CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS

1. Introduction

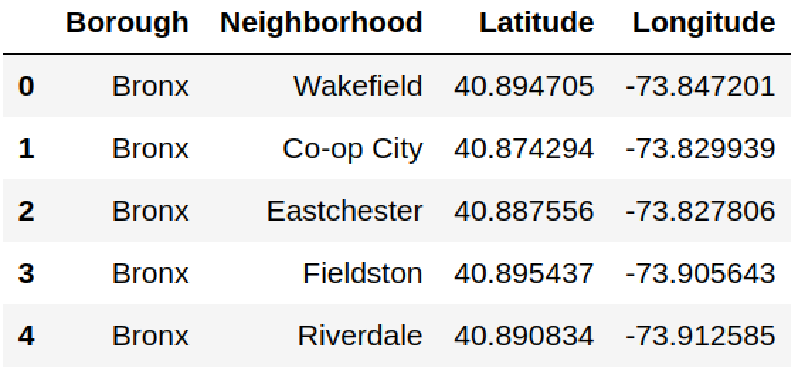
In a globalized world, It is very common for a company to transfer an employee to another city or even country, offering better career prospects or salary increase. Or someone could find a better job in other company in other city. No matter the reasons, moving to other city is always very stressful, and people are afraid about not get used with the new city . The problem is worse when the person is married with children. To minimize the problems, it is better to move a place similar to his last location, with compatible venues, like schools, restaurants, swimming pools, gyms, coffee-shops, supermarkets, etc. To help people in this situation, the goal of this project is to develop a system to find out which neighborhoods are similar to the current location. We will simulate a situation where a person is moving from New York to Toronto and vice versa.

2. Data Preparation

2.1 New York City Data

New York city Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and longitude coordinates of each neighborhood.

Luckily, this dataset exists for free on the web and here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572. New York data can be downloaded from this link. This data would be *json* format and it can be transformed into *pandas dataframe*. New York datataset is showed in the figure 1, and the location are represented in the figure 2.



