IBM Applied Data Science Capstone Project

The Battle of the Neighbourhoods: Toronto and New York

By: Ngai Cai Hua Kevin

Contents

Introduction/Business Problem 3
Data 3
Methodology 4
Results / Observations 5
Recommendations 11
Conclusion 11
References 11

Introduction/Business Problem

Toronto and New York are big vibrant metropolitan cities with large population of people going about their day-to-day lives. In addition, there are many tourists that visit the cities; foreigners who also work and study in the cities. Hence, these cities consist of people with diverse background with people from all over the world.

Nonetheless, there are both similarities and differences between both the cities. The **business problem** is to identify and determine these similarities and differences based on their venues/nearby attractions in their neighbourhoods. The objective is to gain valuable insights on Toronto and New York by using machine learning technique such as clustering to group similar types of venues together. From here, we can pinpoint some of the similarities and the distinct features between the cities.

Target audience: Toronto and New York Citizens. Potential tourists that would love to visit these two locations. Potential immigrants or people who would like to study, work or setup businesses in the respective cities.

It can also serve as a factor for foreigners to choose between Toronto and New York for various purposes such as education, work, business, holidays.

Data

There are two different datasets used for retrieving the neighbourhoods and location data; one for Toronto and another for New York. From here, we use the FourSquare API to gather the venues data for the respective neighbourhoods in these cities.

Toronto

The dataset used for Toronto consists of a list of postal code, followed by the Borough and Neighbourhood and their respective location in terms of longitude and latitude. Data for Canada location details were scraped from the Wikipedia page and input into a pandas dataframe. However, longitude and latitude data were taken from the excel file from the week 3 assignment. Data cleaning (filtering) was also performed on the pandas dataframe with the purpose of removing the neighbourhoods that do not belong to Toronto.

Next, we are able to obtain the surrounding venues (and their details) of the neighbourhoods in Toronto via FourSquare API, by providing the following inputs to FourSquare:

- List of Neighbourhoods in Toronto
- List of Longitude
- List of Latitude

The results returned contains the various venues near the list of neighbourhoods within a radius of 500 metres and a limit of up to 100 venues per neighbourhood.

New York

The dataset used for New York is retrieved from a JSON file provided in week 3 lab. We proceed to extract the neighbourhoods and location data in New York from the JSON file and input the data into a pandas dataframe.

Next, we are able to obtain the surrounding venues (and their details) of the neighbourhoods in New York via FourSquare API, by providing the following inputs to FourSquare:

- List of Neighbourhoods in New York
- List of Longitude
- List of Latitude

The results returned contains the various venues near the list of neighbourhoods within a radius of 500 metres and a limit of up to 100 venues per neighbourhood.

With both the neighbourhood and nearby venues data for Toronto and New York on hand, we can perform "clustering" separately for both cities.

Last but not least, we can perform a comparison between Toronto and New York based on the clustering results and analyse the similarities and differences between both cities.

Methodology

The methodology used for the project is "Clustering".

Clustering is an unsupervised machine learning technique. In this project, we are using K-means clustering algorithm, which enables unlabelled data to be grouped according to their similarity.

K-means partitions the data into k-number of clusters with each cluster being unlabelled. The number of clusters is represented by "k" and is determined by the data scientist/analyst. For this project, we will be diving the data into 5 different clusters for both Toronto and New York respectively. Hence, k=5.

Prior to perform clustering, we need to perform data cleaning and transformation. Data on Toronto has been filtered out from the Canada dataset and input into a pandas dataframe. For New York case, the data has already been provided in JSON file. Any unassigned or NaN rows are dropped

Next, we have filtered out columns such as "Neighbourhood", "Latitude" and "Longitude" from the Toronto and New York dataframes respectively. From here onwards, we have established a connection to FourSquare API to retrieve the nearby venues in the neighbourhoods for both cities.

