Mark Polinkovsky
May 18, 2020

# Finding a Similar Neighborhood to Move to in NYC and Toronto

## Introduction

People moving to a different city often want to settle in a neighborhood that is similar to the one they have been living in. Because the person is new to the city, they may not be familiar with the various neighborhoods and their individual characters. How can they choose a neighborhood to make sure it fits their familiar lifestyle? Given their home neighborhood, its "venues" can be analyzed and matched to similar neighborhoods in the new city. This would be a valuable service for real estate agents to provide for their customers, to help them narrow down the neighborhood choices. It will make the customer happy by helping them find the ideal neighborhood for their lifestyle, and will help the real estate agent by improving the quality of service they offer, and by helping them narrow their focus to the desired neighborhoods.

## Data

For the customer's current location, the current address will be needed. The address will be used to find its latitude and longitude. The data needed in the new cities is the list of neighborhoods and their postal codes, to find their latitude and longitude. For the new cities, the Foursquare data on venues near the neighborhood centers will be needed. The same information for venues around the customer's current address will be needed. For this project, we will focus on Toronto and New York City as the new cities, since we have the neighborhood information already.

The new cities' neighborhoods will be clustered using k-means clustering, based on the types of the top 30 venues in that neighborhood. The original neighborhood's venue types for the top 30 venues will also be calculated. The K-means model will then be used to predict which cluster of neighborhoods in each of the new cities the original address matches best with. This way, the neighborhoods in the new city that belong to that cluster will be returned as matches by similarity to the customer's old neighborhood.

## Methodology

Since we are focusing on New York City and Toronto, we need information about the neighborhoods in both cities. For Toronto, I used the previous week's data set that matches the postal codes with the neighborhood names. From there, I used Geocoder to query Arcgis for the latitude and longitude coordinates of each neighborhood. For New York City, I found two datasets on neighborhoods, the most comprehensive being the NYC Open Data set: https://data.cityofnewyork.us/City-Government/Neighborhood-Names-GIS/99bc-9p23
This data set provided over 200 neighborhoods in the city, along with their latitude and longitude coordinates.

Once we have the coordinates for each neighborhood, we can use the Foursquare API to get a list of the top venues in the neighborhood. I opted for a radius of 750 m, and the 30 most popular venues for both cities. At this point, we can do one-hot encoding of the venues by their type, and group them in each neighborhood.

We can now proceed to the K-means clustering of the neighborhoods in each city. The first thing is to determine how many clusters to use. For this, we can use the Elbow method to see where the decrease in cluster distance slows with the increasing number of clusters. This will give us a reasonable value for the number of clusters to use in our model.

