Predicting the Perfect Neighborhood

Anurag Gupta

April 09, 2020

1. Introduction

1.1 Background

Often times, people come across situations which forces them to move to a different city or even a different country, or maybe they do so of their own free will. But in any case it is important that the neighborhood we choose to live in is as close as possible to the neighborhood we want it to be. i.e. our new neighborhood should be similar to our old one as much as possible. It allows us to adapt more easily as well as the places we like will be in close proximity.

1.2 Problem

If we try to do it manually it will require us to check each and every neighborhood, find whether hundreds of venues such as restaurants, exist in close proximity or not. What is the distance from the schools, hospitals and many more factors. Doing this will require a lot of man-hours, and it still might not be effective enough. Thus, using a machine learning algorithm is a much better option.

1.3 Interest

Obviously the real estate agencies and the customers who are looking to switch homes will be interested in this project. The user will provide the neighborhood location where he/she used to live, and the model will predict which neighborhoods are the most similar to their last neighborhood along with a Similarity Score. This model can be used by real estate agencies to provide customers with the best housing options in the city of Toronto.

2. Managing Data

2.1 Data Sources

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, and 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 returned top 100 venues for each neighborhood.

Wikipedia page: link

2.2 Data Acquisition and Cleaning

DF1 : Data from Wikipedia page was scraped using BeautifulSoup library. The main table from the page was extracted and the data from the table was added to a pandas dataframe. Note that this dataframe was grouped by postal codes. The postal codes whose Borough were 'Not Assigned' were removed from the dataframe and those rows which had empty 'Neighborhood' column were also dropped. So finally this dataframe had 3 columns ('Postal Code, Borough, Neighborhoods) and 103 rows/entries.

	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)
```

DF2: Then a new dataframe was created and all the neighborhoods in the previous dataframe were split and appended to this dataframe. This dataframe was grouped by neighborhoods and had 200 rows/entries. Logitude and Latitude Columns were concatenated to this dataframe and the values were obtained using geopy library. Those neighborhoods whose location coordinates could not be obtained and those whose latitude and longitude values were redundant were dropped from the dataframe. This table now contained 158 rows/entries.

	Postal_Code	Borough	Neighborhood	Latitude	Longitude
0	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498
1	M1S	Scarborough	Agincourt	43.785353	-79.278549
2	M1V	Scarborough	Agincourt North	43.808038	-79.266439
3	M9V	Etobicoke	Albion Gardens	43.741665	-79.584543
4	M8W	Etobicoke	Alderwood	43.601717	-79.545232

```
Shap (158, 5)
```

DF3: Now a new table was created which stored all the venues, their latitude and longitude and their category obtained from the foursquare API. One of the categories was neighborhood itself so the rows with category as 'Neighborhood' were dropped. The table had a total of 7979 entries/rows and 9 columns. Neighborhood with no venues were dropped from DF2.

	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Venue	VLatitude	VLongitude	Category
0	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	Richmond Station	43.651569	-79.379266	American Restaurant
1	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	Dineen Coffee	43.650497	-79.378765	Café
2	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	GoodLife Fitness Toronto 137 Yonge Street	43.651242	-79.378068	Gym
3	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	Beerbistro	43.649419	-79.377237	Gastropub
4	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	Pilot Coffee Roasters	43.648835	-79.380936	Coffee Shop

```
(7979, 9)
Number of neighborhoods : 158
Number of venue categories : 355
```

DF4: A new dataframe was created, whose columns were the categorical variables of column 'Category'. The dataframe was grouped by the neighborhood names taking the mean along the column axis. This dataframe would serve as the basis for model creation. The dataframe finally had 158 entries and 356 columns.

	Neighborhood	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	...
0	Adelaide	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	Agincourt North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
3	Albion Gardens	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	Alderwood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

5 rows × 356 columns

```
(158, 356)
```

DF5: The dataframe obtained from user given neighborhood had 79 venues which were converted to categorical variables and grouped together, the unknown columns/categories were dropped and those which were in DF4 were added. Finally it had 1 row and 355 columns.

	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	Wenger's	28.633412	77.218292	Bakery
2	Connaught Place, New Delhi	28.631383	77.219792	Starbucks	28.632011	77.217731	Coffee Shop
3	Connaught Place, New Delhi	28.631383	77.219792	Farzi Cafe	28.632581	77.221125	Molecular Gastronomy Restaurant
4	Connaught Place, New Delhi	28.631383	77.219792	Rajdhani Thali	28.629999	77.220401	Indian Restaurant

```
(240, 7)
```

	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...
0	0	0	0	0	0	0	0	0	0	0	...

1 rows × 355 columns

