The Battle of Neighborhoods

Jason Zhao

July 6, 2021

1. Introduction

1.1. Background

According to UN Department of Economic and Social Affairs (DESA), there are approximately 272 million international migrants worldwide in 2020. Although under the COVID-19 pandemic, this number stays the same as in 2019, about 3.5% of world's total population are migrants.

Toronto and New York, two well-known big cities in Canada and US, are the economic centers and financial capitals of their countries. In 2020, Toronto welcomed 130,000 immigrants while New York welcomed 400,000 immigrants. A comparison of the two cities regarding their living quality can be made to help potential international immigrants to decide on which city they should choose.

1.2. Introduction

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 people to find a similar neighborhood in the new city as the one they lived in before. In this project, machine learning tools are used to cluster Toronto and Shanghai neighborhoods in order to recommend the neighborhoods which are the best options for migrants based on surrounded essential facilities such as school, hospital, restaurants, and stores etc.

This report is mainly for but not limited to people who are planning to move from one city to another. This recommendation system can also be beneficial to stakeholders who are interested in citing a new business in a new city.

2. Data

The following datasets are used for this project:

- Toronto city data: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv
- New York city data: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

- Foursquare search features are used to collect neighborhood venues information, which are used to explore and compare geographical locations of Toronto and New York.

3. Methodology

3.1. Obtain Coordinates

Two DataFrames named Toronto_neighborhoods_df and newyork_df are created using a csv file mentioned in the Data section. Geospatial coordinates (latitude and longitude) are included in both DataFrames for Toronto and New York neighborhoods.

A new DataFrame named df_toronto_newyork is a merged dataset for both cities. Table 1 shows the configuration of this dataset.

```
There are total 409 neighbourhoods in Toronto and New York.
```

	City	Borough	Neighborhood	Latitude	Longitude
0	Toronto	North York	Parkwoods	43.753259	-79.329656
1	Toronto	North York	Victoria Village	43.725882	-79.315572
2	Toronto	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	Toronto	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	Toronto	Queen's Park	Ontario Provincial Government	43.662301	-79.389494

Table 1: Merged dataset for Toronto and New York

3.2. Data Visualization

Now we have obtained information of boroughs, neighborhoods and their respective geospatial coordinates for Toronto and New York. Folium library is then installed and used to plot a map of all the neighborhoods in each city. The neighborhoods belong to the same borough are plotted with the same color, shown in Figure 1.

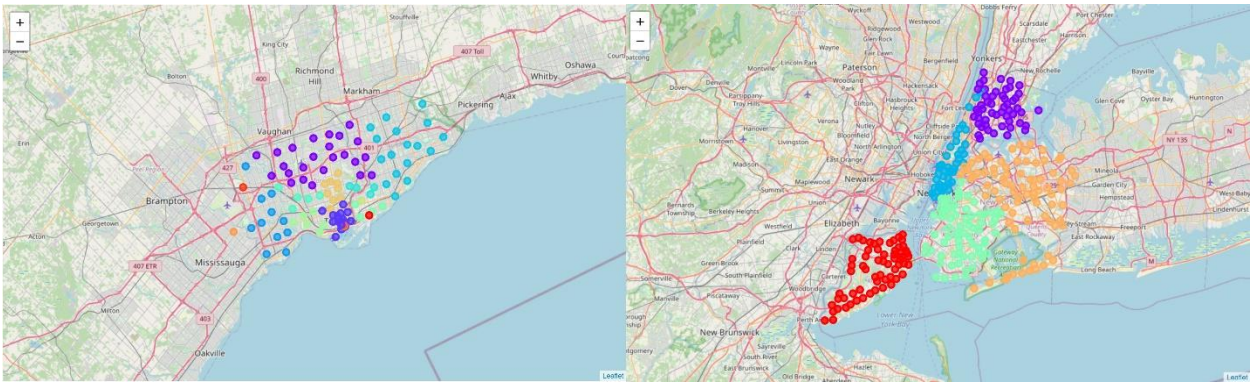

Figure 1: Maps of Toronto (left) and New York (right) neighborhoods

3.3. Foursquare and API Search Feature

Foursquare API provides access to massive dataset of location data and venues information including address, images, tips, ratings and comments. In this project, Foursquare API and Geopy data are used to locate nearby venues within 500 meters of each neighborhood in Toronto and New York. After defining Foursquare user 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. Figure 2 shows an example with Toronto's venue information.

