Capstone Presentation:
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.

- The major purpose of this project, is to suggest a better neighborhood in a new city for the person who are shifting 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.

Data Management

- The neighborhood names, boroughs and postal codes for Toronto were scraped from the Wikipedia page.
- The latitude and longitude coordinates of the neighborhoods was acquired using geopy library
- The venue details, categories and latitude and longitude were acquired using foursquare API. The neighborhoods were scanned in a 1000 meter radius for venues. The API only return top 100 venues for each neighborhood. After cleaning 2692 entries were obtained

Main data frame

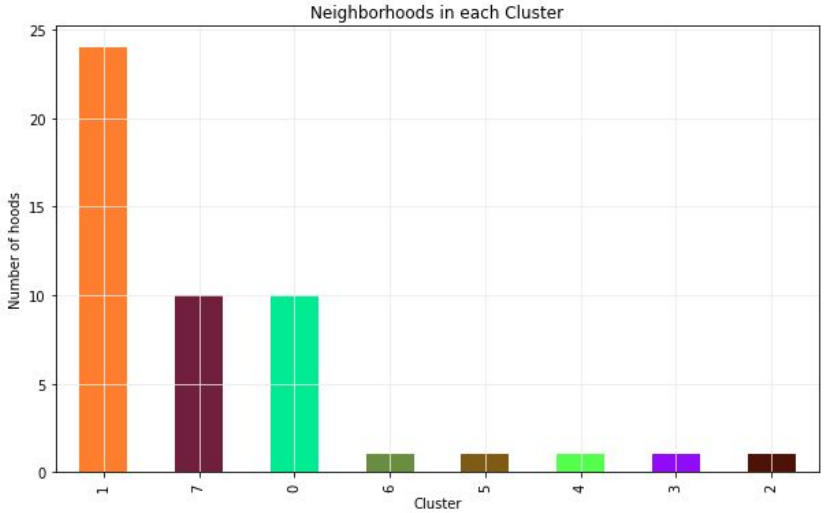
	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Venue	VLatitude	VLongitude	Category
0	M1S	Scarborough	Agincourt	43.785353	-79.278549	One2 Snacks	43.787048	-79.276658	Asian Restaurant
1	M1S	Scarborough	Agincourt	43.785353	-79.278549	Tim Hortons	43.785637	-79.279215	Coffee Shop
2	M1S	Scarborough	Agincourt	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
```

Assigning cluster to neighborhood

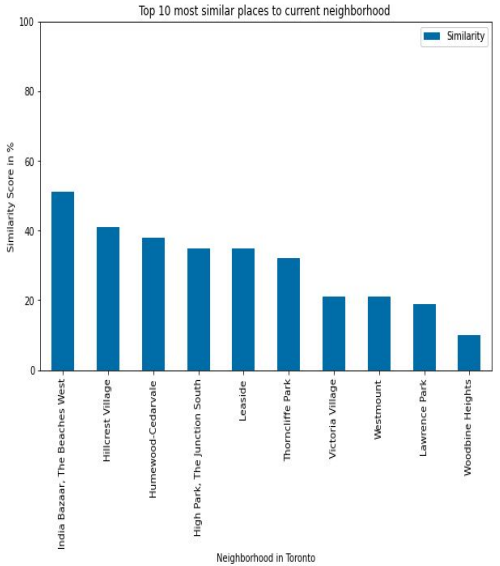