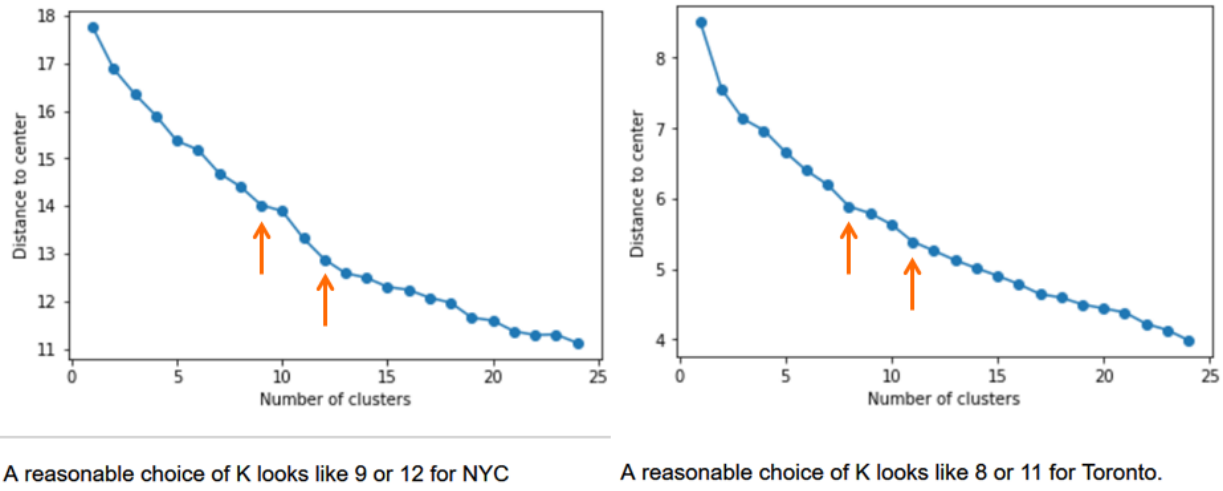



*Fig 1:* Using the Elbow method to find the optimal number of clusters for neighborhoods in NYC and Toronto

The number of clusters were chosen as K=9 and K=8 for NYC and Toronto, respectively. These values minimize the number of single-neighborhood clusters, while preserving the general structure of the large groups of clusters.

Once we have our K-means clustering algorithm tuned, we can use it to generate the clustering model for each city.

Finally, we can turn to the customer's current address. Geocoder can again be used to query the Arcgis API to get the latitude and longitude. The Foursquare API can be used to get a list of venues around the address. After one-hot encoding of the venues by type, we can use our previously-trained models to predict which cluster the current address will fit into.

