Capstone Project: Finding a Better Place in Toronto

Yonas Berhe
October 31, 2020

Introduction:

The purpose of this Project is to help people in exploring better facilities in a new neighborhood. It will help people making smart and efficient decision on selecting great neighborhood out of numbers of other neighborhoods Toronto.

Lots of people are migrating to various states of Canada and needed lots of research for good housing prices and reupdated schools for their children. This project is for those people who are looking for better neighborhoods. For ease of accessing to Cafe, School, Super market, medical shops, grocery shops, mall, theatre, hospital, like minded people, etc.

This Project aim to create an analysis of features for a people migrating to Toronto to search a best neighborhood as a comparative analysis between neighborhoods. It will help people to get awareness of the area and neighborhood before moving to a new city, state, country or place for their work or to start a new fresh life.

Problem
The major purpose of this project, is to suggest a better neighborhood in a new city for the person who are shiffting there. Social presence in society in terms of like minded people. Connectivity to the airport, bus stand, city center, markets and other daily needs things nearby. This machine learning approach in searching for a new neighborhood will make it more effect and efficient.

Foursquare API
This project would use Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

Machine learning Approach:
To compare the similarities neighborhoods, we decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods in a big city like New York and Toronto. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm as well as KNN Classification to classify the new location given by user

Data Description

Data Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare API Data

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meter.
The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

- Neighborhood
- Neighborhood Latitude
- Neighborhood Longitude
- Venue
- Name of the venue e.g. the name of a store or restaurant
- Venue Latitude
- Venue Longitude
- Venue Category

Data Acquisition and Cleaning

Data was scraped from Wikipedia using the Beautiful soap library. This data frame was grouped by postal code. The borough with no postal code was removed and the column with empty neighborhoods were dropped. So finally the data frame has 3 columns as show below.

	Postal_Code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park / Harbourfront
3	M6A	North York	Lawrence Manor / Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government

```
Shape :  (103, 3)
```

The new data frame was created by splitting all the neighborhoods in the previous data frame and were split and appended. Longitude and latitude columns were concatenated and the values were obtained using geopy library. Neighborhoods without location data were dropped and as well as redundant values were dropped. The table now contains the following.

	Postal_Code	Borough	Neighborhood
0	M1S	Scarborough	Agincourt
1	M8W	Etobicoke	Alderwood, Long Branch
2	M3H	North York	Bathurst Manor, Wilson Heights, Downsview North
3	M2K	North York	Bayview Village
4	M5M	North York	Bedford Park, Lawrence Manor East

```
(99, 3)
```

The third data frame was created storing all the venues including their latitude and longitude and their category obtained from the foursquare API. The rows and with neighborhood category was dropped due to redundancy. The new table looks as follows

	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Venue	VLatitude	VLongitude	Category
0	M1S	Scarborough	Agincourt	43.785353	-79.278549	One2 Snacks	43.787048	-79.276658	Asian Restaurant
1	M1S	Scarborough	Agincourt	43.785353	-79.278549	Tim Hortons	43.785637	-79.279215	Coffee Shop
2	M1S	Scarborough	Agincourt	43.785353	-79.278549	In Cheon House Korean & Japanese Restaurant 인천관	43.786468	-79.275693	Korean Restaurant
3	M1S	Scarborough	Agincourt	43.785353	-79.278549	Yummy Cantonese Restaurant 老西関腸粉	43.787568	-79.269585	Cantonese Restaurant
4	M1S	Scarborough	Agincourt	43.785353	-79.278549	Beef Noodle Restaurant 老李牛肉麵	43.785937	-79.276031	Chinese Restaurant

```
(2692, 9)
Number of neighborhoods : 49
Number of venue categories : 290
```

A new data frame was created with columns from the categorical variables of column category and was also grouped by neighborhood names by taking the mean along the column axis. The table looks as follows.

	Neighborhood	Accessories Store	Afghan Restaurant	African Restaurant	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	Argentinian Restaurant	...	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Wine Bar	Wings Joint	Women's Store	Xi Rest
0	Agincourt	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.027778	0.0	0.0	0.0	0.00	
1	Alderwood, Long Branch	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.00	
2	Bayview Village	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.02	
3	Berczy Park	0.0	0.0	0.0	0.03	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.00	
4	Cedarbrae	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.025000	0.0	0.0	0.0	0.00	

5 rows × 291 columns

```
(49, 291)
```

The data obtained from the user given neighborhood and converted to categorical variables and grouped together. I also dropped those added in the previous table.

