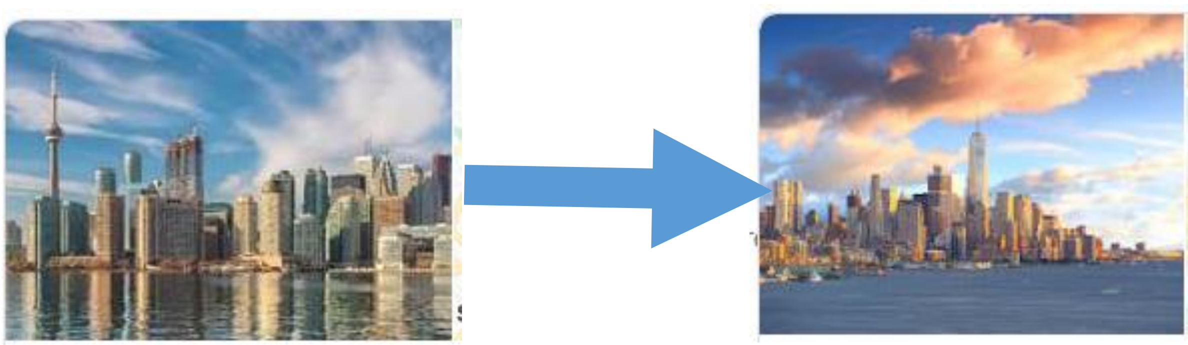
Coursera Capstone Project – The Battle of Neighborhoods

IBM Data Science Professional Certificate

Migrating from Toronto to New York

By: Sameer Mahajan
March 2020



Introduction

When you decide to live in a neighborhood, it is mainly because of quality of life. It depends on your personal choices. These choices are for amenities or venues that are in your neighborhood. These venues could be like gourmet restaurants, pharmacies, parks, schools and so on.

Business Problem

A friend of mine has been settled in Toronto, Canada for the last 25 years. She lives in Caledonia-Fairbanks neighborhood in York Borough (with postal code of M6E to be more specific). She loves her neighborhood based on above mentioned factors. Recently she received a very lucrative job offer from a great company in city of New York with great career prospect. Because of location change she will have to move if she decides to accept the offer. Wouldn't it be great if she is able to decide the neighborhoods in New York that are exactly like her current neighborhood in Toronto? She can then pick a neighborhood that is exactly like her current neighborhood and also very close to her new workplace to move to. This exercise will also help her to make a decision whether to move based on what city of New York has to offer to her liking.

Target Audience of this Project

This project is particularly useful for anyone moving from Toronto to New York City. It can also be used by recruiting agencies to motivate, guide and / or assist employees moving across these cities. The same technique can be applied for move between any two cities by leveraging similar data from those cities.

Data

To go about solving this problem, we need

- List of neighborhoods in these two cities
- Geographical coordinates of latitude and longitude information of these neighborhoods and
- Information about venues (categories like restaurant, park, hospital etc.) around these neighborhoods

For list of neighborhoods we will either get readymade data like https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json for New York City or extract it from Wikipedia pages like https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M for Toronto. Here is some sample data for Toronto:

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Rouge,Malvern,
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union,
2	M1E	Scarborough	Guildwood,Morningside,West Hill,
3	M1G	Scarborough	Woburn,
4	M1H	Scarborough	Cedarbrae,

We will get the geographical coordinates of these neighborhoods using python's geocoder package. Here is some sample data for Toronto:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353

	Postal Code	Latitude	Longitude
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

We will use Foursquare API (https://developer.foursquare.com/) to get information about venues around these neighborhoods. It has data of over 105 million places. It provides many categories of the venue data. Here is some sample data for Toronto:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge,Malvern,	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge,Malvern,	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
2	Highland Creek,Rouge Hill,Port Union,	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
3	Guildwood,Morningside ,West Hill,	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood,Morningside ,West Hill,	43.763573	-79.188711	Marina Spa	43.766000	-79.191000	Spa

We will leverage category of venues to identify similar neighborhoods.

You can refer to notebook https://github.com/sameermahajan/Coursera_Capstone/blob/master/Neighborhoods%20in%20Toronto.ipynb for details on how it can be done for Toronto.