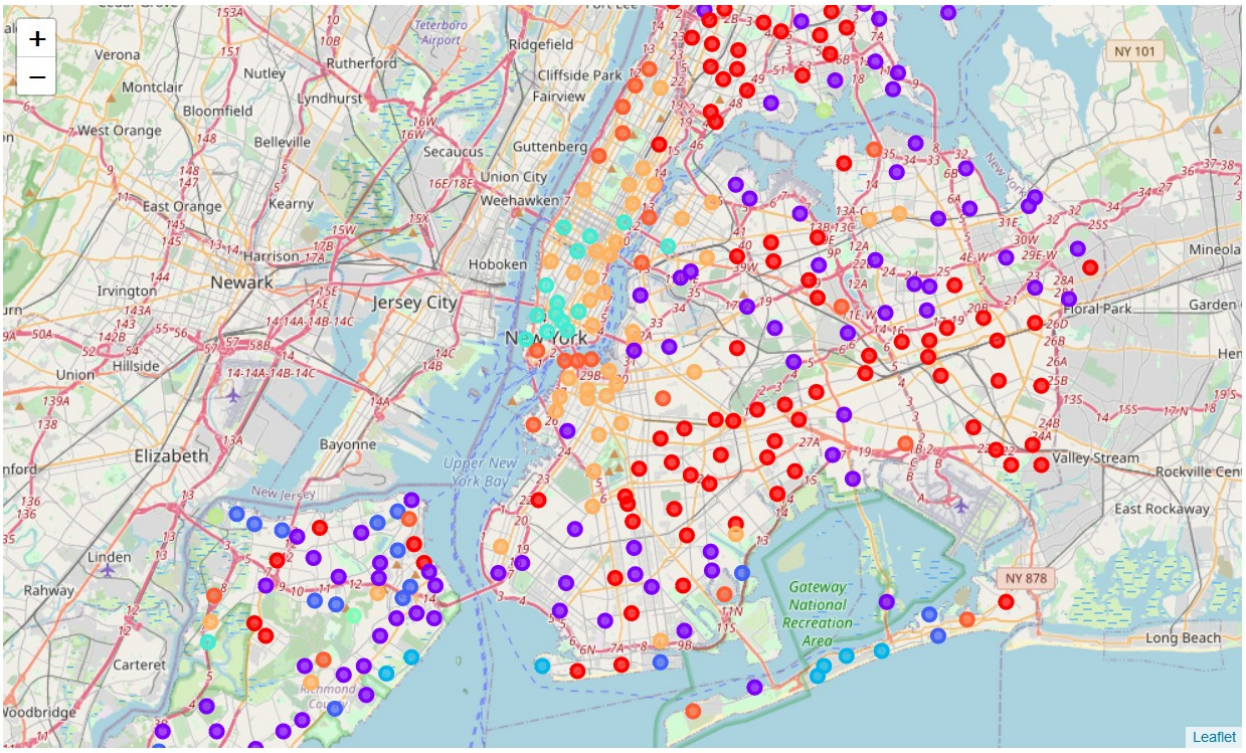



*Fig 2:* Clusters of neighborhoods in NYC