- The cluster to which each neighborhood belonged to was determined using K – Means Clustering.

- The neighborhoods were divided into 7 clusters.



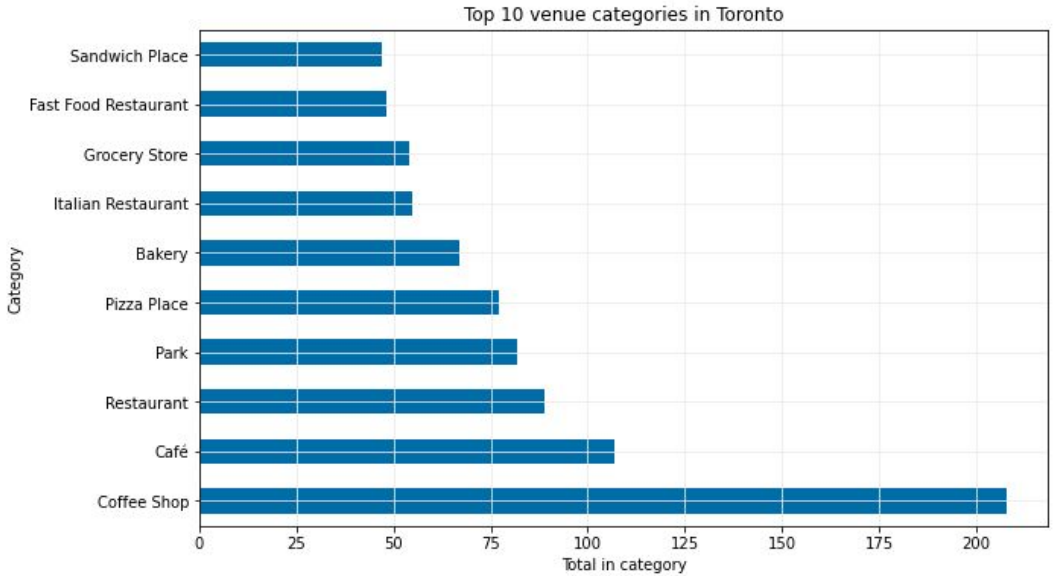
Classify new Neighborhoods

- A similarity score of each neighborhood in the cluster with the current neighborhood was obtained, which is basically the correlation between them.
- The Euclidean distance between the neighborhoods was also obtained.

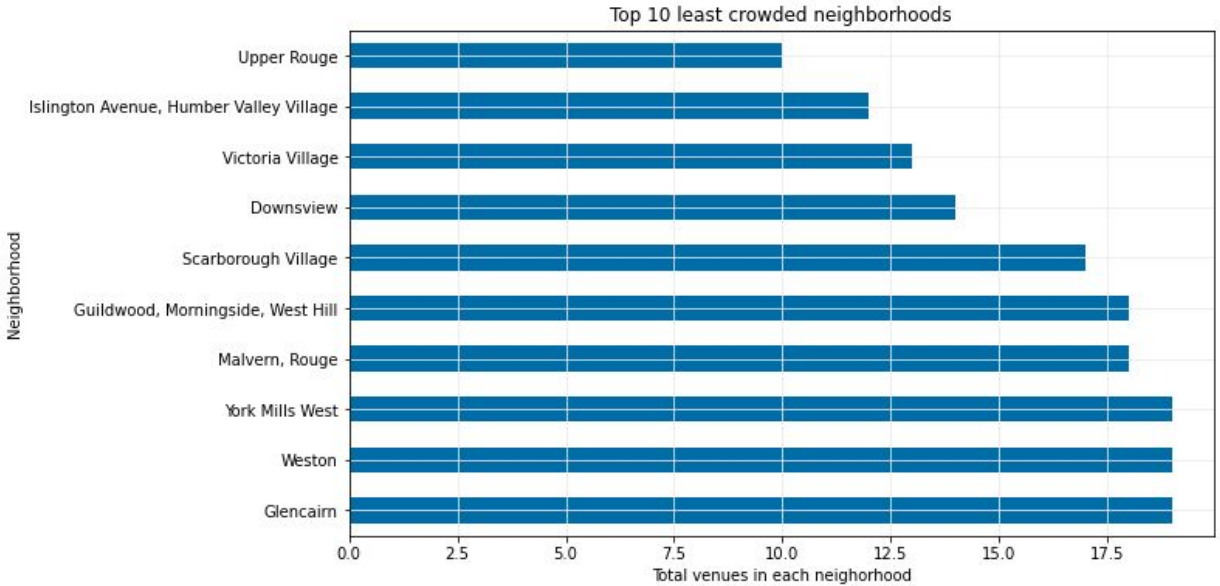


	Neighborhood	Distance	Similarity	Latitude	Longitude
0	India Bazaar, The Beaches West	0.195780	51.0	43.669189	-79.317248
1	Hillcrest Village	0.210850	41.0	43.681695	-79.425712
2	Humewood-Cedarvale	0.219908	38.0	43.688322	-79.428080
3	High Park, The Junction South	0.225607	35.0	43.653867	-79.466864
4	Leaside	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 category



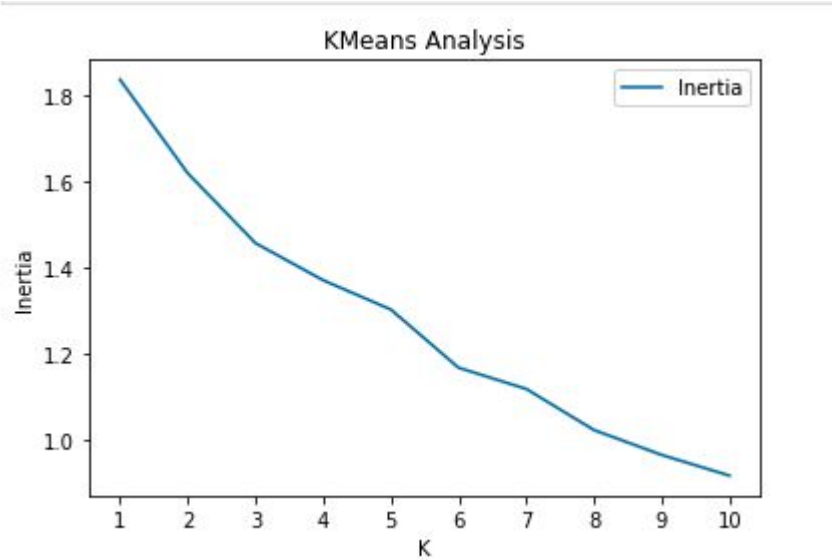
The least Crowded neighborhoods



Map of all clusters

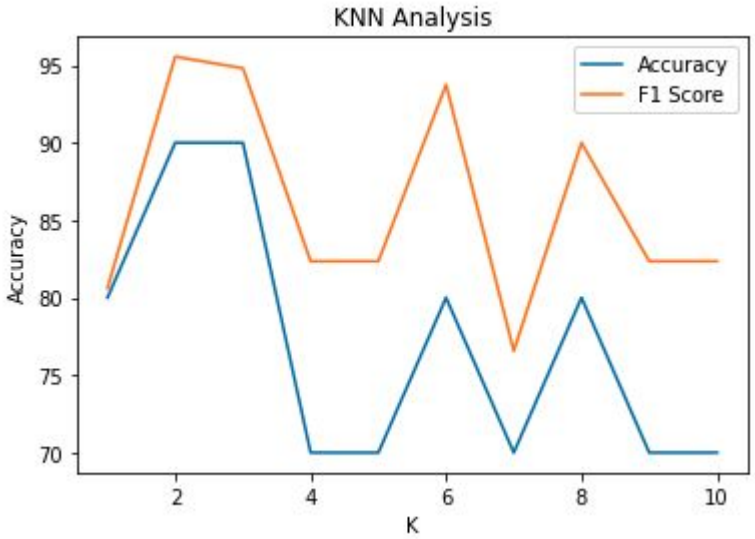
K-means clustering

- K-Means clustering was used to group the similar neighborhoods in Toronto together.
- The neighborhoods were divided into 8 clusters.
- The best K was found using elbow technique.
- The clustering model was made using 290 features.



KNN classification

- KNN classification was used to classify the current neighborhood (given by user) to a cluster.
- The model was evaluated using different values of K and the best K was found to be 3.



	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

KNN classification evaluation report

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

Finally

- In this study I analyzed the neighborhoods present in the city of Toronto, Ontario, Canada. I found out the venues present in each neighborhood and clustered the similar neighborhoods together using K-Means Clustering.
- I found the most common venues in the city of Toronto, which neighborhoods are the most crowded, which are the least crowded, which venues are the most popular.
- I used KNN Classification to predict which cluster a new neighborhood will belong to considering the types of venues present in the neighborhood.
- The classification model had an accuracy of around 64 % using Jacquard similarity score, and 75% using F1 score, which is pretty good considering the small sample size. Thus this model can be used to predict a perfect new neighborhood similar to a given neighborhood.