Introduction

People are traveling around world to different cities. We are so familiar with our own hometown, we know where to go if we want to go shopping, which area is best for night life and so on. Once we go to a different city, we are able to find a particular store to shop, or restaurants for dinner using various mapping tools without difficulty. However it is not easy to search a neighborhoods that fit our needs. For example, Person A might like a place with lots of restaurants but less bar, more education institutes. Person B might like place with lots of recreations but less restaurants. How could we utilized data and machine learning to help us make decision and find appropriate neighborhoods? This is the problem I would like to address in this capstone project. In this project, I am going to use New York City and Toronto as an example, and use Foursquare location data and cluster machine learning to group the location to different group by their venues information. And then try to find the most relevant neighborhoods between the two cities given the preference you provided.

Data

We get neighborhood data for New York and Toronto from https://cocl.us/new_york_dataset and https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The New York data is Jason file and Toronto data is obtained using BeautifulSoup web scraping. In addition, we obtained Neighborhood and associated venues information from Foursquare.

Methodology

We first obtain the Borough, Neighborhood, Latitude and Longitude information from web using Json phaser and BeautifulSoup. The snapshots of the New York (left) and Toronto (right) result is shown below.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Table 1

	Borough	Neighborhood	Latitude	Longitude
0	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	Scarborough	Woburn	43.770992	-79.216917
4	Scarborough	Cedarbrae	43.773136	-79.239476

Table 2

Then we retrieve Venue information of each neighborhood in New York and Toronto for Toronto from Foursqure as shown below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Category ID
0	Rouge,Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant	4bf58dd8d48988d16e941735
1	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar	4bf58dd8d48988d116941735
2	Guildwood,Morningside,West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place	4bf58dd8d48988d1ca941735
3	Guildwood,Morningside,West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store	4bf58dd8d48988d122951735
4	Guildwood,Morningside,West Hill	43.763573	-79.188711	Marina Spa	43.766000	-79.191000	Spa	4bf58dd8d48988d1ed941735

Table 3: Toronto Venues

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Category ID
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop	4bf58dd8d48988d1d0941735
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy	4bf58dd8d48988d10f951735
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop	4bf58dd8d48988d1c9941735
3	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898276	-73.850381	Caribbean Restaurant	4bf58dd8d48988d144941735
4	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station	4bf58dd8d48988d113951735

Table 4: New York Venues

Since the venues obtained are sub categories, we mapping back the venues to main categories which are 'Arts & Entertainment', 'College & Education', 'Event', 'Food', 'Nightlife', 'Outdoors & Recreation', 'Professional', 'Residence', 'Shops' and 'Travel'. And to one-hot encode to get features of each Neighborhood for New York and Toronto as shown below.

	Neighborhood	Arts & Entertainment	College & Education	Food	Nightlife	Outdoors & Recreation	Professional	Shops	Travel
0	Allerton	0	0	18	0	1	0	12	2
1	Annadale	0	0	5	1	0	0	3	1
2	Arden Heights	0	0	3	0	0	0	2	0
3	Arlington	0	0	3	0	0	0	1	1
4	Arrochar	0	0	10	0	2	0	4	3

Table 5: New York Neighborhood Features

	Neighborhood	Arts & Entertainment	College & Education	Food	Nightlife	Outdoors & Recreation	Professional	Shops	Travel	Cluster Labels
0	Adelaide,King,Richmond	8	0	64	6	6	1	11	4	4
1	Agincourt	0	0	2	1	1	0	0	0	0
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	0	0	1	0	2	0	0	0	0
3	Albion Gardens,Beaumond Heights,Humbergate,Jam...	0	0	5	0	0	0	4	0	3
4	Alderwood,Long Branch	0	0	4	1	4	0	1	0	0

Table 6: Toronto Neighborhood Features

Based on features obtained above, we use K-means clustering method to divide the neighborhoods into 5 different clusters. Each cluster represent neighborhoods with similar features. The clustered neighborhoods of New York and Toronto are shown below. The clustering is quite meaningful. For example, the blue are in New York are prosperous tourist ares with lost of entertainment, foods and etc. The blue area in Toronto are also the similar commercial area. After obtain the clustered, we an analyze the similarity and dissimilarity among clusters which is be discussed below.