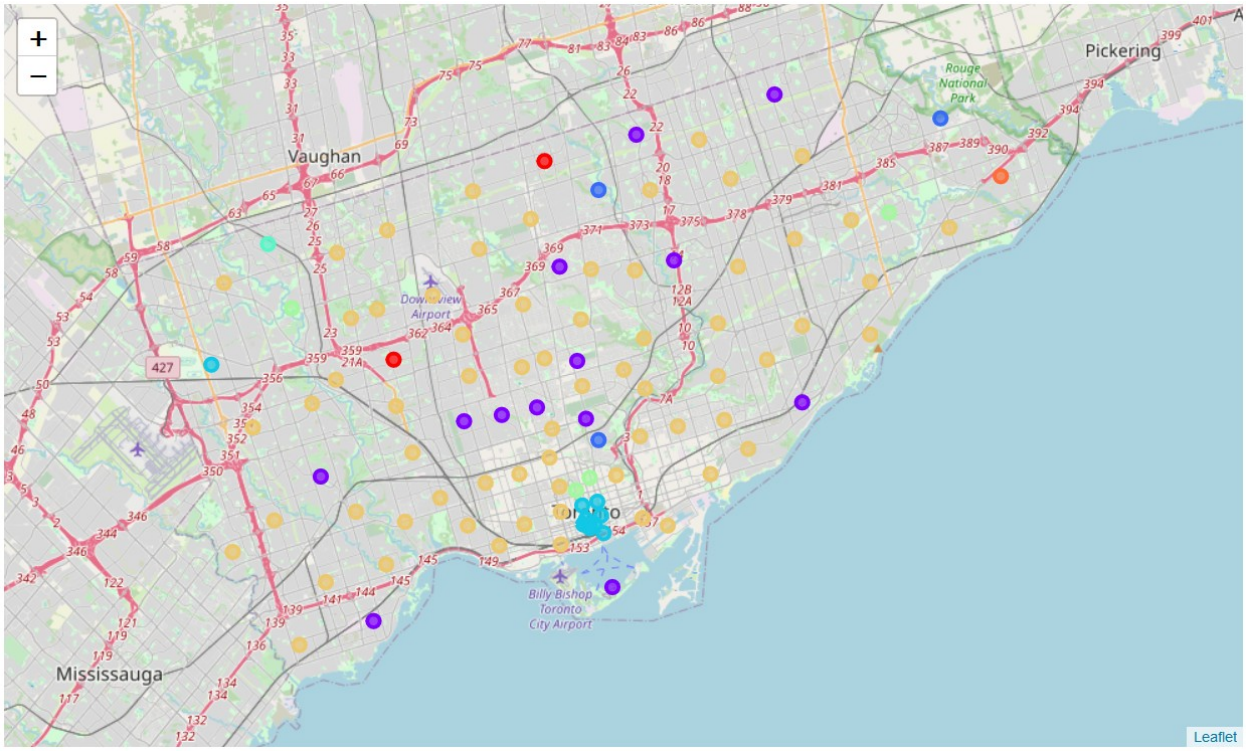
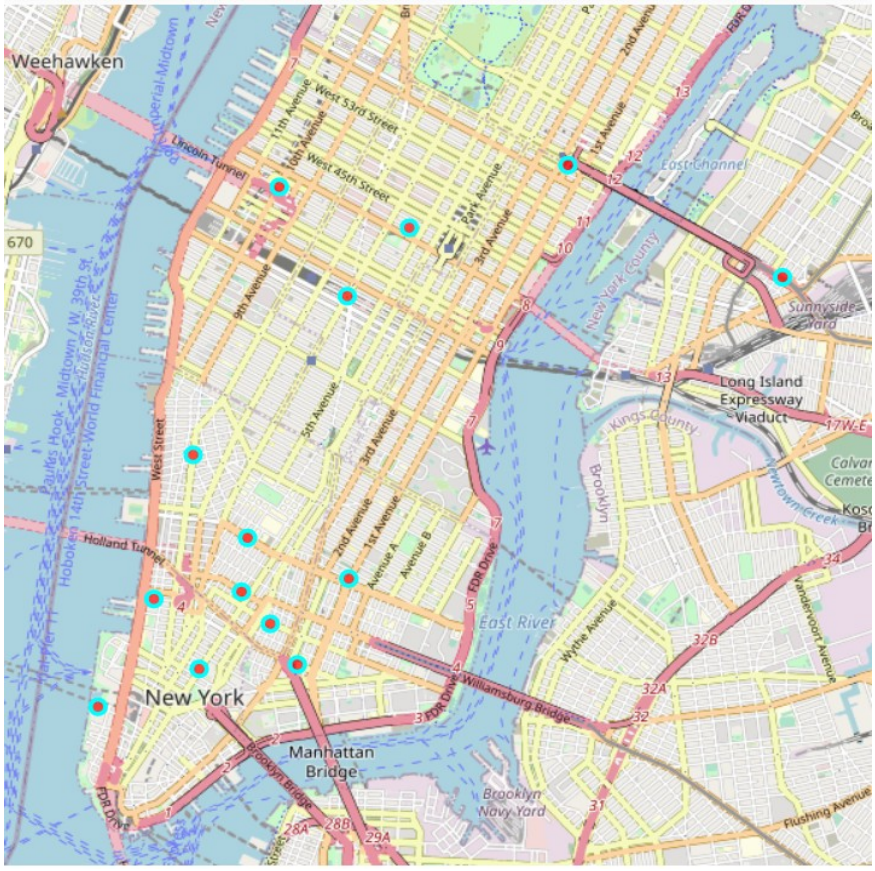


