**The Battle of Neighborhoods**

**Lingmei Kong**

**Introduction**

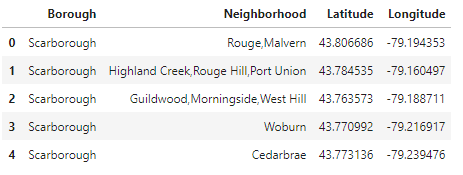
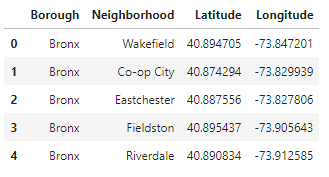
People are traveling around world to different cities. We are so familiar with our own hometown, we know where to go if we want to go shopping, which area is best for night life and so on. Once we go to a different city, we are able to find a particular store to shop, or restaurants for dinner using 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.

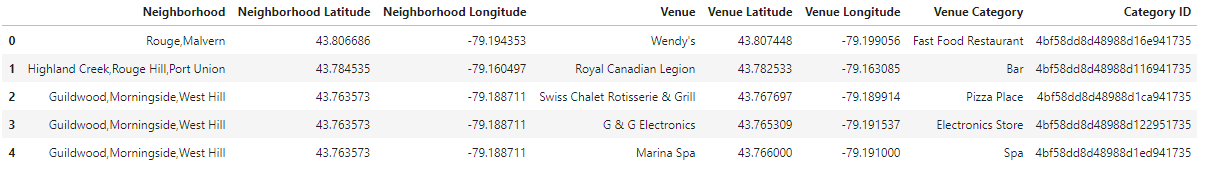
**Methodology**

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.



***Table 1 Table 2***

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

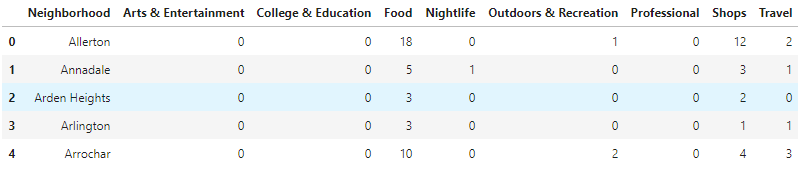


***Table 3: Toronto Venues***

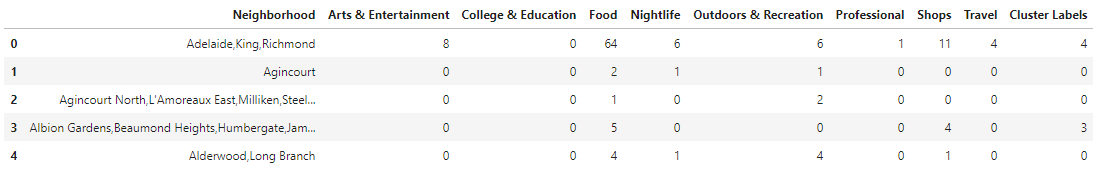


***Table 4: New York Venues***

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.