	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

Figure 1 – New York city dataset samples

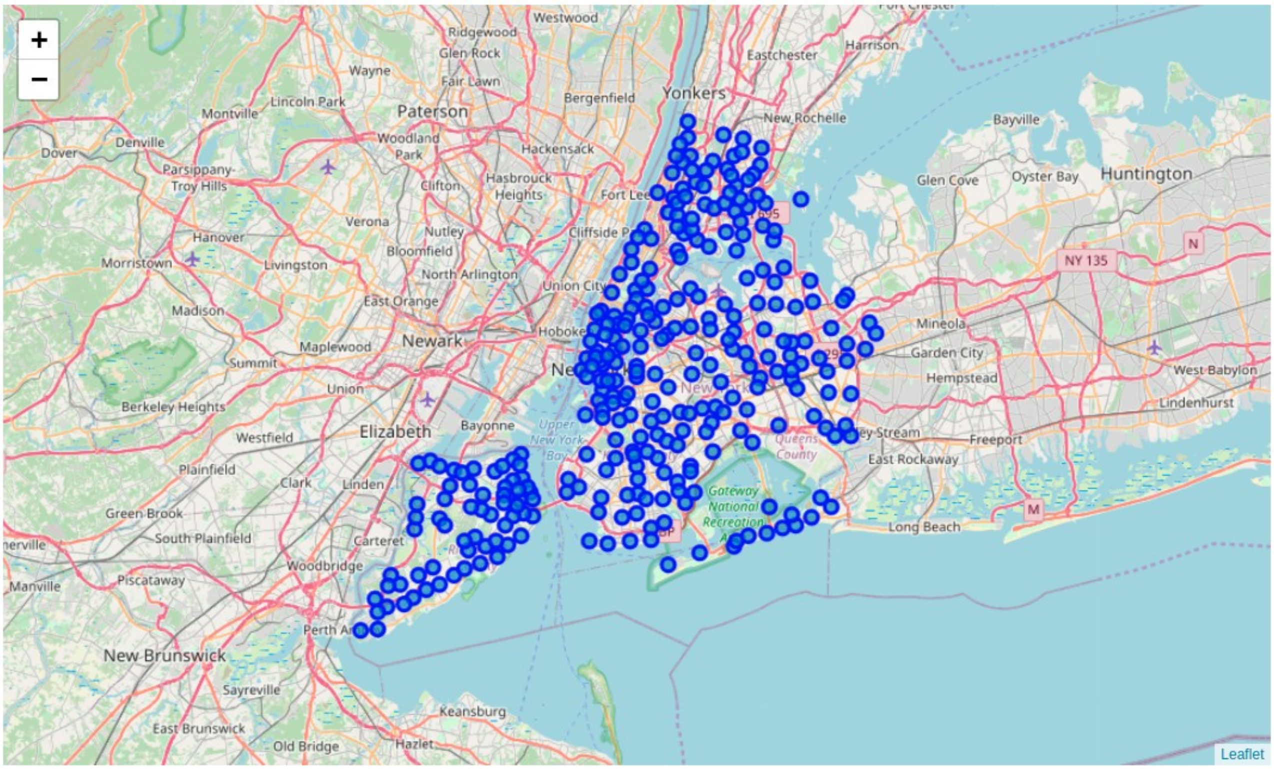


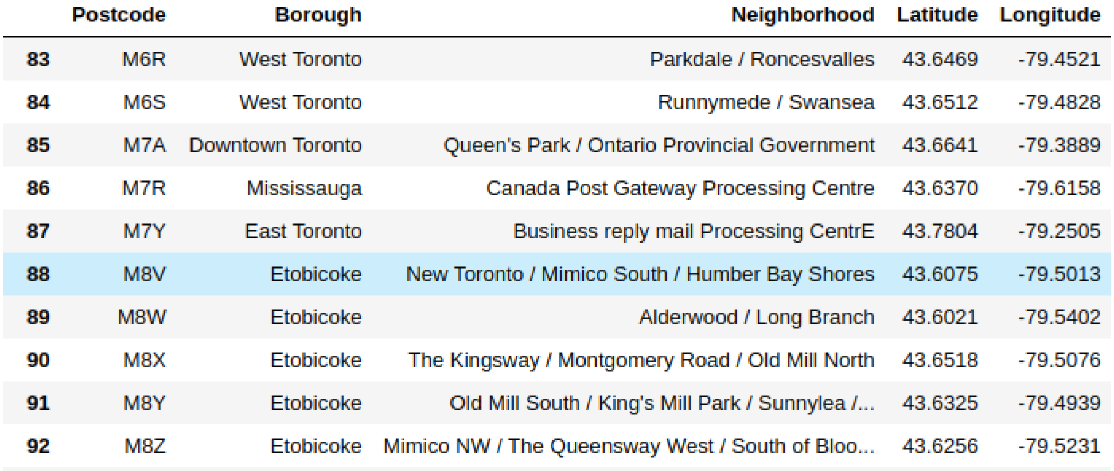
Figure 1 – New York city neighborhoods location

2.2 Toronto City Data

Unlike New York, the neighborhood data is not readily available on the internet. So we needed to use a notebook to build the code to scrape the following Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, in order to obtain the data that is in the table of postal codes and to transform the data into a pandas dataframe. We have to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas dataframe so that it is in a structured format like the New York dataset. There are some rows where “Not assigned” is written. Drop the rows where borough is “Not assigned” (only). If Neighborhood is “Not assigned” but Borough is assigned then make corresponding Borough as Neighborhood. We Merged the rows if Postal Code and Borough of two or more rows are the same and merged neighborhood will be separated by comma “,”. Latitude and longitude information can be downloaded from here. We tried to use Geocode but it was not working! So I used pgeocode. I preferred to do this instead of loading the file from the link. Geocode failed to find the M7R coordinates. So, I got the coordinates from Google and inserted in the table.

Finally this data and Toronto data can be merged together. After both data are ready we used Foursquare API to get the venues near each neighborhood.

Toronto datataset is showed in the figure 3, and the location are represented in the figure 4.



