A PLACE LIKE HOME

IBM COURSERA CAPSTONE PROJECT

By
Siddarth Nataraj

Problem Description

- This project emphasizes the need to compare neighborhoods and find similarity between them.
- This helps people who are moving from one place to another find a place to settle quickly and easily.
- Here I’ve taken one neighborhood from Toronto city and comapred it with all the neighborhoods in New York city to find out the ones that are similar.

Toronto Data

- The Toronto neighborhood data is obtained from a Wikipedia page and a CSV file.
- For convenience to identify the neighborhood in the later section of the problem, a city column has been added with a singular value “Toronto”.

	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

New York Data

- The New York Neighborhood data is obtained from a JSON file.
- The columns are similar to Toronto data and a city column is added with value “New York”.

	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

Methodology

- 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.
- The number of venues obtained for each neighborhood is determined.

Venues for each neighborhood

```
[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

Dataframe with top 10 venues in each neighborhood

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

K-Means clustering

- 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.
- The “North york” neighborhood is identified and from the cluster label column, it is found out that it belongs to cluster 4.

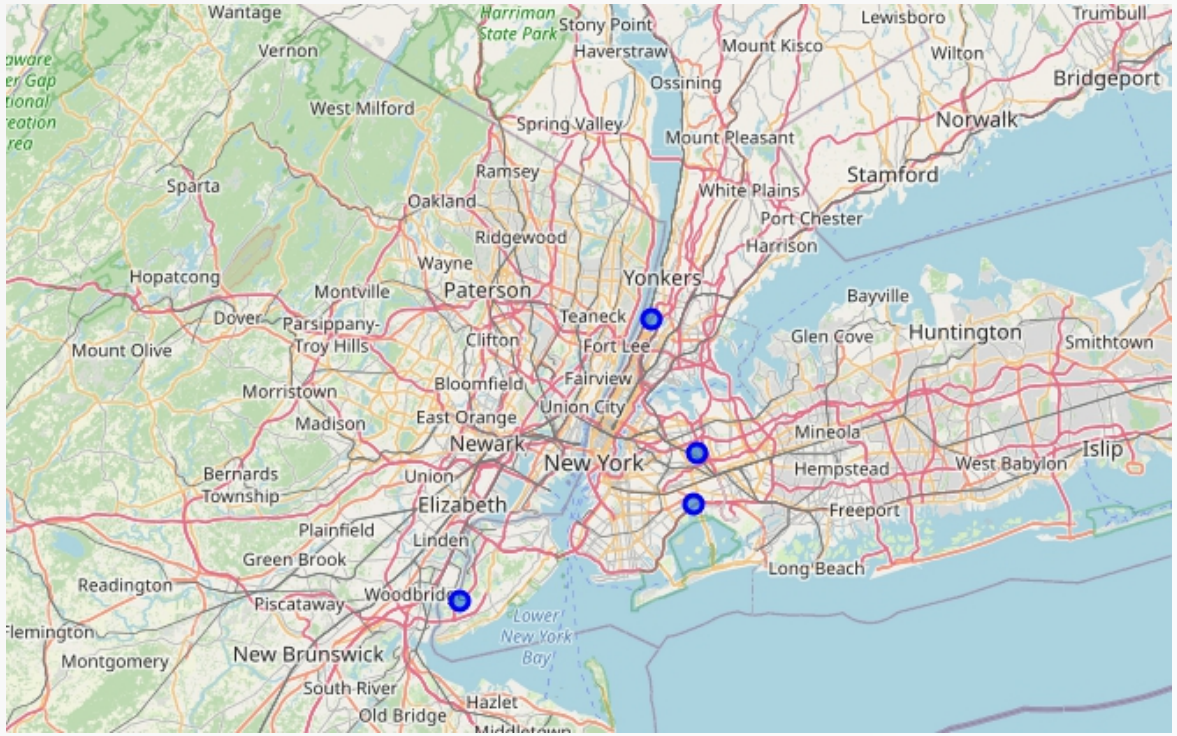
```
[50]: merged.loc[merged['City']=='Toronto']
```

[50]:

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

Visualization of the 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.



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.
- Likewise this whole process can be applied to any other neighborhoods data.
- 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.