Methodology

This project applies various data science techniques like

- Web scraping to gather data from Wikipedia page

- Working with APIs (Foursquare)
- Data cleaning
- Data wrangling
- Machine learning (k means clustering)
- Map visualization (folium)

to arrive at the solution.

First we get the list of neighborhoods in Toronto from Wikipedia page of https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M . We perform web scraping from this page to get the data. We get the geographical coordinates of these neighborhoods using geocoder python package. After gathering this data we plot a map of these neighborhoods using folium package. This map performs a simple validation of correctness of our data.

We then use Foursquare APIs to get top 100 venues within 500 meter radius of these neighborhoods. We check how many venues are returned for each neighborhood as a validation of this data. We take the mean of the frequency of occurrence of each venue by category to provide it as a feature to k means Clustering.

We perform k means clustering to categorize the neighborhoods into clusters of different characteristics based on categories of nearby venues.

We plot a graph of count of most common venue categories in each cluster to identify the type of the cluster.

We perform similar analysis for neighborhoods in New York.

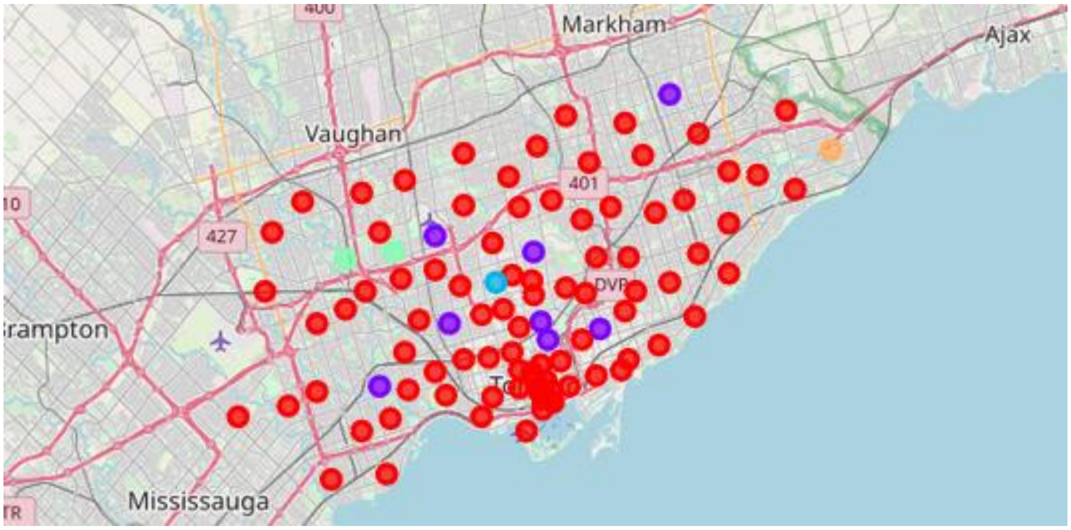
We compare the results of analysis of Toronto neighborhoods with that of New York.

We identify neighborhood cluster in New York that matches with neighborhood cluster in Toronto from where the user is moving. We recommend neighborhood from this identified cluster to the user based on its proximity to her new work place.

Results

The results of k means clustering of Toronto neighborhoods indicate that there are five clusters namely

- Cluster 0: Quite a few neighborhoods fall under this category. It has quite a few diverse venues like coffee shops, Cafes, Grocery Stores, Pizza Places, Parks etc.
- Cluster 1: 8 neighborhoods fall under this category. It has mostly Parks.
- Cluster 2: Only 1 neighborhood falls under this category. It has Garden.
- Cluster 3: Only 2 neighborhoods fall under this category. It has Food Truck and Baseball Field.
- Cluster 4: Only 1 neighborhood falls under this category. It has Bar.