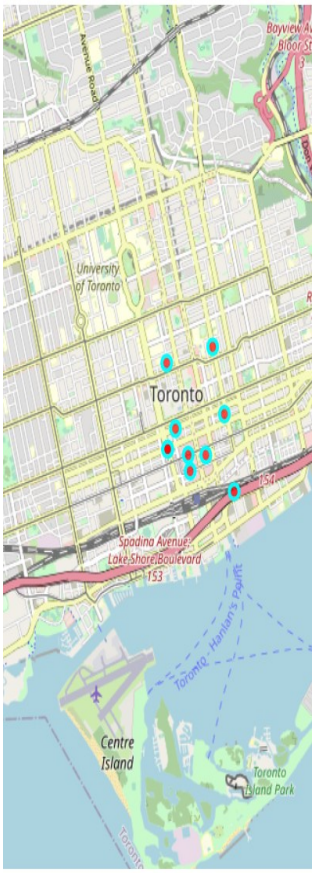

*Fig. 3:* Clusters of neighborhoods in Toronto

**Results / Discussion**

I used the addresses for various landmarks around the world to find the most similar neighborhoods in NYC and Toronto.

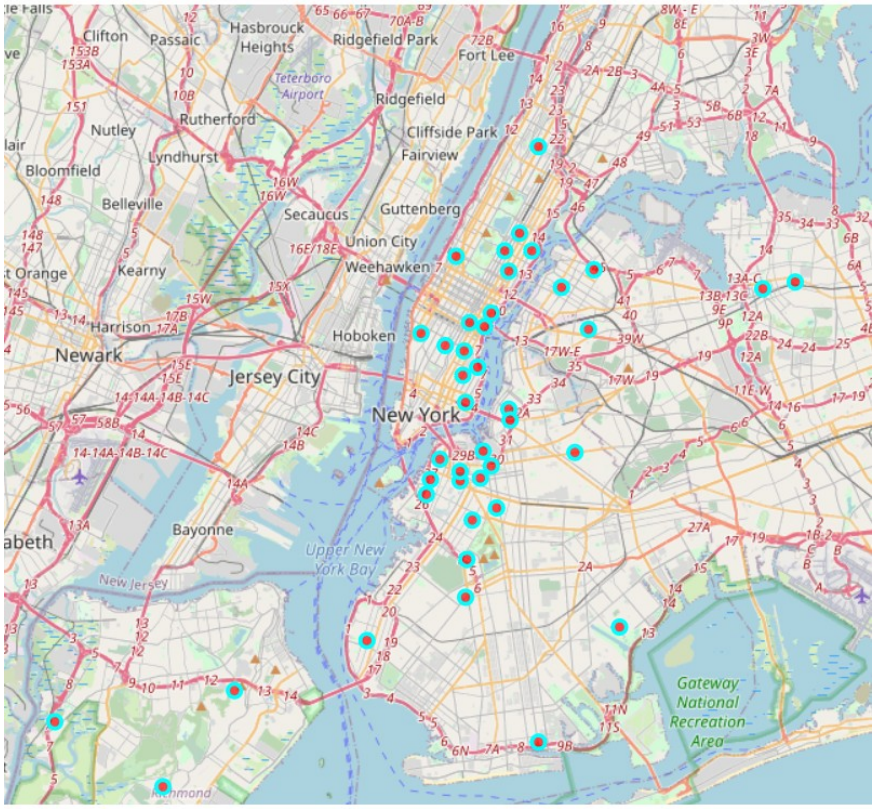
1. Looking at Chicago City Hall at 121 N LaSalle St, Chicago, IL 60602, the map of matching NYC and Toronto neighborhoods are:

