The Battle of Neighborhoods

(IBM Applied Data Science Capstone Final Project Report)

Introduction

Study and analysis of data by various methods is critical for any activity in today's world. This project is to help a friend to identify a neighborhood location in Toronto city like his current location in New York city with shopping malls, parks and food courts. My friend can decide his location to relocate based on the outcome of my analysis. This is achieved with the help of Machine Language Algorithms.

Business Problem

In this project, I want to help my friend to find the best neighborhood in Toronto which is like his current place of residence with amenities like shopping malls, parks and food courts. The challenge is to find a suitable neighborhood which is close enough meeting his specific requirements.

Data

A list of neighborhoods in New York and Toronto is downloaded and their respective location. The data is extracted from the below sources

New York neighborhoods:
https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json

Toronto neighborhoods:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Transform the data into Pandas dataframe, use Geopy Python package to get the latitude and the longitude of all the neighborhoods of Toronto, use Folium Python library to map the neighborhoods, use Foursquare API to get information about venues around the neighborhoods. Look for a group of venues in walking distance of each neighborhood like shopping malls, parks and food courts.

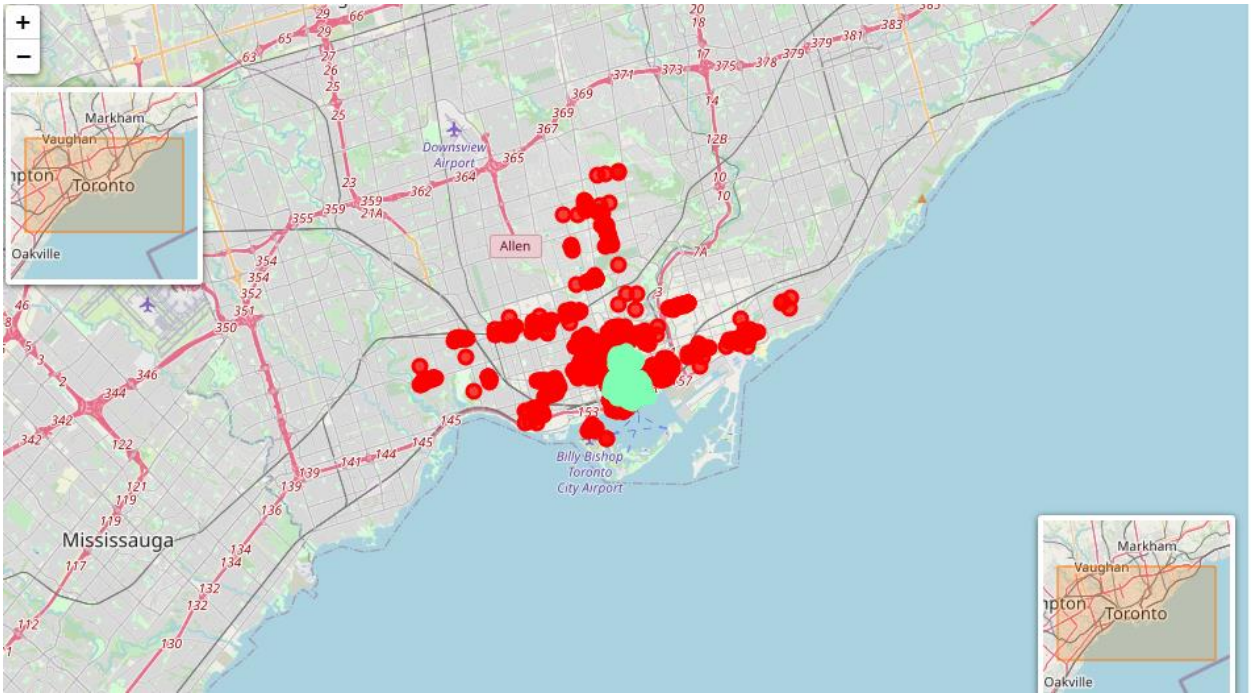
Methodology

For the purpose of doing unsupervised learning to find similarities between neighborhoods, clustering algorithm K-Means is used. K-Means clustering helps to classify the neighborhoods based on the nearby venues.To make the decision, the neighborhoods are sorted in each cluster.

	Neighborhood	Shopping Mall	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bath Beach	0.000000	0	40.599519	-73.998752	Bensonhurst Park	40.597065	-73.998340	Park
66	Mill Basin	0.000000	0	40.615974	-73.915154	Key Food	40.617269	-73.909776	Supermarket
66	Mill Basin	0.000000	0	40.615974	-73.915154	MTA B3, B100 (Ave U/East 66th St)	40.617206	-73.911299	Bus Station
...	...	...	...	...	...	...	...	...	...
24	Commerce Court, Victoria Hotel	0.010000	2	43.648198	-79.379817	Jump	43.648147	-79.378752	American Restaurant
95	Stn A PO Boxes	0.010417	2	43.646435	-79.374846	Joe Fresh	43.644285	-79.369771	Clothing Store
95	Stn A PO Boxes	0.010417	2	43.646435	-79.374846	Loblaws	43.645427	-79.369789	Grocery Store