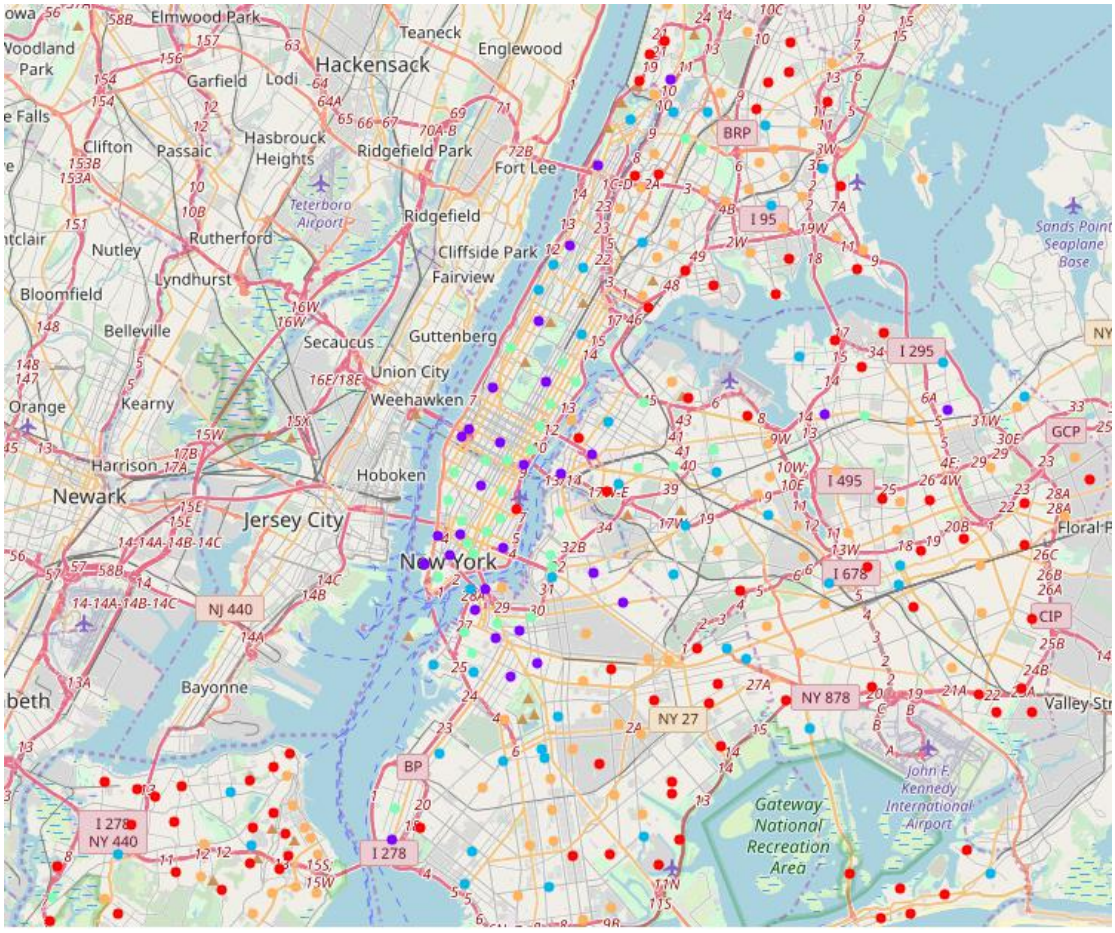

Figure 1: New York Neighborhood Cluster

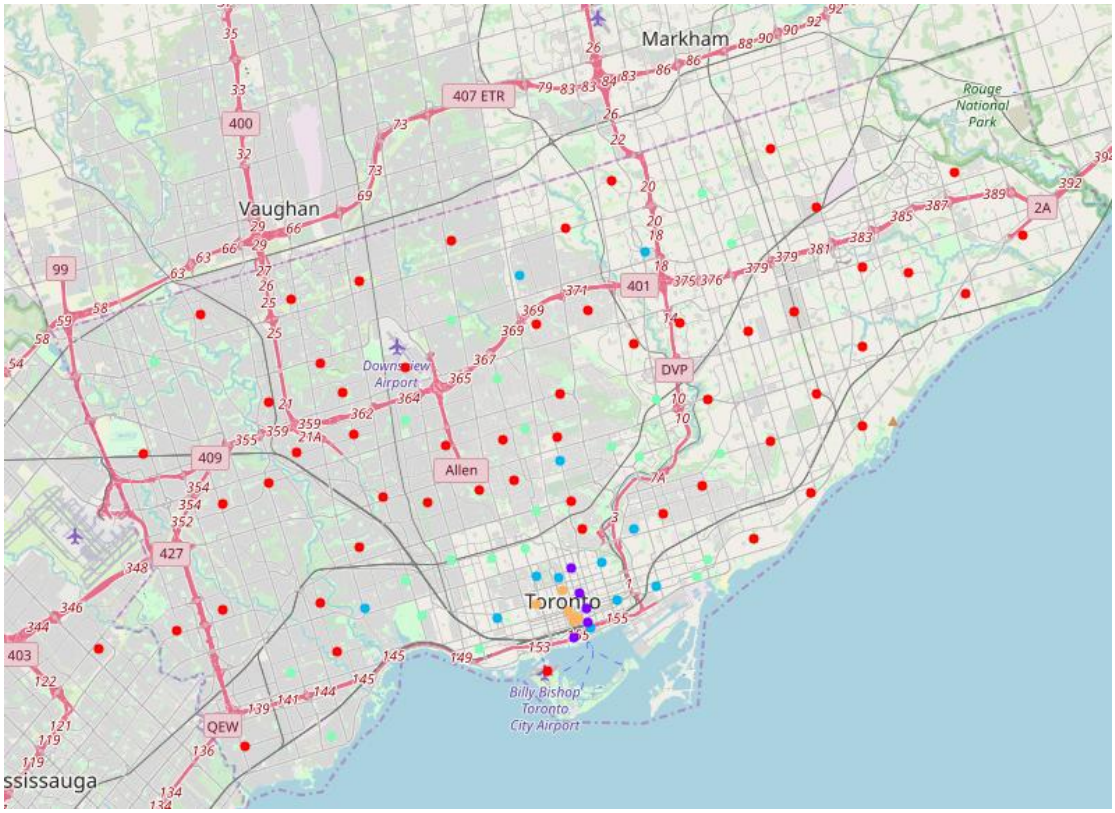

Figure 2: Toronto Neighborhood Cluster

Results & Discussion

To represent each clusters using the main categories, For each cluster, we calculated average number of venues in each category per neighborhood and scale each main category to one. And obtain the features space of cluster in Toronto and New York as shown below (table 7 - table 8). The correlations between each cluster in both Toronto and New York are small, indicating our clustering is stressful (table 9 - table 10). By giving a preference in each categories and using the feature vectors, we calculate the pairwise correlation between Toronto and New York. From the correlation table and map (table 11 and table 12), we can see that cluster in NY [0, 1, 2, 3, 4] is mapped to [[0, 4], 1, 3, 2, 0] in Toronto.

Cluster Labels	0	1	2	3	4
Arts & Entertainment	0.006799	0.523513	0.130878	0.027195	0.311615
College & Education	0.032967	0.362637	0.604396	0.000000	0.000000
Food	0.009392	0.335147	0.166943	0.056984	0.431533
Nightlife	0.006841	0.413852	0.172438	0.038478	0.368391
Outdoors & Recreation	0.073672	0.391691	0.174460	0.101299	0.258876
Professional	0.015852	0.435931	0.072655	0.039630	0.435931
Shops	0.024483	0.411306	0.224423	0.135767	0.204021
Travel	0.060740	0.396710	0.052199	0.037963	0.452388

Table 7: Toronto cluster feature vector

Cluster Labels	0	1	2	3	4
Arts & Entertainment	0.033403	0.545969	0.106890	0.274768	0.038970
College & Education	0.000000	0.311645	0.100419	0.437308	0.150628
Food	0.026936	0.297431	0.174073	0.412467	0.089093
