IBM COURSERA CAPSTONE PROJECT

A PLACE LIKE HOME

May 2021

By

Siddarth Nataraj

1. Introduction

This project emphasizes the need to compare neighborhoods and find similarity between them. One main use case for this in the real world is that, if someone is moving from one city to another, they may want to settle in a place that's similar to their old neighborhood. But it's not feasible for the person to visit each neighborhood in the new city to observe and find out the one that they need. As Foursquare API contains humungous amounts of location data, we can leverage that to explore and find out similar neighborhoods in a matter of minutes. In this project, I have compared one neighborhood from Toronto city to all neighborhoods in New York city to find out the ones that are similar. This way a person who's moving into form Toronto to New York, can easily find the place that they like to live.

2. Data

The Toronto neighborhood data is obtained from a Wikipedia page where the actual data is extracted from a table which includes Borough and Neighborhood columns. The Latitude and Longitude data for those neighborhoods are obtained from a CSV file. These two data-frames are joined on the Postal-code column. For convenience to identify the neighborhood in the later section of the problem, a city column has been added with a singular value "Toronto". The resulting data-frame looks like below.

	City	Borough	Neighborhood	Latitude	Longitude
0	Toronto	North York	Parkwoods	43.753259	-79.329656
1	Toronto	North York	Victoria Village	43.725882	-79.315572
2	Toronto	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	Toronto	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	Toronto	Queen's Park	Ontario Provincial Government	43.662301	-79.389494

Figure 1. Toronto data-frame

The New York Neighborhood data is obtained from a JSON file. The columns are similar to Toronto data and like above, a city column is added with value "New York". The resulting data-frame looks like below.

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

Figure 2. New York data-frame

The Foursquare API is used to get the venues in each neighborhood to compare them and identify which are similar to the neighborhood selected from the Toronto data.

3. Methodology

As my Foursquare account is only a Sandbox tier account, 950 daily API calls limit wasn't enough for me to compare all the Toronto neighborhoods to the New York neighborhoods. So I decided to go for a minimized version by taking one neighborhood from Toronto data and find it's similar neighborhoods in New York.

The "Parkwoods" neighborhood from "North York" borough is taken from Toronto data and inserted into a new data-frame along with the New York data-frame. The nearby venues are obtained for the new neighborhoods in the data-frame with the radius set as 300. We get the below number of venues.

```
[33]: venues.shape

[33]: (4937, 7)
```

Figure 3. The dimension of Venues data-frame

The number of venues obtained for each neighborhood is determined as below.

```
[34]: venues.groupby('Neighborhood').count()
```

[34]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Allerton	17	17	17	17	17	17
Annadale	2	2	2	2	2	2
Arden Heights	5	5	5	5	5	5
Arlington	1	1	1	1	1	1
Arrochar	10	10	10	10	10	10
...	...	...	...	...	...	...
Woodhaven	12	12	12	12	12	12
Woodlawn	14	14	14	14	14	14
Woodrow	19	19	19	19	19	19
Woodside	43	43	43	43	43	43
Yorkville	37	37	37	37	37	37

289 rows × 6 columns

Figure 4. Venues vs Neighborhood

In the above there are 379 unique venue categories. Each neighborhood is analyzed and rows are grouped by neighborhood and by taking the mean of the frequency of occurrence of each category. Top 10 venues in each neighborhood is put into a data-frame using a custom function.

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Allerton	Pizza Place	Discount Store	Breakfast Spot	Donut Shop	Spa	Fried Chicken Joint	Supermarket	Bus Station	Fast Food Restaurant	Gas Station
Annadale	Bakery	Train Station	Women's Store	Entertainment Service	Ethiopian Restaurant	Event Service	Event Space	Eye Doctor	Factory	Falafel Restaurant
Arden Heights	Deli / Bodega	Pharmacy	Coffee Shop	Playground	Bus Stop	Women's Store	Farmers Market	Ethiopian Restaurant	Event Service	Event Space
Arlington	Grocery Store	Women's Store	Fast Food Restaurant	Ethiopian Restaurant	Event Service	Event Space	Eye Doctor	Factory	Falafel Restaurant	Farm
Arrochar	Pizza Place	Deli / Bodega	Bus Stop	Liquor Store	Italian Restaurant	Cosmetics Shop	Bagel Shop	Fast Food Restaurant	Event Service	Event Space

Figure 5. Top 10 venues for all neighborhoods

Clustering is done on the above data-frame using the K-Means methodology. The data-frame is divided into 6 clusters. The cluster label for each row is inserted to indicate to which cluster the respective neighborhood belongs.

4. Results

Using the city column, the Toronto neighborhood that was earlier selected, is identified and from the cluster label column, it is found out that it belongs to cluster 4.

```
[50]: merged.loc[merged['City']=='Toronto']
```

[50]:

	City	Borough	Neighborhood	Latitude	Longitude	Cluster Labels
0	Toronto	North York	Parkwoods	43.7533	-79.3297	3.0

Figure 6. The Toronto neighborhood cluster

When the cluster 4 is investigated, it is found out that four other neighborhoods from New York has been clustered within. These neighborhoods are visualized on the New York map.

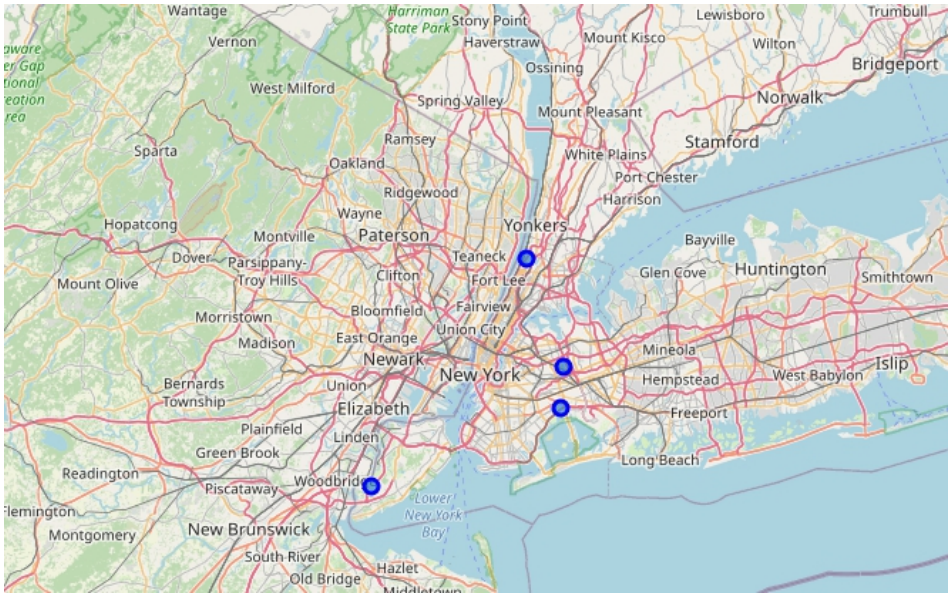

Figure 7. A map of the similar New York neighborhoods

5. Conclusion

The Parkwoods neighborhood is clustered with four other neighborhoods from New York in cluster 4. This means that these neighborhoods have some similar properties. So if a person is moving from Parkwoods to New York, these four neighborhoods would be a good option to consider settling in. Likewise this whole process can be applied to any other neighborhoods data and therefore a person who's moving need not go through a nightmare of physically visiting innumerable places to find the one that resembles their home.