	Postcode	Borough	Neighborhood	Latitude	Longitude
83	M6R	West Toronto	Parkdale / Roncesvalles	43.6469	-79.4521
84	M6S	West Toronto	Runnymede / Swansea	43.6512	-79.4828
85	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government	43.6641	-79.3889
86	M7R	Mississauga	Canada Post Gateway Processing Centre	43.6370	-79.6158
87	M7Y	East Toronto	Business reply mail Processing CentrE	43.7804	-79.2505
88	M8V	Etobicoke	New Toronto / Mimico South / Humber Bay Shores	43.6075	-79.5013
89	M8W	Etobicoke	Alderwood / Long Branch	43.6021	-79.5402
90	M8X	Etobicoke	The Kingsway / Montgomery Road / Old Mill North	43.6518	-79.5076
91	M8Y	Etobicoke	Old Mill South / King's Mill Park / Sunnylea /...	43.6325	-79.4939
92	M8Z	Etobicoke	Mimico NW / The Queensway West / South of Bloo...	43.6256	-79.5231

Figure 3 – Toronto Dataset samples

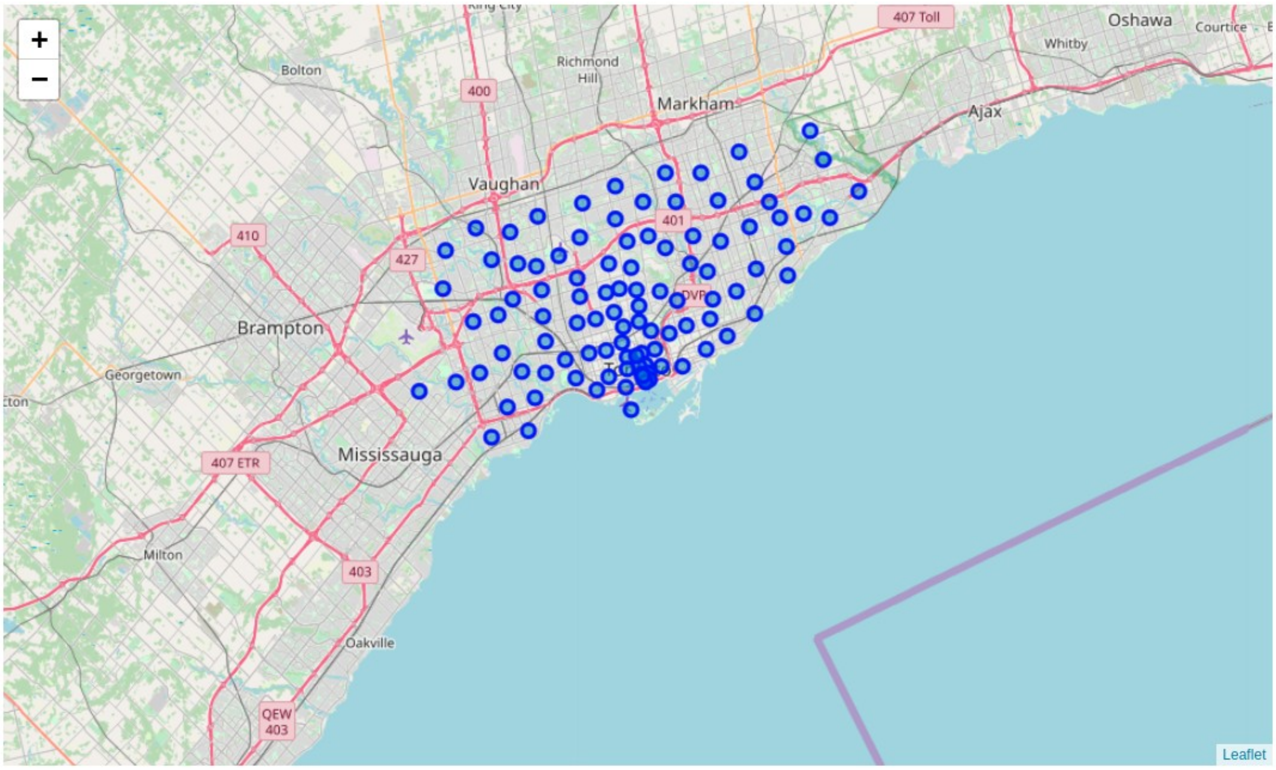
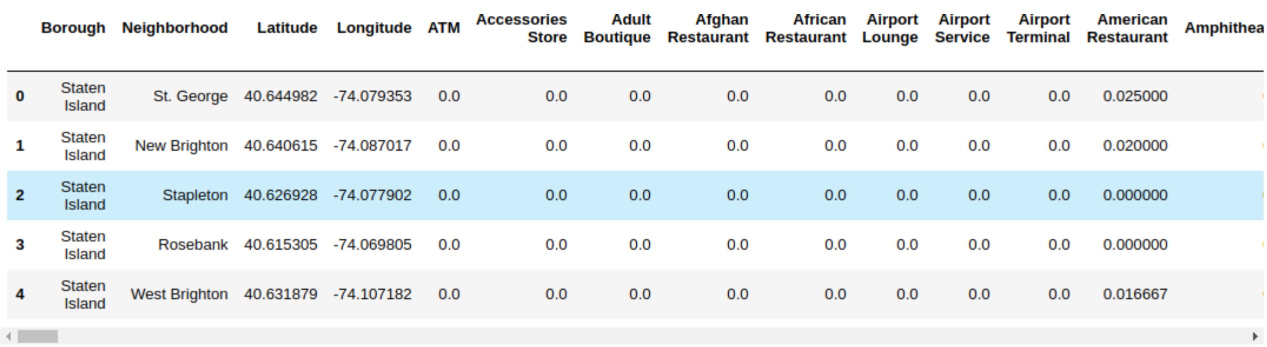