	City	Borough	Neighborhood	Latitude	Longitude	Venue	Venue Category
0	Toronto	North York	Parkwoods	43.753259	-79.329656	KFC	Fast Food Restaurant
1	Toronto	North York	Parkwoods	43.753259	-79.329656	Brookbanks Park	Park
2	Toronto	North York	Parkwoods	43.753259	-79.329656	Variety Store	Food & Drink Shop
3	Toronto	North York	Victoria Village	43.725882	-79.315572	Victoria Village Arena	Hockey Arena
4	Toronto	North York	Victoria Village	43.725882	-79.315572	Portugril	Portuguese Restaurant

Figure 3: Toronto venues

After merging Toronto and New York venues dataset into one single dataframe, we can find a lot of different types of restaurants. In order to make our result simpler, we define all types of restaurants into "Restaurant" only.

3.4. Machine Learning Models

K-means clustering is a machine learning algorithm that group unsupervised data based on the similarity. In this project, K-means model is applied to segment and cluster all neighborhoods in Toronto and New York based on the similarity of the venue type. Elbow method is used here to determine the correct value of K. In order to do this, different values of K are tested and selected based on the sum of squared errors (SSE). The elbow point of the line chart is determined as the optimal K for clustering, where K equals to 6. Below shows the K-means clustering.

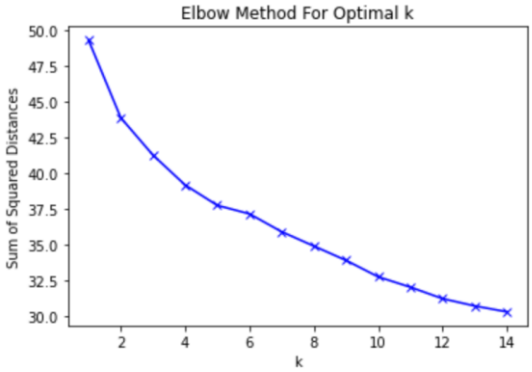

Figure 4: K-means clustering.

4. Results

Finally, we got the summarization below:

```
The number of neighborhoods in cluster 1 is 212
The number of neighborhoods in cluster 2 is 2
The number of neighborhoods in cluster 3 is 7
The number of neighborhoods in cluster 4 is 27
The number of neighborhoods in cluster 5 is 19
The number of neighborhoods in cluster 6 is 137
```

Where:

- Cluster 1 contains mainly restaurants and coffee shops.
- Cluster 2 contains mainly playgrounds and yoga studios.
- Cluster 3 contains mainly parks and convenience stores.
- Cluster 4 contains mainly restaurants and other snake stores.
- Cluster 5 contains mainly grocery stores, construction sites and baseball playfields.
- Cluster 6 contains many different living essential stores and entertainment venues.

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

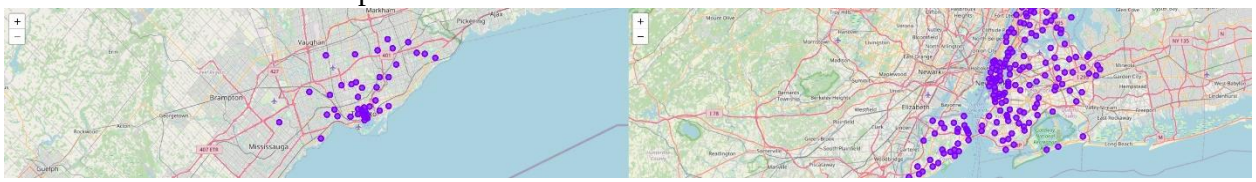

Figure 5: Cluster 1 (Left: Toronto; Right: New York)