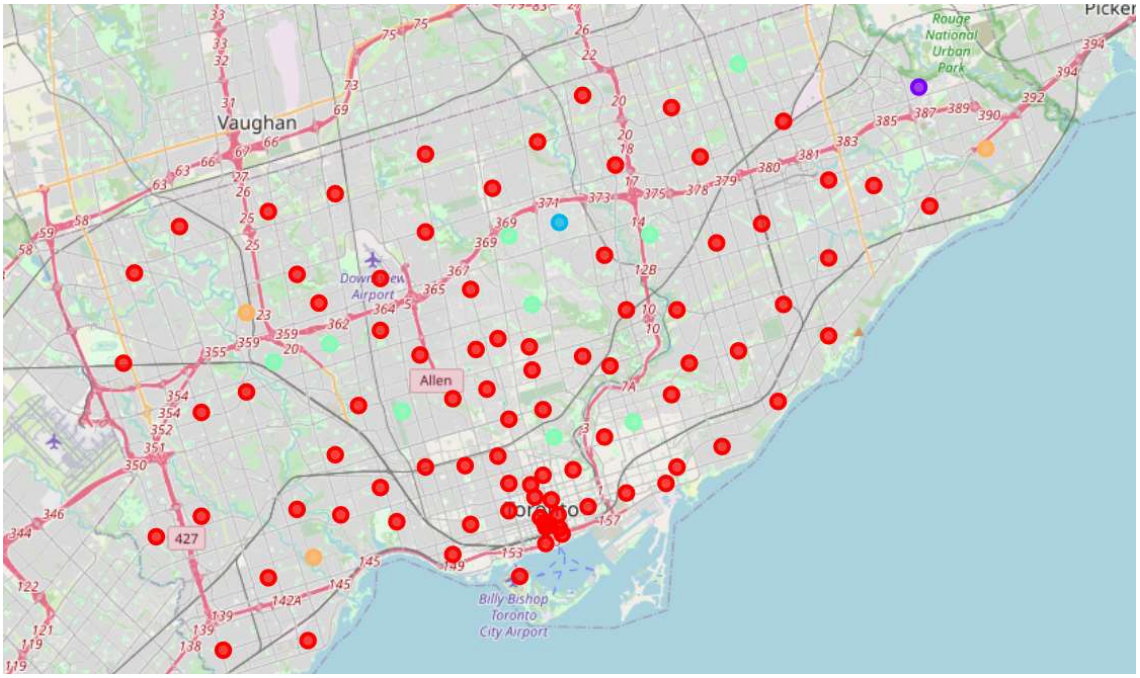
From the data retrieved via the FourSquare API, we can find the frequency and type of venues in the neighbourhoods. Henceforth, we performed data transformation using the "One hot encoding" technique to convert the respective types of venues into columns with

numerical values of “0” or “1”. “1” represents that it is in the neighbourhood, while “0” states otherwise.

Upon performing the necessary data cleaning and transformation, we can proceed to perform “Clustering”, and thus analyse the results thereafter.

Results / Observations

Map of Toronto Clusters



There are 5 clusters partitioned for the city of Toronto, from cluster 0 to 4.

	Borough	Cluster	1st Most Common Venue	2nd Most Common Venue
1	North York	0.0	Pizza Place	Hockey Arena
2	Downtown Toronto	0.0	Coffee Shop	Bakery
3	North York	0.0	Clothing Store	Accessories Store
4	Downtown Toronto	0.0	Coffee Shop	Yoga Studio
7	North York	0.0	Gym	Japanese Restaurant
8	East York	0.0	Pizza Place	Gastropub

Figure 1 Toronto Cluster 0

For Cluster 0, it is represented by red circles/dots on the map. Cluster 0 consists of the most number of locations with a substantial amount of food and drink outlets. Places such as Pizza and coffee shops are quite popular as evident in Figure 1.

For Cluster 1, it is represented by purple circles/dot on the map. Cluster 1 consists of a location in Scarborough borough with the top 2 common venues being restaurants as evident in Figure 2.

	Borough	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
6	Scarborough	1.0	Fast Food Restaurant	Dumpling Restaurant	Discount Store

Figure 2 Toronto Cluster 1

For Cluster 2, it is represented by blue circles/dot on the map. Cluster 2 consists of a location in North York borough with the top 2 common venues being Martial Arts School and Yoga Studio as evident in Figure 3.

	Borough	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
45	North York	2.0	Martial Arts School	Yoga Studio	Eastern European Restaurant

Figure 3 Toronto Cluster 2

For Cluster 3, it is represented by green circles/dot on the map. Cluster 3 is grouped, with the top 2 common venues being Park and Playground as evident in Figure 4.

