IBM Data Science

Capstone Project Report

The Battle of Neighborhoods

Submitted By

Anamitra Saikia

1. Introduction/Business Problem

New York and Toronto are some of the most popular cities in the world. People migrate to other places in search of better opportunities and standards of living. A large number of people immigrate to these cities every year. Some people move from Toronto to New York and vice versa. A person moving from one city to another prefers to stay in a similar locality in the new place with similar facilities and venues in the locality, which helps to accommodate to the new place easily and faster. It also makes life easier. The given project compares different localities and presents the localities in both the cities which are similar in terms of venues and facilities available like shops, coffee shops, transportation, services, etc. It will help people migrating from one city to another to find a locality with similar venues, points of interest, and facilities.

2. Data

Toronto Data

The postal codes, borough, and neighborhood data are scrapped from the Wikipedia [link] using BeautifulSoup. After cleaning and formatting the dataset, it merged with Longitude and Latitude Dataset provided by Coursera [Link]. With the help of Foursquare API, the venues nearby a given neighborhood is obtained.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029
7	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848

Fig: Toronto dataset

New York Data

The postal codes, borough, and neighborhood data are obtained from the given source (Cognitive Class) [link]. The data is in JSON which will be converted to a pandas data frame. The data of the borough, neighborhood, latitude, and longitude will be used in this project.

	Borough	Neighbourhood	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
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

Fig: New York dataset

3. Methodology

The coordinates of the locations are collected from trusted sources and used to collect the venues around the locations using Foursquare API. For this, a Foursquare developer account is needed. There are different types of venues in the cities of New York and Toronto.

Toronto venues:

	PostalCode	Borough	Neighbourhood	Latitude	Longitude	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Amphitheater	Anim Shelt
0	M9W	Etobicoke Northwest	Clairville, Humberwood, Woodbine Downs, West H...	43.706748	-79.594054	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0
1	M8V	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321	0.0	0.0	0.0	0.0	0.0	0.0	0.066667	0.0	0
2	M8W	Etobicoke	Alderwood, Long Branch	43.602414	-79.543484	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0
3	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0
4	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0

Fig: Toronto venues

New York venues:

	Borough	Neighbourhood	Latitude	Longitude	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Ampl
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	0.0	0.015625	
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	0.0	0.026316	
2	Staten Island	Stapleton	40.626928	-74.077902	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	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	0.0	0.024390	
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	0.0	0.020408	

Fig: New York venues

It is found that out of 321 unique venue types in Toronto and 489 unique venue types in New York there are 293 venue types common to both the cities which will be used to find the most similar locations.

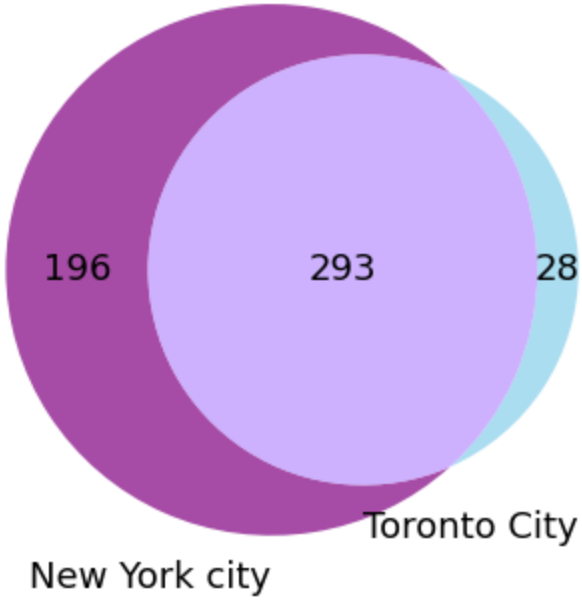
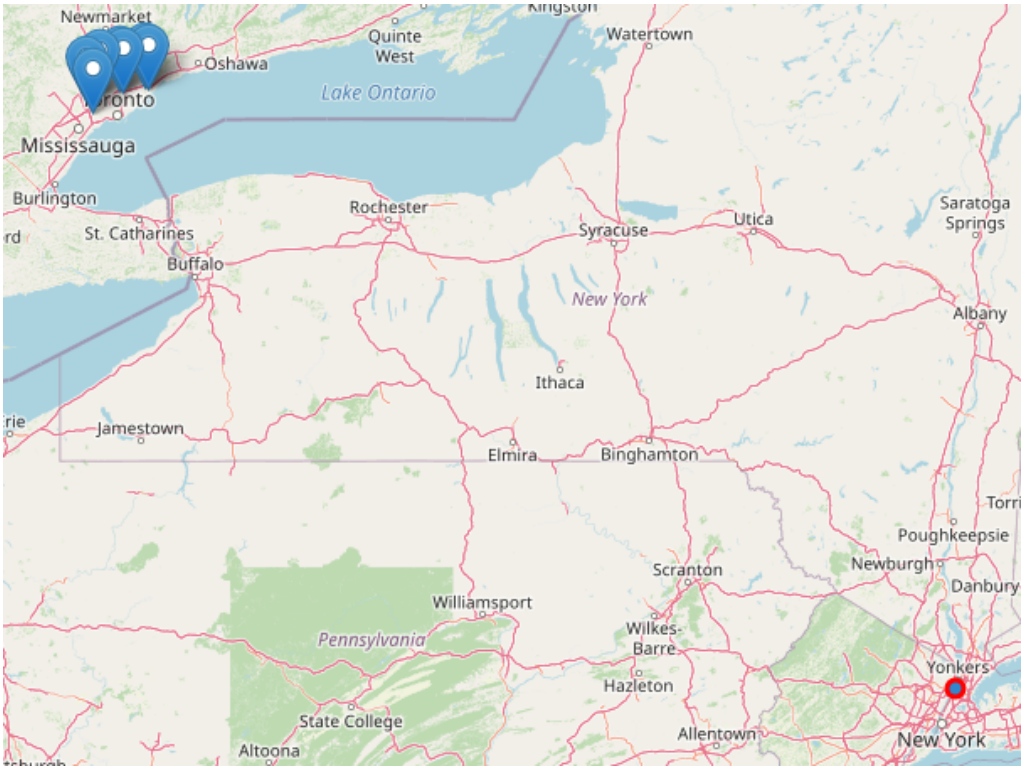