Figure 4 – Toronto city neighborhoods location (just postcode with M)

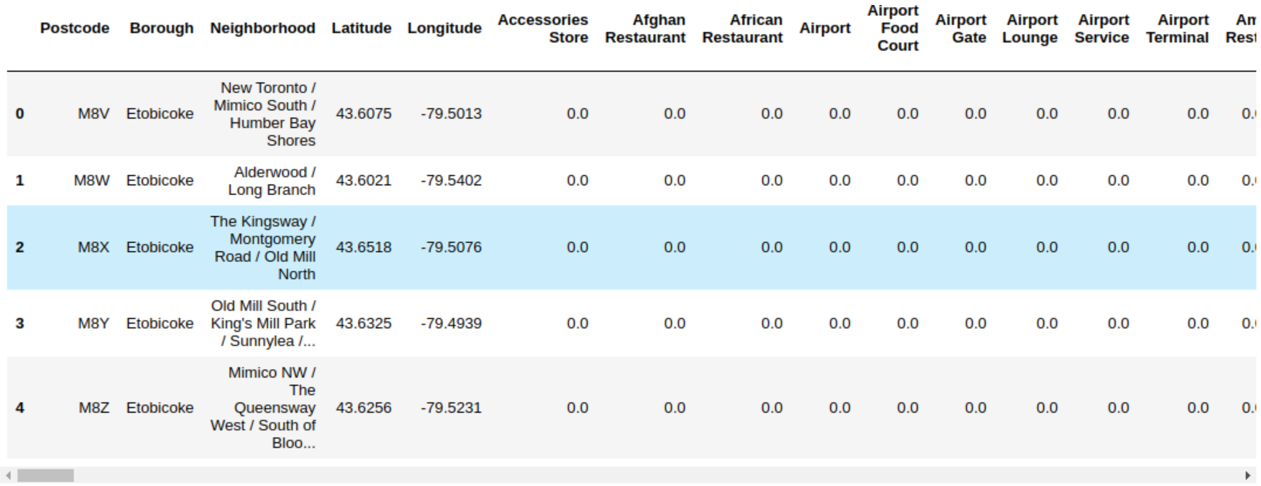
2.3 Using Foursquare to Get Venue Data

Foursquare was used to get venues information from each neighborhood. First, it has been necessary to obtain the client ID and the secret to get access to the online API. Figures 5 and 6 show samples of the New York and Toronto venues summary, respectively.