	Borough	Cluster	1st Most Common Venue	2nd Most Common Venue
0	North York	3.0	Park	Food & Drink Shop
21	York	3.0	Park	Women's Store
35	East York	3.0	Intersection	Park
49	North York	3.0	Bakery	Park
61	Central Toronto	3.0	Park	Swim School
64	York	3.0	Park	Yoga Studio
66	North York	3.0	Park	Convenience Store
84	Scarborough	3.0	Playground	Park
90	Downtown Toronto	3.0	Park	Playground

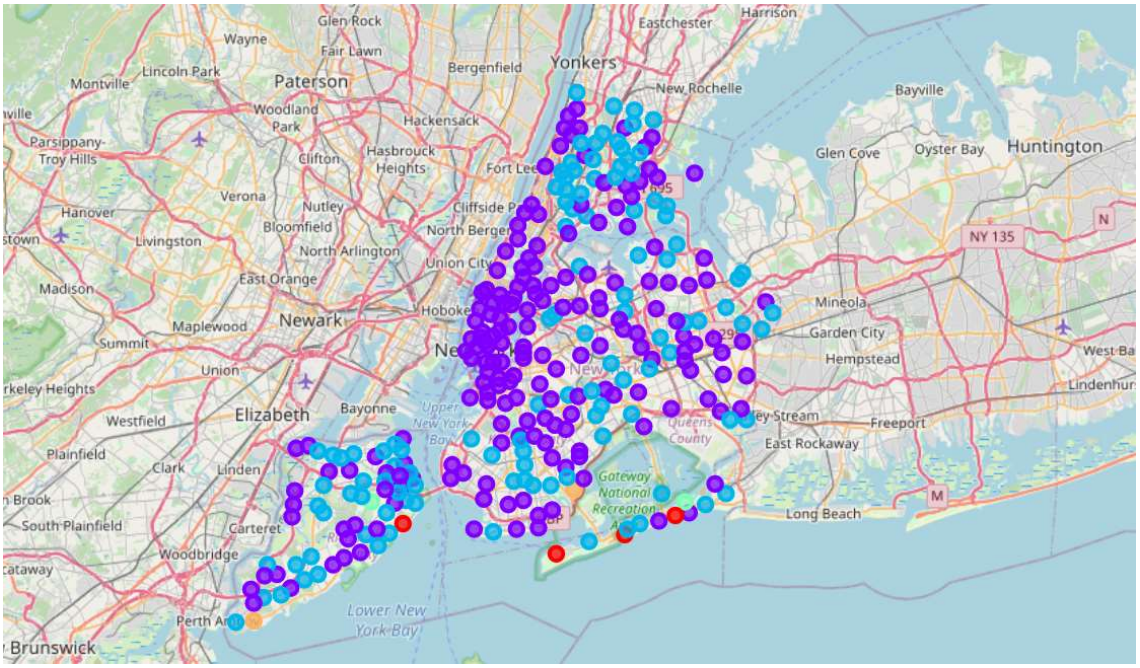
Figure 4 Toronto Cluster 3

For Cluster 4, it is represented by orange circles/dot on the map. Cluster 4 consists of 3 locations in Scarborough, North York and Etobicoke boroughs respectively. The top 2 common venues being visited are Construction & Landscaping and Baseball Field as evident in Figure 5.

	Borough	Cluster	1st Most Common Venue	2nd Most Common Venue
12	Scarborough	4.0	Construction & Landscaping	Bar
57	North York	4.0	Baseball Field	Yoga Studio
100	Etobicoke	4.0	Construction & Landscaping	Baseball Field

Figure 5 Toronto Cluster 4

Map of New York Clusters



There are 5 clusters partitioned for the city of New York, from cluster 0 to 4.

For Cluster 0, it is represented by red circles/dots on the map. Cluster 0 is grouped, with the most common venue being Beach as evident in Figure 6.

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
172	Breezy Point	0	Beach	Monument / Landmark	Trail
179	Neponsit	0	Beach	Yoga Studio	Food
204	South Beach	0	Pier	Deli / Bodega	Beach
302	Hammels	0	Beach	Deli / Bodega	Bus Stop

Figure 6 New York Cluster 0

For Cluster 1, it is represented by purple circles/dots on the map. Cluster 1 is grouped with a variety of different types of venues as evident in Figure 7. They include places such as Bus Station, Food Truck, Gym, Pizza Place, Nightclub, etc.