	Neighborhood	Latitude	Longitude	Venue	VLatitude	VLongitude	Category
0	Connaught Place, New Delhi	28.631383	77.219792	Connaught Place | कनॉट प्लेस (Connaught Place)	28.632731	77.220018	Plaza
1	Connaught Place, New Delhi	28.631383	77.219792	Starbucks	28.632011	77.217731	Coffee Shop
2	Connaught Place, New Delhi	28.631383	77.219792	Jain Chawal Wale	28.630052	77.217649	Food Truck
3	Connaught Place, New Delhi	28.631383	77.219792	Rajdhani Thali	28.629999	77.220401	Indian Restaurant
4	Connaught Place, New Delhi	28.631383	77.219792	Fabindia	28.632012	77.217729	Clothing Store

```
(80, 7)
```

	Accessories Store	Afghan Restaurant	African Restaurant	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	Argentinian Restaurant	Art Gallery	...	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Wine Bar	Wings Joint	Women's Store	Xinjiang Restaurant
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0.0125	0

1 rows × 290 columns

```
(1, 290)
```

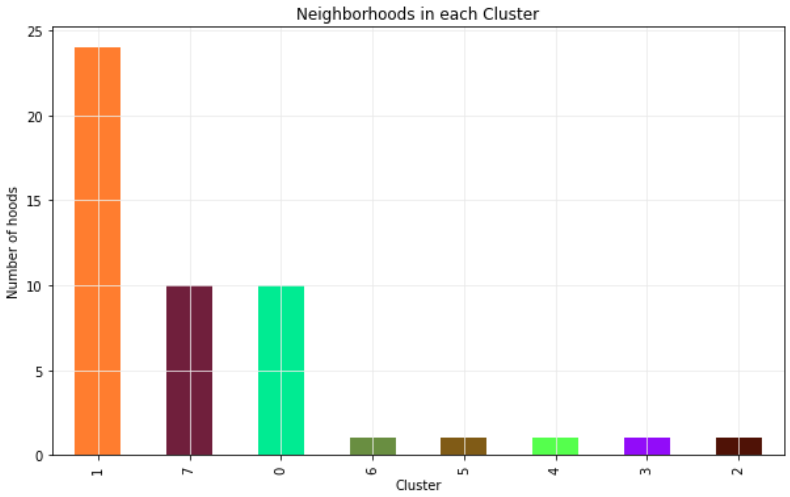
The columns in from above table was used to create the mode. And total of 290 features were selected.

Cluster assignment of neighborhoods

The clusters determination was using the K-means clustering. The neighborhoods were divided in 7 clusters.

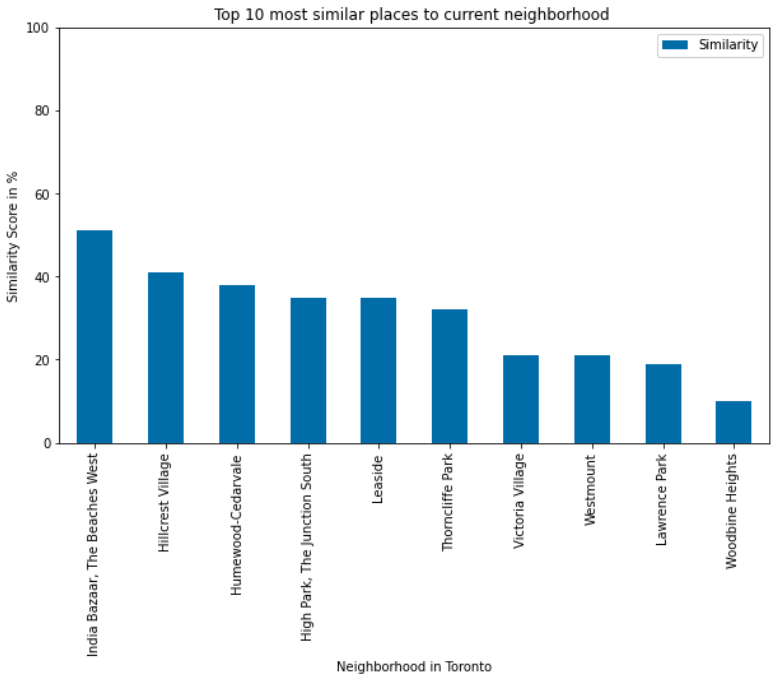
	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Cluster
0	M1S	Scarborough	Agincourt	43.785353	-79.278549	4
1	M8W	Etobicoke	Alderwood, Long Branch	43.601717	-79.545232	0
2	M2K	North York	Bayview Village	43.769197	-79.376662	1
3	M5E	Downtown Toronto	Berczy Park	43.647984	-79.375396	1
4	M1H	Scarborough	Cedarbrae	43.756467	-79.226692	0

```
(49, 6)
```