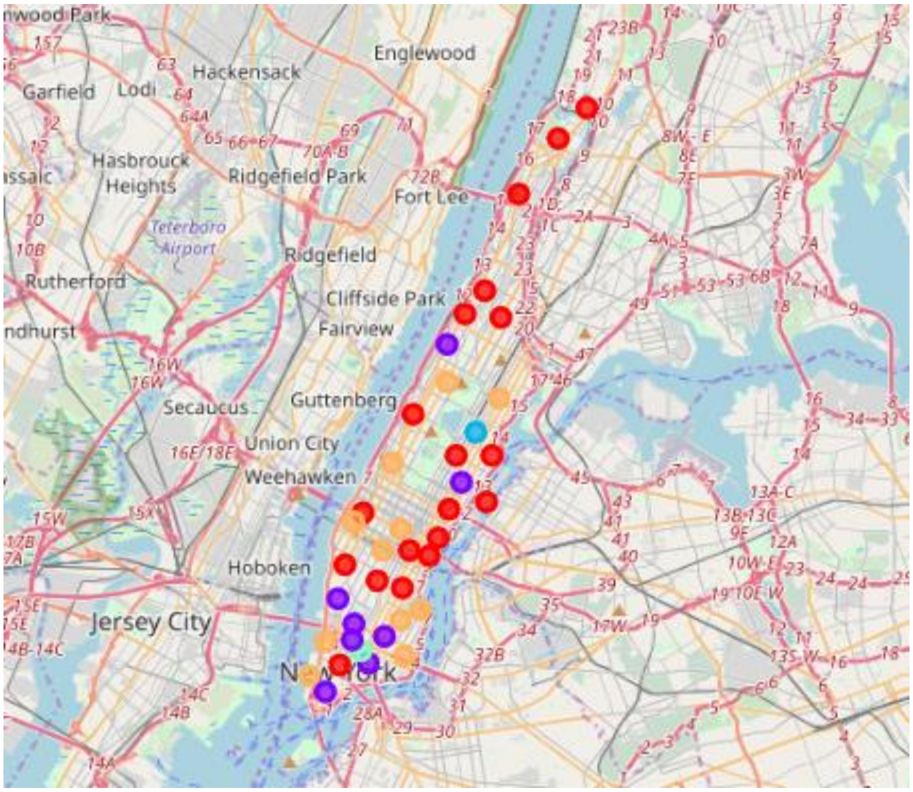


Our neighborhood of interest namely Caledonia-Fairbanks falls in cluster 1 with top 10 venue categories as:

Park	Market	Women's Store	Dumpling Restaurant	Drugstore	Donut Shop	Doner Restaurant	Eastern European Restaurant	Dance Studio	Distribution Center

The results of k means clustering of New York City neighborhoods indicate that there are five clusters namely

- Cluster 0: It consists of single neighborhood with Italian restaurant.
- Cluster 1: It consists of diverse venues like Coffee Shops, Italian Restaurants, cafes, fitness centers, theater etc..
- Cluster 2: It consists of venues like Parks, Italian restaurants, Cafés, Pizza Place, Theater, American Restaurant, Gym.
- Cluster 3: It consists of venues like Mexican and Korean Restaurant, Plaza, Hotel and Bar.
- Cluster 4: It consists of single neighborhood with Pizza Place.



These two cities are similar in terms of their neighborhoods. We can map these neighborhood clusters as:

Toronto	New York City
Cluster 0	Cluster 1
Cluster 1	Cluster 2
Cluster 2	Cluster 0
Cluster 3	Cluster 3
Cluster 4	Cluster 4

As you can easily see the closest type of neighborhoods for my Toronto based friend in New York city is cluster 2 You can also verify the similarity looking at the neighborhoods and their top 10 common venue categories as below:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	2	Gym	American Restaurant	Sandwich Place	Coffee Shop	Yoga Studio	Deli / Bodega	Supplement Shop	Steakhouse	Shopping Mall	Seafood Restaurant
2	Manhattan	Washington Heights	40.851903	-73.936900	2	Café	Bakery	Grocery Store	Mobile Phone Shop	Supplement Shop	Sandwich Place	Mexican Restaurant	Coffee Shop	Liquor Store	Spanish Restaurant
8	Manhattan	Upper East Side	40.775639	-73.960508	2	Italian Restaurant	Exhibit	Art Gallery	Bakery	Coffee Shop	Gym / Fitness Center	Juice Bar	Hotel	French Restaurant	Pizza Place
9	Manhattan	Yorkville	40.775930	-73.947118	2	Italian Restaurant	Coffee Shop	Gym	Bar	Deli / Bodega	Pizza Place	Sushi Restaurant	Japanese Restaurant	Mexican Restaurant	Diner
11	Manhattan	Roosevelt Island	40.762160	-73.949168	2	Park	Coffee Shop	Deli / Bodega	Sandwich Place	Scenic Lookout	Gym	Dry Cleaner	Baseball Field	Liquor Store	Outdoors & Recreation
14	Manhattan	Clinton	40.759101	-73.996119	2	Theater	Italian Restaurant	Gym / Fitness Center	American Restaurant	Coffee Shop	Spa	Sandwich Place	Hotel	Wine Shop	Cocktail Bar
20	Manhattan	Lower East Side	40.717807	-73.9	2	Café	Coffee Shop	Pizza Place	Art Gallery	Cocktail Bar	Japanese	Bakery	Ramen	Chinese	Nightclub