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Co-op City	1	Bus Station	Accessories Store	Pizza Place	Grocery Store	Fast Food Restaurant
3	Fieldston	1	Medical Supply Store	River	Plaza	Fish Market	Exhibit
4	Riverdale	1	Park	Food Truck	Gym	Home Service	Moving Target
5	Kingsbridge	1	Pizza Place	Bar	Sandwich Place	Latin American Restaurant	Mexican Restaurant
6	Marble Hill	1	Discount Store	Gym	Coffee Shop	Sandwich Place	Yoga Studio
9	Williamsbridge	1	Bar	Soup Place	Caribbean	Nightclub	Fruit & Vegetable

Figure 7 New York Cluster 1

For Cluster 2, it is represented by blue circles/dots on the map. Cluster 2 is also grouped with a variety of different types of venues as evident in Figure 8. They include places such as Pharmacy, Bus Station, Pizza Place, Ice Cream Shop, etc. Both Clusters 1 and 2 consist of the majority of the neighbourhoods with multiple venues in New York.

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Wakefield	2	Pharmacy	Donut Shop	Pizza Place	Ice Cream Shop	Laundromat
2	Eastchester	2	Bus Station	Caribbean Restaurant	Deli / Bodega	Diner	Cosmetics Shop
7	Woodlawn	2	Deli / Bodega	Pub	Pizza Place	Playground	Bar
8	Norwood	2	Pizza Place	Chinese Restaurant	Bank	Park	Deli / Bodega
11	Pelham Parkway	2	Bus Station	Italian Restaurant	Pizza Place	Frozen Yogurt Shop	Home Service
13	Bedford Park	2	Diner	Mexican Restaurant	Chinese Restaurant	Pizza Place	Deli / Bodega

Figure 8 New York Cluster 2

For Cluster 3, it is represented by green circles/dots on the map. Cluster 3 consists of 2 neighbourhoods (Somerville and Todt Hill) with the top 3 common venues being Park, Yoga Studio and Exhibit as evident in Figure 9.

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
192	Somerville	3	Park	Yoga Studio	Exhibit
203	Todt Hill	3	Park	Yoga Studio	Exhibit

Figure 9 New York Cluster 3

For Cluster 4, it is represented by orange circles/dots on the map. Cluster 4 consists of 2 neighbourhoods (Mill Island and Butler Manor) with the most common venue being Pool as evident in Figure 10.

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
76	Mill Island	4	Pool	Yoga Studio	Eye Doctor
238	Butler Manor	4	Pool	Baseball Field	Food

Figure 10 New York Cluster 4

Similarities between Toronto and New York

Both Toronto and New York have many food and drink restaurants/outlets. Pizza Place is a popular venue for both cities. There is a wide variety of food available and bars for people to relax after a long day of work or during weekends.

Leisure and recreational venues such as Park, Pool, Beach, etc are among the common venues. People can take leisure walks in the parks and beaches. Fitness venues are also popular in both cities as Yoga Studios and Gyms are frequently visited. People can choose to run, swim, workout and play sports in these venues.

Differences between Toronto and New York

Based on the results returned, there is no huge disparity between both cities based on the types of venues in both cities.

Nonetheless, if we perform a comparison of both the maps, we can deduce that there are more venues in New York than Toronto. New York is a larger city than Toronto, with a land area of 783.8 km² as compared to 630.2 km². Moreover, the population in New York is also substantially larger than Toronto.

The location points (venues) in Toronto are more spread out in comparison to New York's location points. This comparison indicates that there is significantly less clustering in Toronto, which may imply than Toronto is not as busy as New York; lesser activities for the people in Toronto to participate as compared to New York.

Recommendations

For tourists, both cities provide many good locations to visit. Tourists can go sight-seeing, visit different restaurants, etc. There is a wide variety of restaurants available in both cities; hence, they can cater to people all around the world.

For potential business owners, they can setup their business in either cities; although it also depends on the type of business that they will be doing. With that being said, New York may offer more business opportunities, mainly due to its large population and strong economy. On the other hand, we can also argue that setting up and maintaining a business in New York is very difficult as they are many competitors in the city.

Conclusion

In conclusion, we have performed a comparison of Toronto and New York to determine the similarities and differences of both cities in terms of popular and commonly frequented venues. The methodology used is via machine learning, to cluster the venues based on their similarities. With the results produced, we can perform data analysis to determine the similarities and differences. Moreover, we can drill down (perform next-level analysis) to understand the reason for the similarities and differences. However, the next-level analysis is not covered in this project. Henceforth, we can also make insightful recommendations to the target audience based on our analysis.

References

List of postal codes of Canada: M. In Wikipedia. Retrieved October 26, 2020, from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M