4392 rows × 9 columns

Map the cluster neighborhoods



Results

The results of the clusters are as below

Cluster 1

	Shopping Mall	Venue	Venue Latitude	Venue Longitude	Venue Category
0	0.0	Bensonhurst Park	40.597065	-73.998340	Park
66	0.0	Key Food	40.617269	-73.909776	Supermarket
66	0.0	MTA B3, B100 (Ave U/East 66th St)	40.617206	-73.911299	Bus Station
...	...	...	...	...	...
33	0.0	Dumbo Archway	40.703056	-73.987985	Monument / Landmark
30	0.0	Kumo Sushi	40.640900	-73.965067	Japanese Restaurant
30	0.0	Cortelyou Gourmet Deli	40.641665	-73.962767	Deli / Bodega

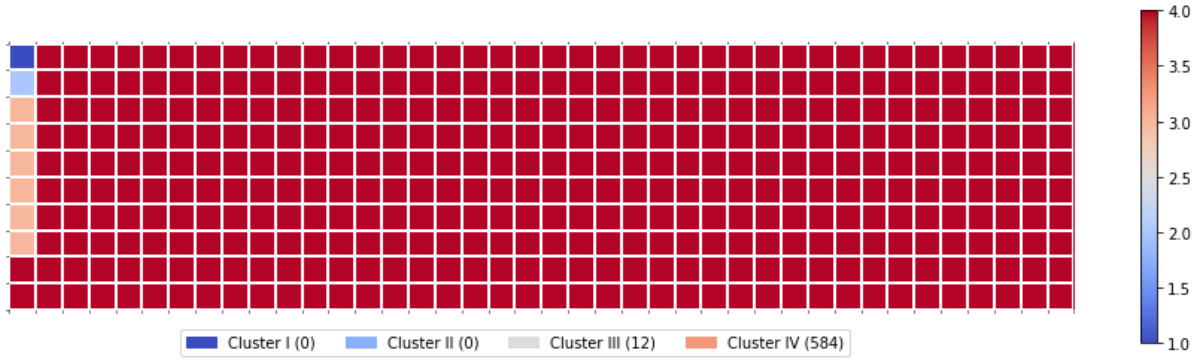
Cluster 2

	Shopping Mall	Venue	Venue Latitude	Venue Longitude	Venue Category
94	0.083333	MTA Bus - B82, B82 LTD, B83, BM2, BM5	40.646296	-73.879967	Bus Station
94	0.083333	Dunkin'	40.646776	-73.883176	Donut Shop
94	0.083333	CVS pharmacy	40.647316	-73.882990	Pharmacy
...	...	...	...	...	...
94	0.083333	Starrett City Shopping Center	40.649342	-73.884329	Shopping Mall
94	0.083333	Fresh Creek Natural Reserve	40.643372	-73.880271	River
94	0.083333	50% Half Cards	40.647151	-73.883584	Convenience Store

Cluster 3

	Shopping Mall	Venue	Venue Latitude	Venue Longitude	Venue Category
102	0.010000	Maple Leaf Square	43.642925	-79.380892	Plaza
102	0.010000	The Ritz-Carlton	43.645330	-79.387089	Hotel
102	0.010000	Boxcar Social Temperance	43.650557	-79.381956	Bar
...	...	...	...	...	...
24	0.010000	Jump	43.648147	-79.378752	American Restaurant
95	0.010417	Joe Fresh	43.644285	-79.369771	Clothing Store
95	0.010417	Loblaws	43.645427	-79.369789	Grocery Store

In the above image it is obvious that cluster algorithm is not segmenting the neighborhoods for location areas. It is possible to see which neighborhoods within Brooklyn, New York are more like the neighborhoods within Toronto. Those neighborhoods that are similar among them belong to the same cluster. Hence, they have the same color in the image above.



Discussion

This work is useful only for those who live in Brooklyn,New York or in the neighborhoods near the center of Toronto,Canada. That is because there is a limited amount of data we can request using de Foursquare API.

Conclusion

The K-Means clustering algorithm is used for finding similiraties between all the neighborhoods listed in the feature matrix. Results show that there are 2 major groups and 2 minor groups. In this work a segmentation between two different countries is done. This segementation involves the neighborhoods in Brooklyn, New York and the neighborhoods near to the center of Toronto. The data is downloaded and the venues around the neighborhoods is selected using the Foursquare API. One Hot Encoding is used for converting the categories of the venues into a feature matrix. Then, all venues are grouped by neighborhoods and at the same time the mean is calculated. Hence, the resulting features used are the frequency of occurrence from each category in a neighborhood.

The description of the clusters is as below

Cluster

I: Neighborhoods that have around parks, bus station and Restaurant places.

II: Neighborhoods that have around shopping mall, bus station and convenience store

III: Neighborhoods that have around Plaza, hotels, bar and stores

IV: Neighborhood that has nothing as per requirement

My friend can now decide and move from Brooklyn, New York to Toronto, Canada using this system to get a notion or idea about what is the best suitable place for him.