	Borough	Neighborhood	Latitude	Longitude	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphithea
0	Staten Island	St. George	40.644982	-74.079353	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.025000	
1	Staten Island	New Brighton	40.640615	-74.087017	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	
2	Staten Island	Stapleton	40.626928	-74.077902	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
3	Staten Island	Rosebank	40.615305	-74.069805	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
4	Staten Island	West Brighton	40.631879	-74.107182	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.016667	

Figure 5 – New York Venues dataset



	Postcode	Borough	Neighborhood	Latitude	Longitude	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	Am Resi
0	M8V	Etobicoke	New Toronto / Mimico South / Humber Bay Shores	43.6075	-79.5013	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.
1	M8W	Etobicoke	Alderwood / Long Branch	43.6021	-79.5402	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.
2	M8X	Etobicoke	The Kingsway / Montgomery Road / Old Mill North	43.6518	-79.5076	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.
3	M8Y	Etobicoke	Old Mill South / King's Mill Park / Sunnylea /...	43.6325	-79.4939	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.
4	M8Z	Etobicoke	Mimico NW / The Queensway West / South of Bloo...	43.6256	-79.5231	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.

Figure 6 – Toronto Venues dataset

3. Methodology

New York and Toronto dataset are very similar, but some columns are different. First, it was analyzed how many venue categories are common in both cities dataset. The result was:

- Number of common venue categories in both data are :312
- Number of different venue categories in New York city are : 164
- Number of different venue categories in Toronto city are : 31

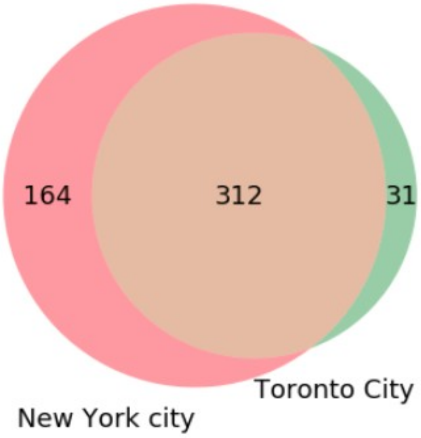


Figure 7 - Venn-Diagram of Venue Categories. 312 categories are the same

As we had 312 common categories, we used only these categories to compare the venues of both cities.

The measure the similarity of the locations, we used the Cosine similarity.

4. Results

We had implemented a function where the input are the current city , current borough current neighborhood and the number of most similar locations.

First, we simulated the case where someone is moving from Moving from Wingate, Brooklyn, NY, to Toronto. The result is represented in the figure 8 and 9.