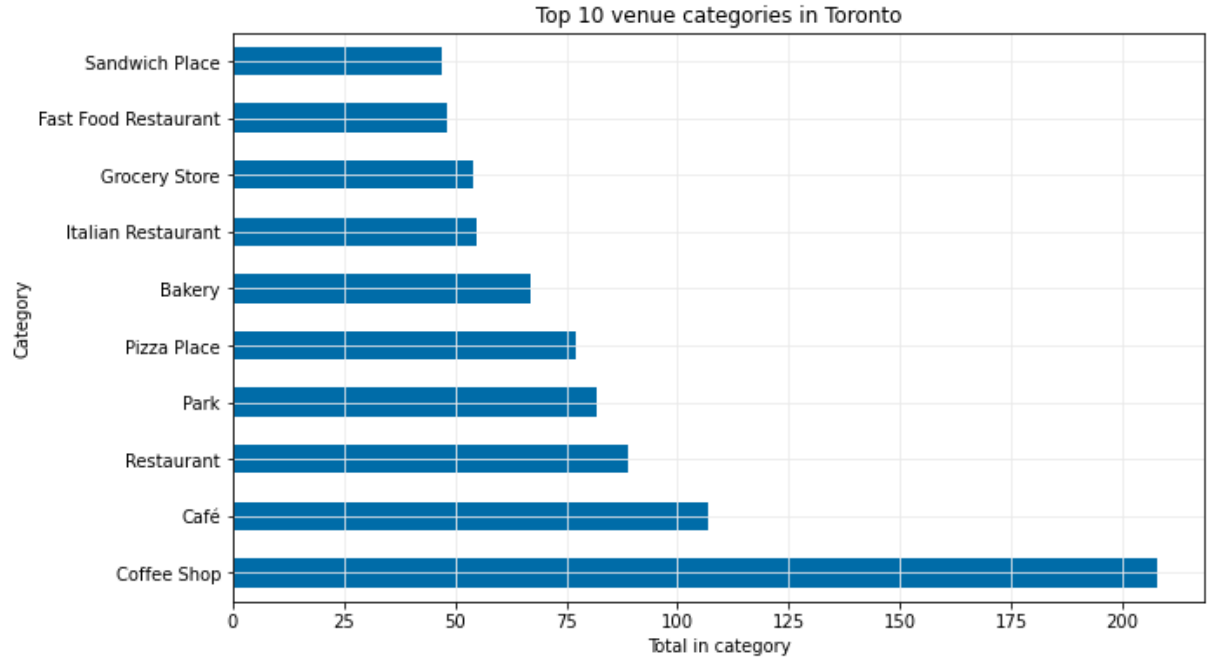


Classifying current neighborhood to a cluster

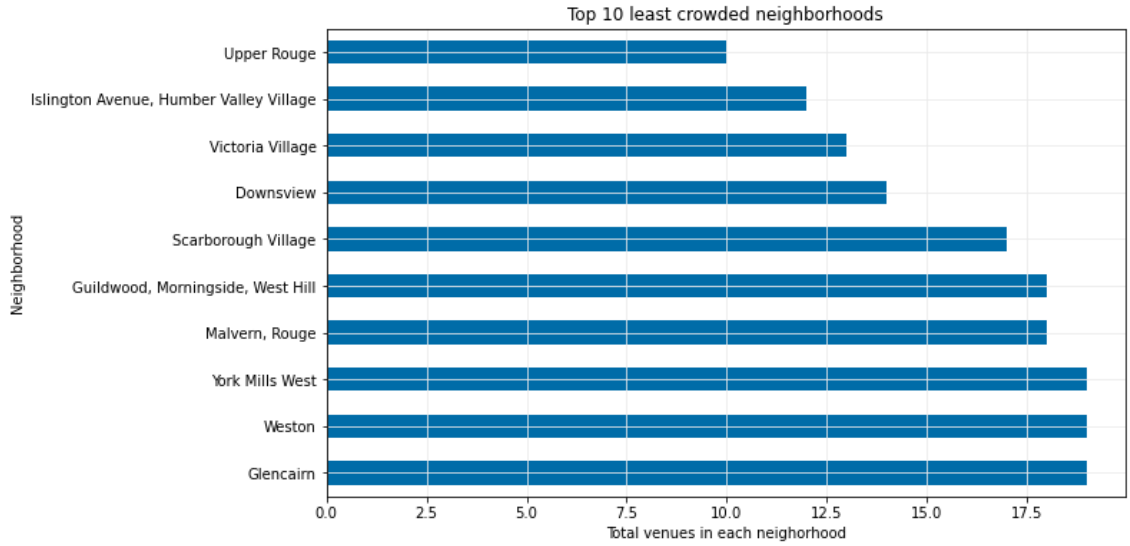
	Neighborhood	Distance	Similarity	Latitude	Longitude
0	India Bazaar, The Beaches West	0.195780	51.0	43.669189	-79.317248
1	Hillcrest Village	0.210850	41.0	43.681695	-79.425712
2	Humewood-Cedarvale	0.219908	38.0	43.688322	-79.428080
3	High Park, The Junction South	0.225607	35.0	43.653867	-79.466864
4	Leaside	0.234854	35.0	43.704798	-79.368090
5	Thorncliffe Park	0.259389	32.0	43.704553	-79.345407
6	Victoria Village	0.309330	21.0	43.732658	-79.311189
7	Westmount	0.269337	21.0	43.693640	-79.521043
8	Lawrence Park	0.253394	19.0	43.729199	-79.403252
9	Woodbine Heights	0.294006	10.0	43.699920	-79.319279