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
				80890							Restaurant		Restaurant	Restaurant	
21	Manhattan	Tribeca	40.721522	-74.010683	2	Park	Italian Restaurant	American Restaurant	Café	Spa	Boutique	Wine Bar	Wine Shop	Coffee Shop	Greek Restaurant
22	Manhattan	Little Italy	40.719324	-73.997305	2	Café	Bakery	Bubble Tea Shop	Sandwich Place	Salon / Barbershop	Mediterranean Restaurant	Italian Restaurant	Cocktail Bar	Yoga Studio	Women's Store
25	Manhattan	Manhattan Valley	40.797307	-73.964286	2	Pizza Place	Indian Restaurant	Bar	Yoga Studio	Thai Restaurant	Playground	Coffee Shop	Mexican Restaurant	Deli / Bodega	Vietnamese Restaurant
26	Manhattan	Morningside Heights	40.808000	-73.963896	2	Park	Bookstore	American Restaurant	Coffee Shop	Burger Joint	Sandwich Place	Deli / Bodega	Pub	Café	Seafood Restaurant
28	Manhattan	Battery Park City	40.711932	-74.016869	2	Park	Coffee Shop	Hotel	Wine Shop	Gym	Shopping Mall	Women's Store	Memorial Site	Food Court	Men's Store
31	Manhattan	Noho	40.723259	-73.988434	2	Italian Restaurant	French Restaurant	Hotel	Cocktail Bar	Rock Club	Art Gallery	Sushi Restaurant	Grocery Store	Gift Shop	Pizza Place

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
39	Manhattan	Hudson Yards	40.756658	-74.000111	2	American Restaurant	Italian Restaurant	Coffee Shop	Hotel	Gym / Fitness Center	Café	Thai Restaurant	Burger Joint	Dog Run	Restaurant

You can pick the actual neighborhood to move to base on its proximity to the new work place and / or venues in the neighborhood.

Based on top 10 venues we recommend Battery Park City. Comparing the top 10 venues in these two neighborhoods you can easily conclude that there are quite a few common venues like Parks, Women's Store, Market, and Restaurants making my friend’s transition a smooth one!

Discussions

As you can see Toronto has around 100 neighborhoods with New York City having 40 neighborhoods. Toronto has 266 unique venue categories with different neighborhoods having different number of venues and some neighborhoods even reaching the limit of 100 for the number of venues. There are quite a few Coffee Shops, variety of restaurants, parks, stores etc. New York has 341 unique venue categories. Most of the neighborhoods have reached the limit of 100 with the remaining having overall a very high count. There is a variety of Coffee Shops, various restaurants, fitness centers, parks, theater, hotel and plaza.

Limitations and Further Work

In this project we only consider the factor of frequency of occurrences of venues of certain category to categorize neighborhoods. It can be augmented by many other factors (e.g. other statistical measures like variance, various percentiles etc. of this number, cost of living, family factors, availability of housing, public transportation etc.). We are also relying on data from foursquare for our analysis. We are also using the free developer's account imposing some restrictions on the data that we obtain. The study can be enhanced further w.r.t. both these factors as well.

Conclusion

Toronto and New York City are similar in terms of their neighborhoods being large metropolitan cities of developed countries like Canada and New York, respectively both from North America on the east coast in nearby vicinity. A person hailing from Caledonia-Fairbanks is advised to pick Battery Park City for settling down while moving from Toronto to New York City. You can carry out similar analysis for any neighborhood in these cities or extend it for migration between any other cities / countries by gathering similar data for those two cities.