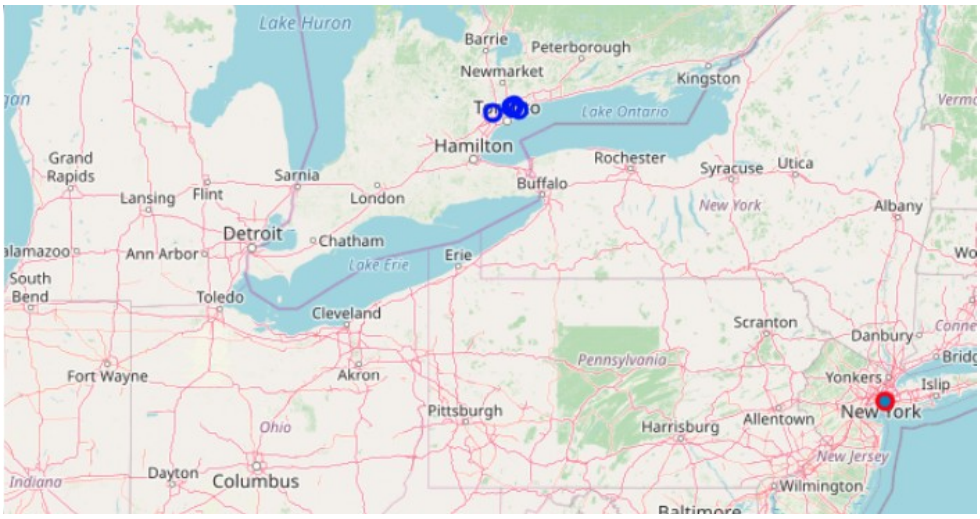


Figure 8 – Five most similar locations In Toronto

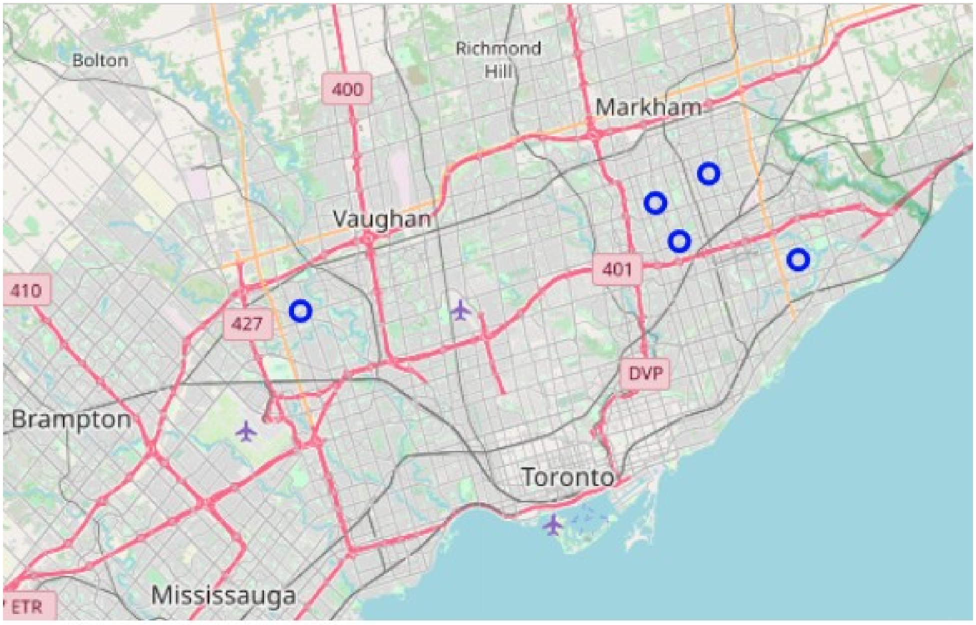


Figure 9 – Five most similar locations in Toronto (zoom)

The finded locations were:

- South Steeles / Silverstone / Humbergate / Jamestown / Mount Olive / Beaumond Heights / Thistletown / Albion Gardens
- Steeles West / L'Amoreaux West
- Milliken / Agincourt North / Steeles East / L'Amoreaux East
- Clarks Corners / Tam O'Shanter / Sullivan
- Woburn

Finally, we simulated the case where someone is moving from Moving from Woburn, Scarborough, to New York .The result is represented in the figure 10 and 11.