Fig: Venn diagram of categories of venues

Finding Similar Locations:

Cosine Similarity is used for finding the most similar location in the other city. Cosine similarity is a metric used to determine how similar the documents are irrespective of their size. The top *n* number of most similar locations will be selected based on user details including latitude, longitude, and the venues near the location. This kind of recommendation comes under an item-item based recommender system.

Let's assume the current location is Bronx, Riverdale, New York. It will get the index, latitude, and longitude of this location from the New York city data. It can get the all common venue categories data of the corresponding index and then we'll multiply it with Toronto city-data. This will give how much each borough is similar to the current location.



Then these values will be sorted and will obtain the top *n* number of boroughs. Then it will be visualized using the folium.

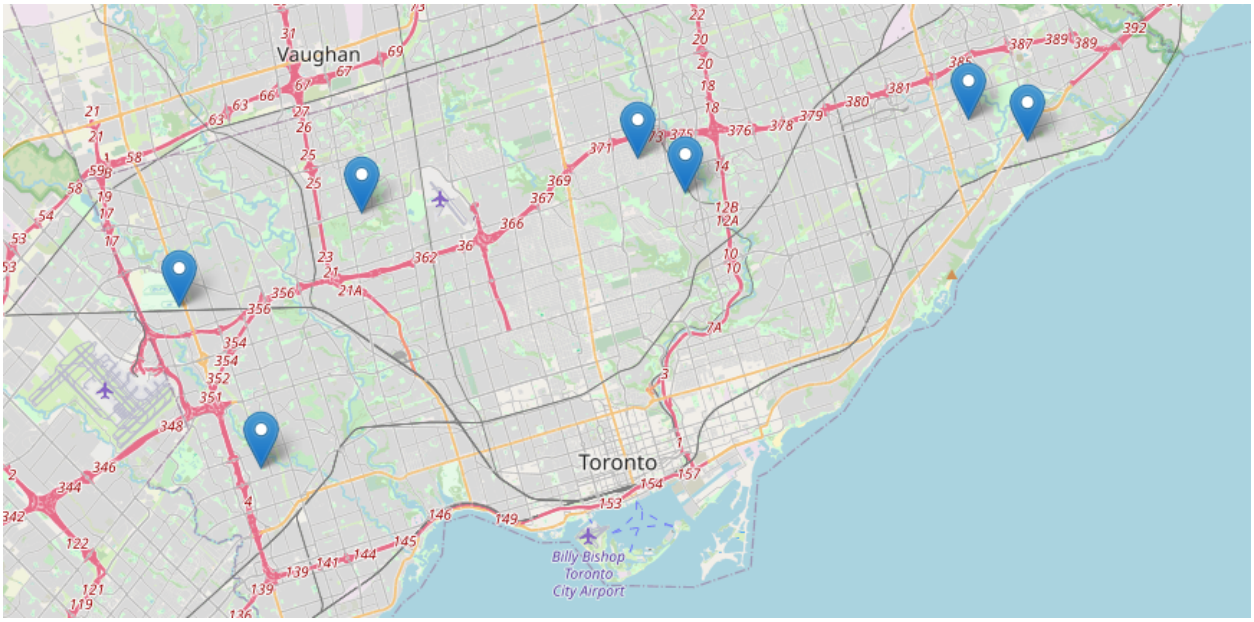


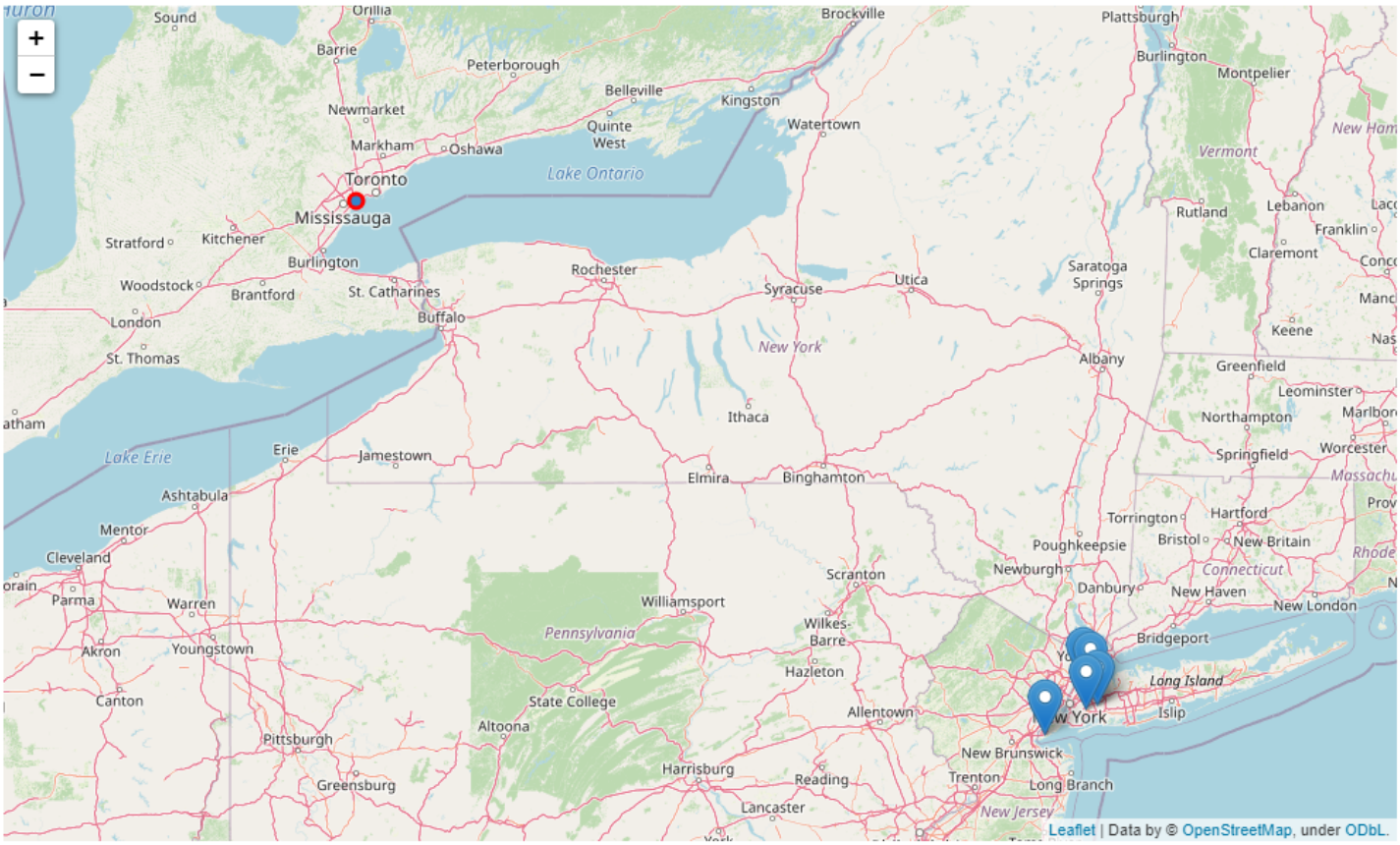
Fig: Locations in Toronto similar to Bronx, NY

4. Results

The given project finds the *n* number of most similar locations in the other city, based on venues of the current city. It will help people to move from one city to another city to find better opportunities and choices without the hassle of finding a new location with all the desired facilities and venues. It takes the information on the index, latitude, and longitude of the current location and outputs the top n locations in the other city which are most similar.

Example:

```
Enter the current City
City: Toronto
Enter the current Borough
Borough: Etobicoke
Enter the current Neigbourhood
Neigbourhood: Alderwood, Long Branch
```



5. Discussion

The proposed project successfully recommends n number of location which are identical to the current location. It helps people to move from Toronto to New York and vice versa. It takes account of the venues of the locations to find the location with similar venues and facilities. But It does not take into consideration of other factors such as climate, social and ethnic composition, the environment which are also important.

6. Conclusion

In this project, Cosine similarity is used to find similar locations in two different cities to help people migrating from one city to another. This model can be implemented for any location or any scale provided the required data is provided. It only considers venues of the location as the factor for finding similar localities but environmental and social factors are also important and play a major role in decision making.