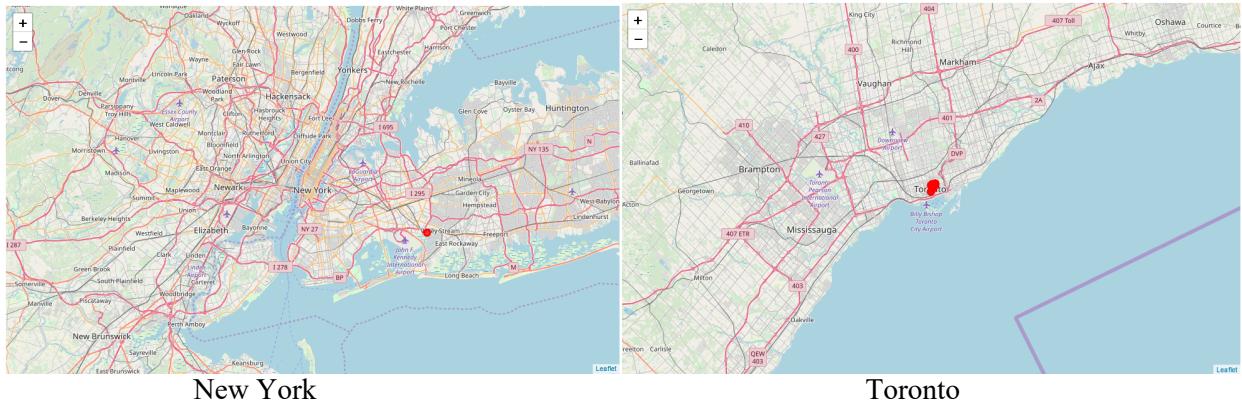
4.8 Cluster 7

Cluster 7 is basically the city center where surrounded by a variety of venues, including theaters, parks, restaurants and bars.



4.9 Cluster 8

Cluster 8 contains primarily Downtown Toronto neighborhoods with access to various venues like restaurants and shopping centers



5. Discussion

Overall, the results are not surprising. Toronto and New York are much more similar than Shanghai and Beijing, because they are much closer in location and have similar cultural background. Despite that, we were able to find some similar neighborhoods among these four cities.

For those who have a job in the city center and don't want to spend much time on transportation, neighborhoods in cluster 1 will be good.

Cluster 7 neighborhoods looks to be the ideal locations for residential living condition since it provides good balance to key venue categories like park, department stores, restaurants and access to public transportation. These neighborhoods are not crowded as city centers, but has a number of outdoor parks, restaurants and supermarkets.

Please note that in cluster 2,3,4 and 8, it doesn't mean Beijing and Shanghai are less bustling. In fact, Beijing and Shanghai have even larger density of population. I think this is because the venue information provider Foursquare is located in USA, thus New York and Toronto have much more information than Beijing and Shanghai. Some venue providers located in China may help to have a more detailed analysis.

6. Conclusion

I hope I convince you that despite the distinct cultural and geographical location differences of New York, Toronto, Beijing and Shanghai, there are several similar neighborhoods in cluster 0, 1, 2, 3, 4, 7and 8. I hope this could provide some useful information for people who are considering moving among Canada, USA and China.

Thanks for reading!