```
(1, 355)
```

2.3 Feature Selection

The columns of table DF4 were used as features to create the model. Thus a total of 352 features were selected (the neighborhood name from DF4 was not used).

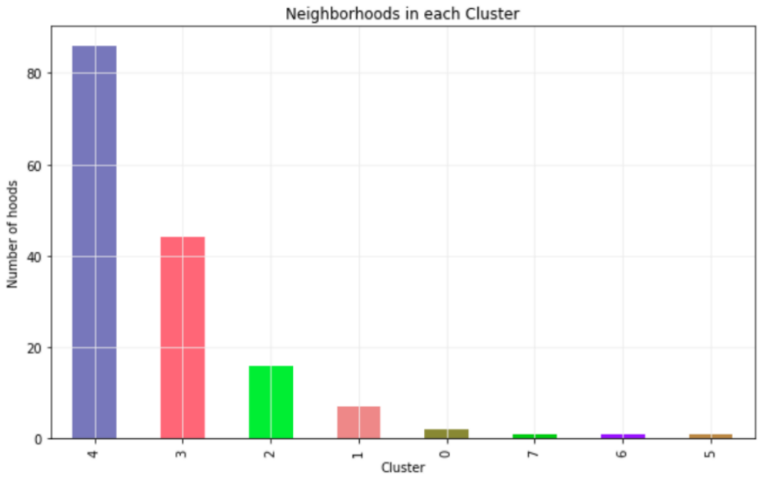
3. Exploratory Data Analysis

3.1 Assigning Cluster to neighborhoods.

The cluster to which each neighborhood belonged to was determined using K – Means Clustering. The neighborhoods were divided into 8 clusters. The mean number of neighborhood per cluster is 19.7 and the standard deviation was 30.5 indication the distribution of neighborhoods into cluster is not uniform. With the top 3 clusters having 86, 44, 16 neighborhoods out of 158. A column with assigned clusters was added to DF4.

	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Cluster
0	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	4
1	M1S	Scarborough	Agincourt	43.785353	-79.278549	1
2	M1V	Scarborough	Agincourt North	43.808038	-79.266439	3
3	M9V	Etobicoke	Albion Gardens	43.741665	-79.584543	3
4	M8W	Etobicoke	Alderwood	43.601717	-79.545232	3