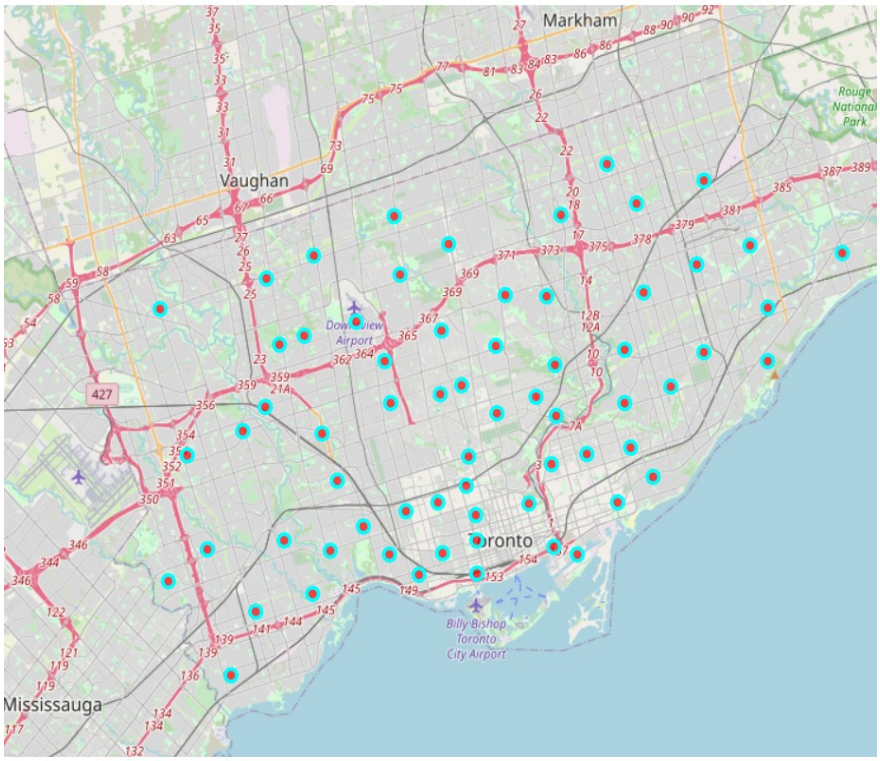


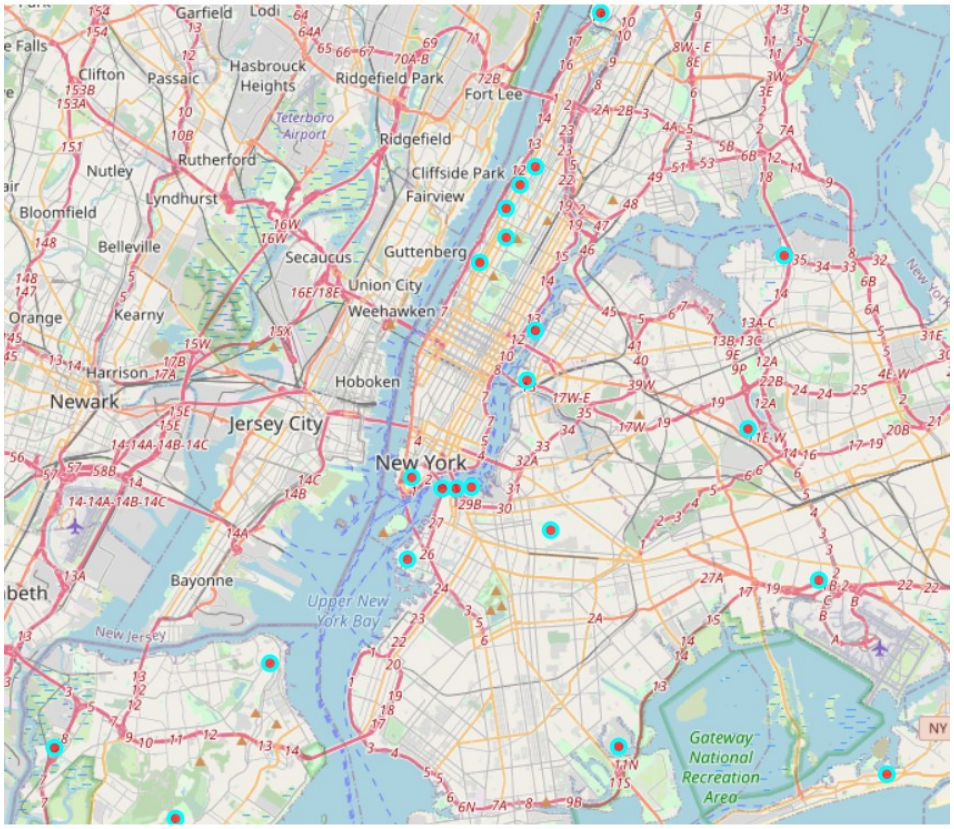

Clearly, these neighborhoods are the political and business centers of the two cities.