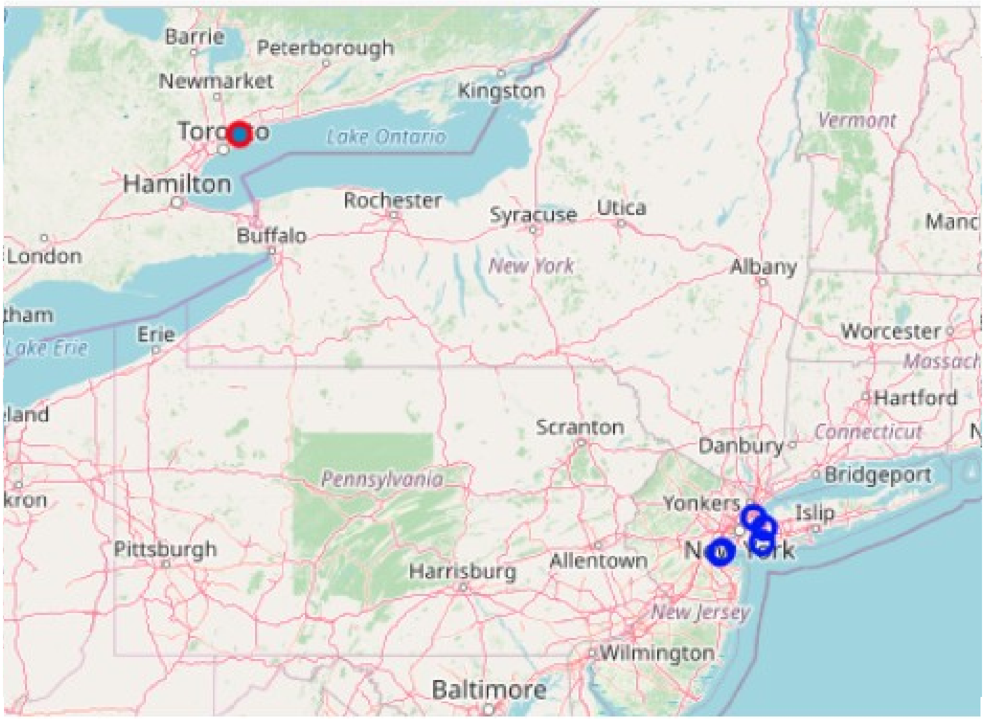


Figure 10 – Five most similar locations in New York

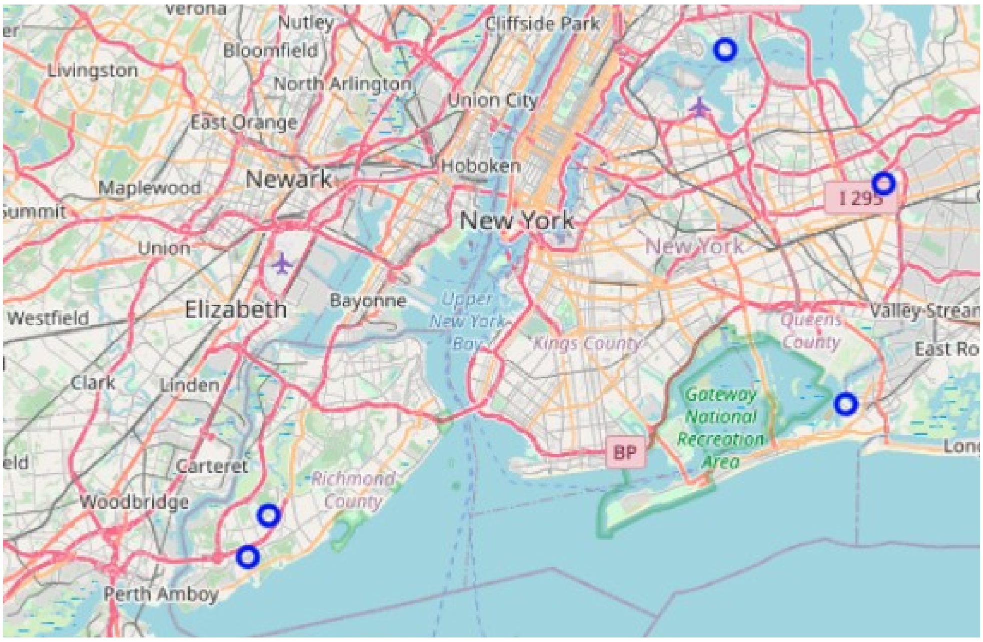


Figure 11– Five most similar locations in New York (zoom)

The finded locations were:

- Clason Point
- Prince's Bay
- Bayswater
- Bellaire
- Arden Heights

5. Discussion

This project used several data analysis tools to help someone to choose a new place to live. The first challenge was to find useful data, and in this case we used location data. Location data is data describing places and venues, such as their geographical location, their category, working hours, full address, and so on, such that for a given location given in the form of its geographical coordinates (or latitude and longitude values) one is able to determine what types of venues exist within a defined radius from that location. A lot of effort was spent to prepare the data in a suitable format, from different sources. Then, online information provided by Foursquare was used, to complement the data about the venues. Besides that, Folium was useful to show the locations, giving a overview of the locations. Finally, cosine similarity to measure the similarity between the locations. This is a very simple method, and in future works someone could try to implement more sophisticated methods.

6.Conclusion

This software is able to help someone who will move to a new city and would like to find a place to live similar to the his hometown. However, first it is necessary to find the right in formations about both the cities. Fortunately, in most of the cases, the data about the neighborhood localization can be found in the internet. Besides that, with the help of the Foursquare API, information about the venues can be download as a dataset. Then, with some adaptations, the comparison can be performed, and the most similar cities determinated.