Nightlife	0.016336	0.320534	0.099594	0.522960	0.040576
Outdoors & Recreation	0.071084	0.430853	0.128292	0.301959	0.067812
Professional	0.050940	0.368876	0.108055	0.423505	0.048625
Shops	0.032492	0.298546	0.190491	0.372748	0.105723
Travel	0.159478	0.320789	0.135852	0.222926	0.160955

Table 8: New York cluster feature vector

Cluster Labels	0	1	2	3	4
Cluster Labels					
0	1.000000	-0.290777	0.004175	0.268190	-0.141711
1	-0.290777	1.000000	-0.375882	-0.098555	0.135739
2	0.004175	-0.375882	1.000000	-0.278178	-0.913703
3	0.268190	-0.098555	-0.278178	1.000000	0.021160
4	-0.141711	0.135739	-0.913703	0.021160	1.000000

Table 9: Toronto cluster correlation

Cluster Labels	0	1	2	3	4
Cluster Labels					
0	1.000000	0.013470	0.116131	-0.744936	0.388064
1	0.013470	1.000000	-0.411633	-0.474702	-0.550057
2	0.116131	-0.411633	1.000000	-0.180539	0.271055
3	-0.744936	-0.474702	-0.180539	1.000000	-0.293758
4	0.388064	-0.550057	0.271055	-0.293758	1.000000

Table 10: New York cluster feature vector

TO	0	1	2	3	4
NY					
0	0.659760	0.011891	-0.577827	0.092849	0.511683
1	-0.008705	0.803229	-0.265294	-0.118272	0.041230
2	0.053134	-0.379861	-0.179012	0.746460	0.119919
3	-0.581879	-0.310630	0.376531	-0.157867	-0.169718
4	0.490063	-0.552161	0.432400	-0.098121	-0.340572

Table 11: New York and Toronto pairwise correlation

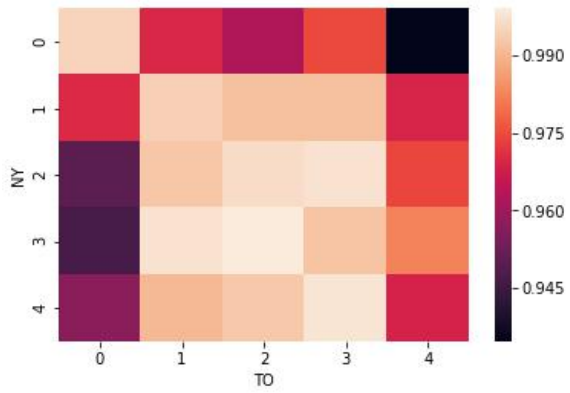

Table 12: New York and Toronto pairwise correlation heatmap

Conclusion

By using the data science method such as web scrapping, data process and machine learning, we can find the most relevant neighborhoods between the New York given the preference you provided.