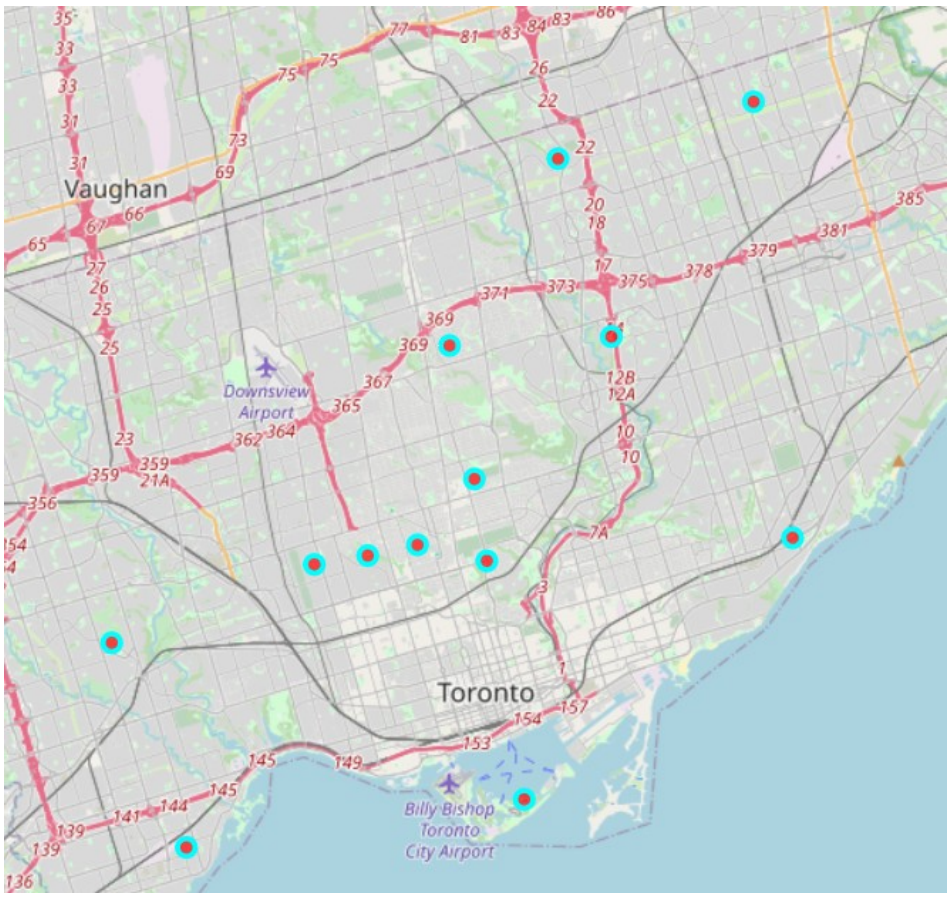
2. Turning to the California Institute of Technology at 1200 E California Blvd, Pasadena, CA. This is a dense suburban area, so we would expect some "inner suburban" neighborhoods to be matched. That turns out to be the case:



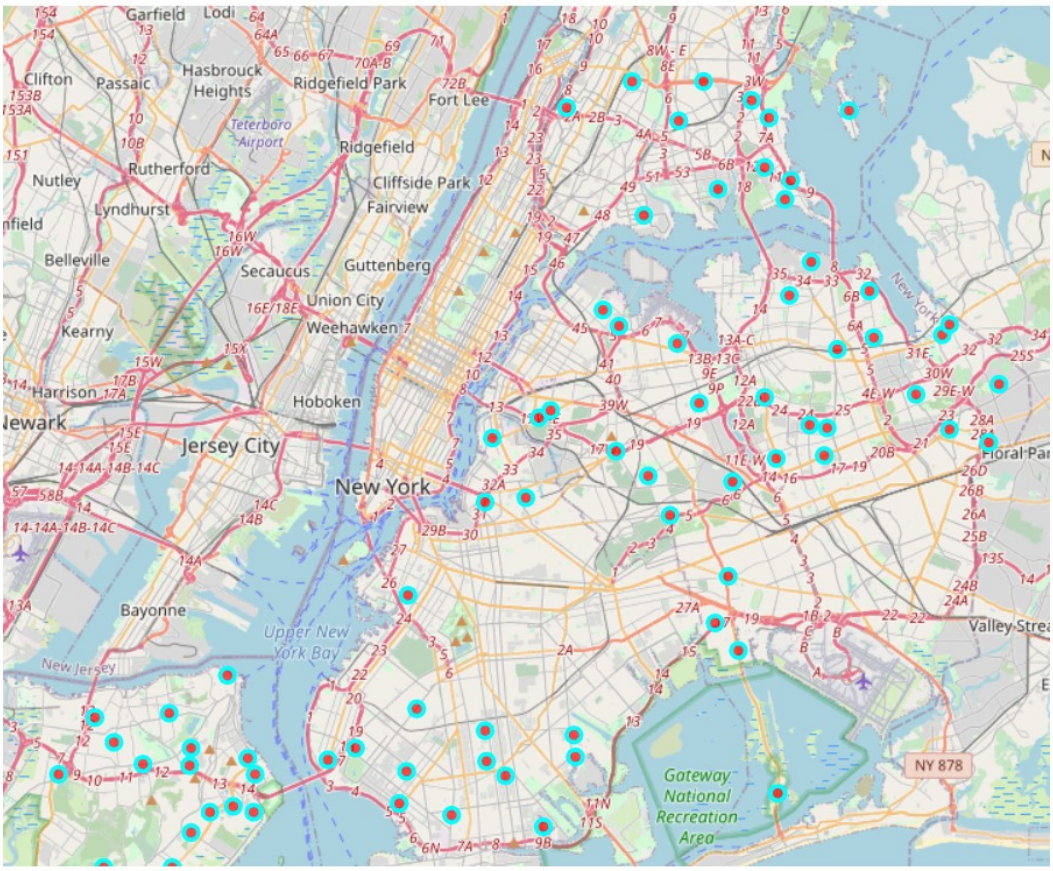



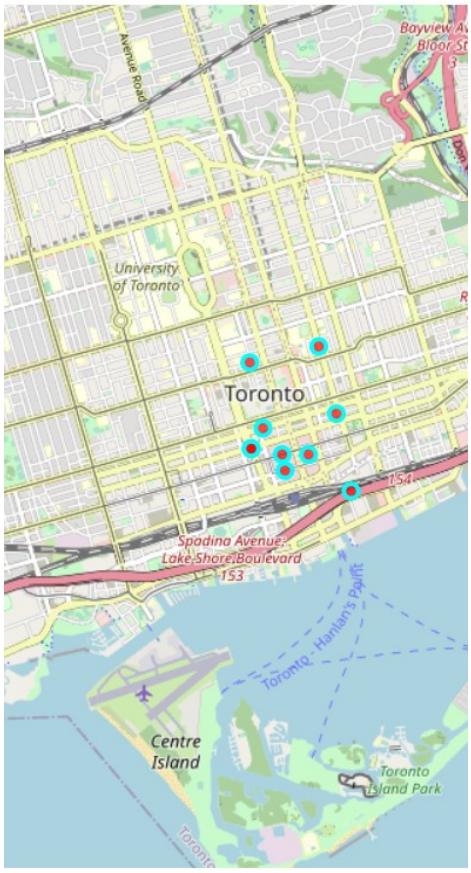
3. Now, we look at a more sparsely-populated area, the village of Santa Murtiola in San Marino: The address of the church there is 47890, San Marino. We see the match with more "park-like" neighborhoods in NYC, and the outer, rural suburbs in Toronto.



4. Finally, we look at the small town of Iqaluit in the remote north of Canada. The address for St. Jude's Anglican Cathedral is 655 Mattaaq, Iqaluit, NU X0A 0H0, Canada. Here we have an interesting result in the cluster predictions. The NYC neighborhoods matched are the more remote ones. In contrast, the Toronto neighborhoods matched are in the center of the city (which we saw before, for Chicago City Hall).




This suggests that the clusters in each city are being selected for different features. In Toronto, the government buildings are playing a bigger part. In NYC, the other features of the neighborhoods are more prominent. This is also an interesting demonstration of the idea that the clusters in the two cities, while having similar features, are not always chosen identically, due to the different weighting of those features.

## Conclusion

We have shown how addresses around the world can be matched with neighborhoods in NYC and Toronto, based on the types of venues in those locations. The neighborhoods in a city can be clustered by those features, and the resulting model can be used to match any location with those neighborhood clusters. A similar analysis can be extended to any other city, as long as neighborhood information is available to get the latitude and longitude coordinates of those neighborhoods.