Top Venue Categories



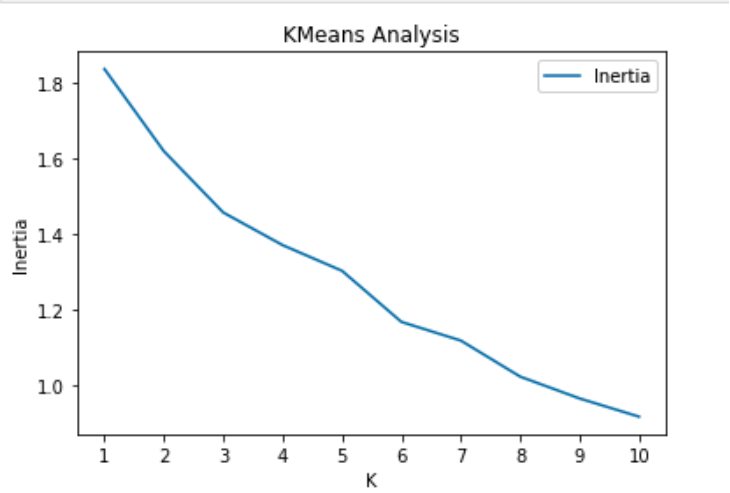
The least crowded neighborhoods



Map plot of all 7 Clusters

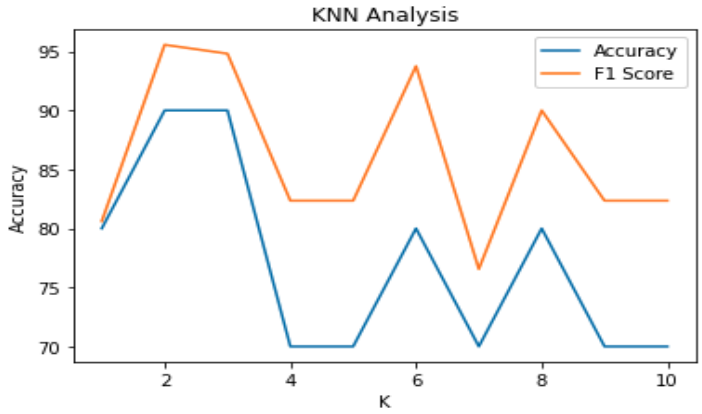
K-Means Clustering

I used a clustering model (K-Means clustering) to group the similar neighborhoods in Toronto together. The neighborhoods were divided into 7 clusters. The best K was found using elbow technique. The clustering model was made using 290 features.



K-Nearest neighbors Classification

Some of the classification accuracy matrices to measure user assigned new data



	K	Accuracy	F1 Score
0	1	80.0	80.62
1	2	90.0	95.56
2	3	90.0	94.81
3	4	70.0	82.35
4	5	70.0	82.35
5	6	80.0	93.75
6	7	70.0	76.56
7	8	80.0	90.00
8	9	70.0	82.35
9	10	70.0	82.35

```
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         2
           1       0.88      1.00      0.93         7
           7       0.00      0.00      0.00         1

   micro avg       0.90      0.90      0.90        10
   macro avg       0.62      0.67      0.64        10
weighted avg       0.81      0.90      0.85        10

The cross validation score is :  0.758
```

Conclusion

This study analyzed neighborhood in the city of Toronto. Using the k-means clustering I was able to find the neighborhoods similar to each other. The most and least common venues are crowded. KNN is used to classify and predict which cluster a new neighborhood will belong considering in the venue present in the neighborhood. The model has an accuracy of 64% and 75% with an F1 score. The number seems fitting for the small sample sized used and would likely predict new neighborhoods similarly in the future.