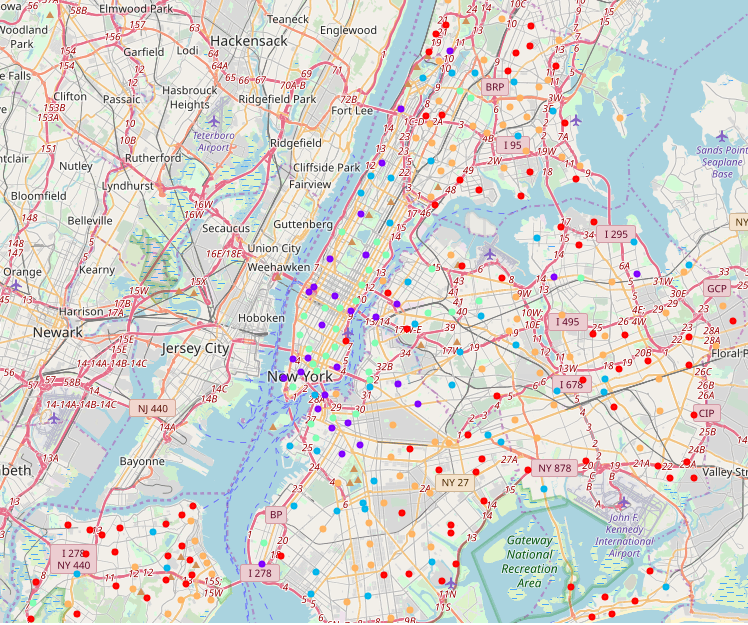


***Table 5: New York Neighborhood Features***

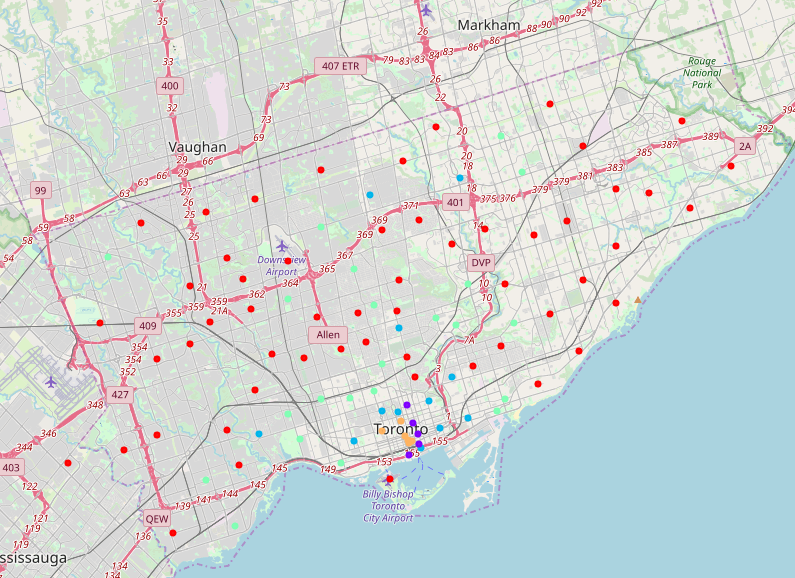


***Table 6: Toronto Neighborhood Features***

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**

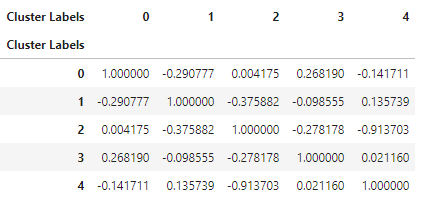
To represent each clusters using the main categories, For each cluster, we calculated average number of venues in each category per neighborhood and scale each main category to one. And obtain the features space of cluster in Toronto and New York as shown below (table 7 - table 8). The correlations between 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 above pairwise correlation, a mapping from NY to Toronto is: NY[0, [1,4] [2, 3]] to TO[0,[1,4], [2,3]]. That means we can see that if a person A interested in cluster 0 in NY, then A should to cluster 0 in Toronto. If A like cluster 4 area in Toronto, A should select cluster 1 or 4 cluster area in NY.



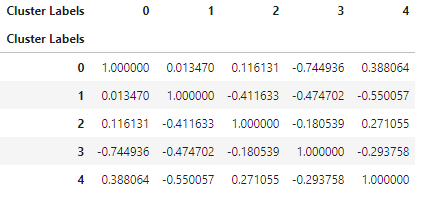
***Table 7: Toronto cluster feature vector***



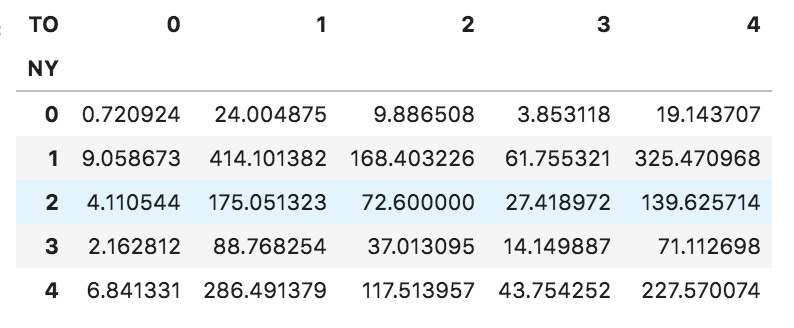
***Table 8: New York cluster feature vector***



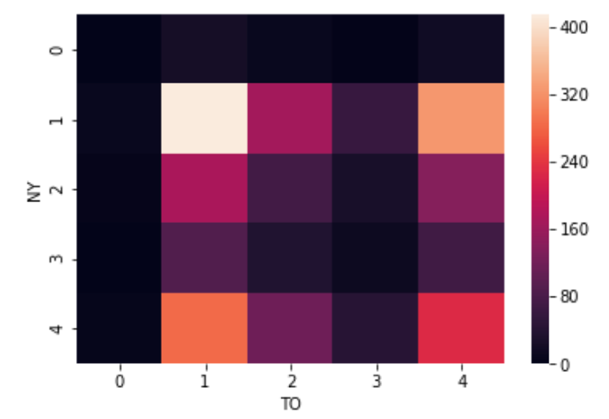
***Table 9: Toronto cluster correlation***



***Table 10: New York cluster feature vector***



***Table 11: New York and Toronto pairwise covariance***



***Table 12: New York and Toronto pairwise correlation heatmap***

**Conclusion**

The project has the following findings: New York has 5 boroughs and 306 neighborhoods. Toronto has 11 boroughs and 103 neighborhoods. New York and Toronto can be divided to 5 cluster area with distinct venues differences. New York cluster 1 and 4 is most prosperous area and both mainly located in Manhattan Toronto cluster 1 and 4 is also most prosperous area and both mainly located in financial district near the water in Toronto. Cluster 0 in both cities are the perimeter zone. Cluster 2 and 3 are local commercial areas. The detailed neighborhood in each can be found in "Neighborhood in each cluster section".