Figure 6: Cluster 2 (Left: Toronto; Right: New York)

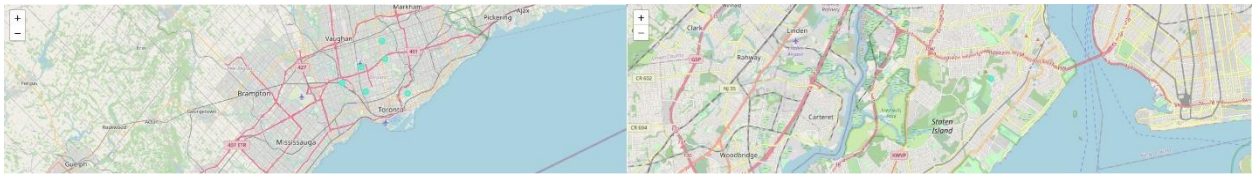

Figure 7: Cluster 3 (Left: Toronto; Right: New York)

Figure 8: Cluster 4 (Left: Toronto; Right: New York)

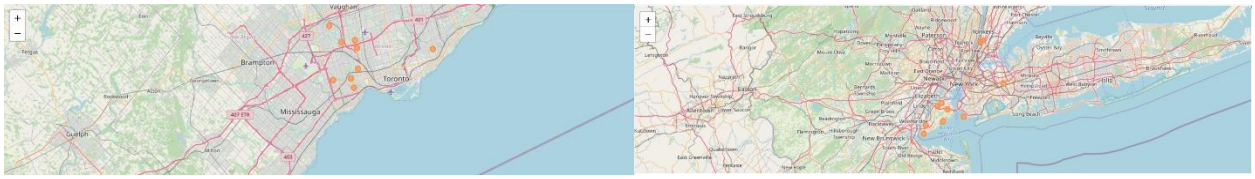

Figure 9: Cluster 5 (Left: Toronto; Right: New York)

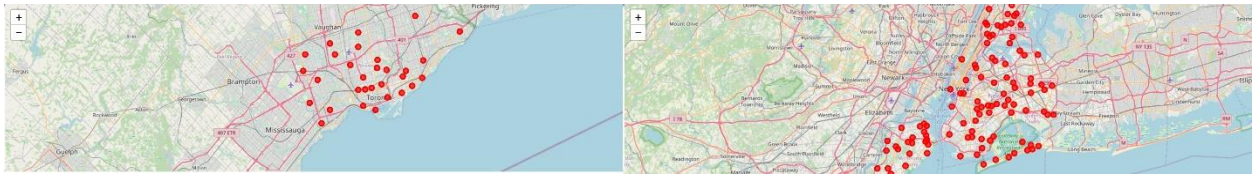

Figure 10: Cluster 6 (Left: Toronto; Right: New York)

5. Discussion

From the results, we can conclude that for those who prefer to settle down in a residential area where surrounded by restaurants, cluster 1 and 4 would be great. For those who want to live in a neighborhood with pleasant views, cluster 2 and 3 would be great choice. For people who want to live in a neighborhood with hospitals, restaurants, banks and many other convenience stores, cluster 6 would be a good choice.

Overall, Toronto is less crowded than New York. For people who want to live a leisurely life, it's better for them to move to Toronto. On the other hand, New York is more prosperous than Toronto for having a lot more varieties of stores in evert neighborhood. Therefore, New York is a better place for people who want to live a high-quality life.

6. Conclusion

I built clustering model in this project to group similar neighborhoods in two big cities – Toronto and New York, based on their nearby venues. This recommendation system can be applied to any other cities besides Toronto and New York. With the globalization, more and more people are becoming international immigrants. This analysis can hopefully help them make a decision on choosing the optimal neighborhood in the destination city that fits their needs.