```
(158, 6)
```

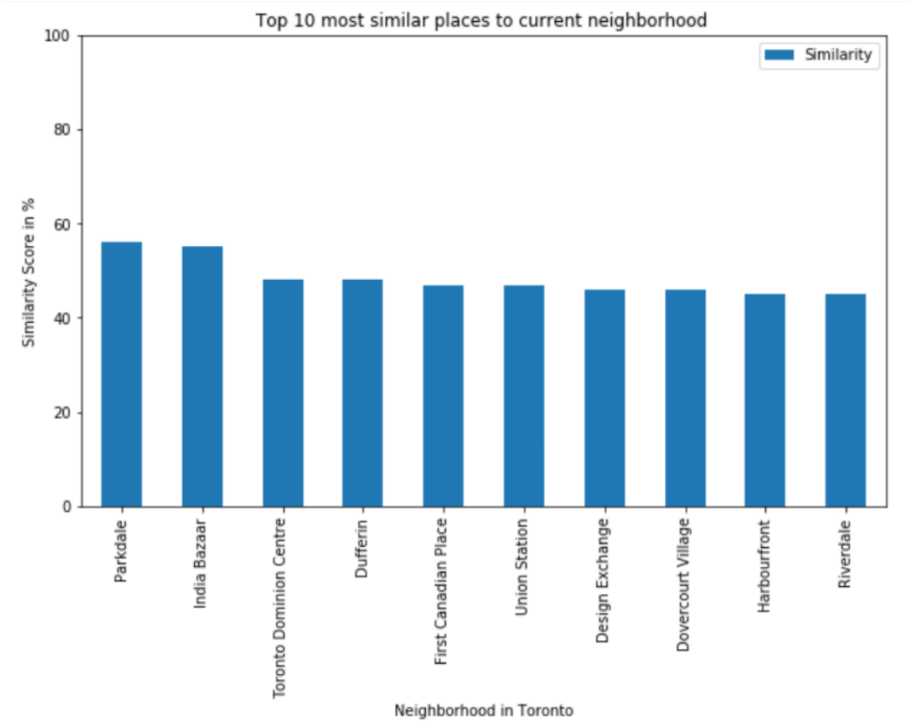


3.2 Classifying current neighborhood to a cluster

The table DF4 served as the basis for training the classification model. The current neighborhood was found to belong to cluster 6. With the most similar neighborhood having a similarity score of 56 %. Cluster 6 contains 16 neighborhoods in it.

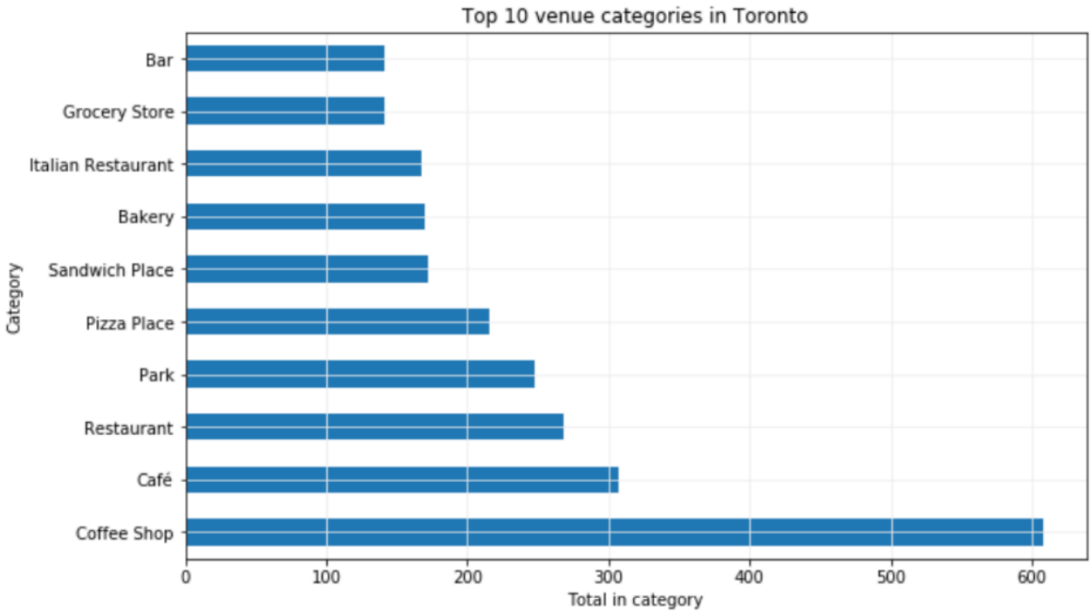
```
This belongs to cluster :  4
```

	Neighborhood	Distance	Similarity
0	Parkdale	0.182004	56.0
1	India Bazaar	0.195937	55.0
2	Toronto Dominion Centre	0.198032	48.0
3	Dufferin	0.209851	48.0
4	First Canadian Place	0.202616	47.0
5	Union Station	0.203119	47.0
6	Design Exchange	0.205274	46.0
7	Dovercourt Village	0.207214	46.0
8	Harbourfront	0.207449	45.0
9	Riverdale	0.211510	45.0



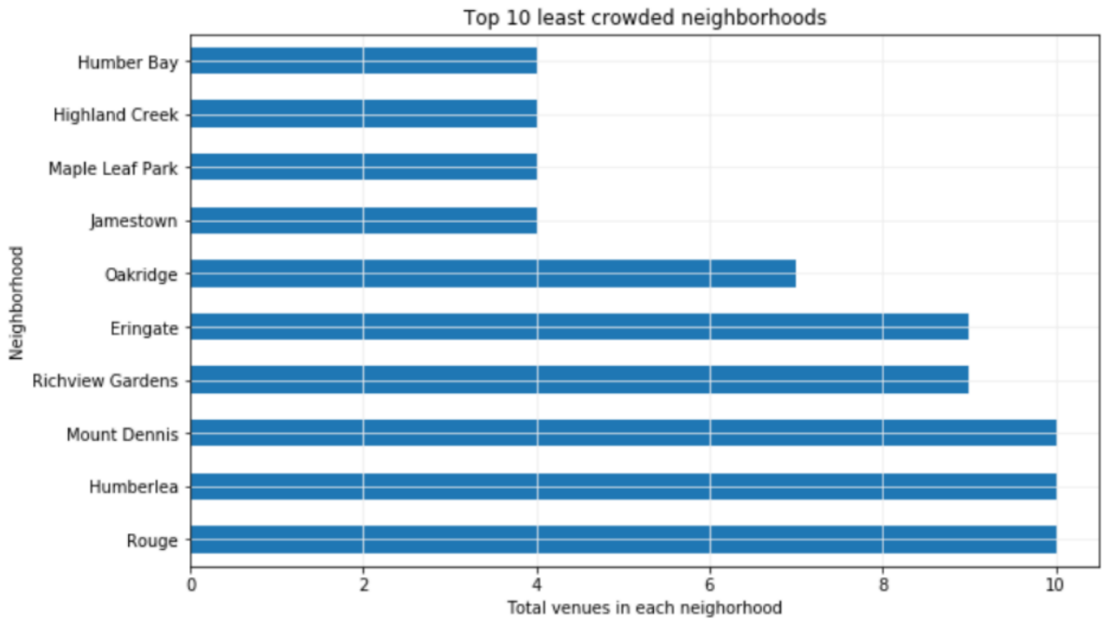
3.3 Top Venue Categories

The top venue categories in Toronto are coffee shops, cafes, restaurants and parks.

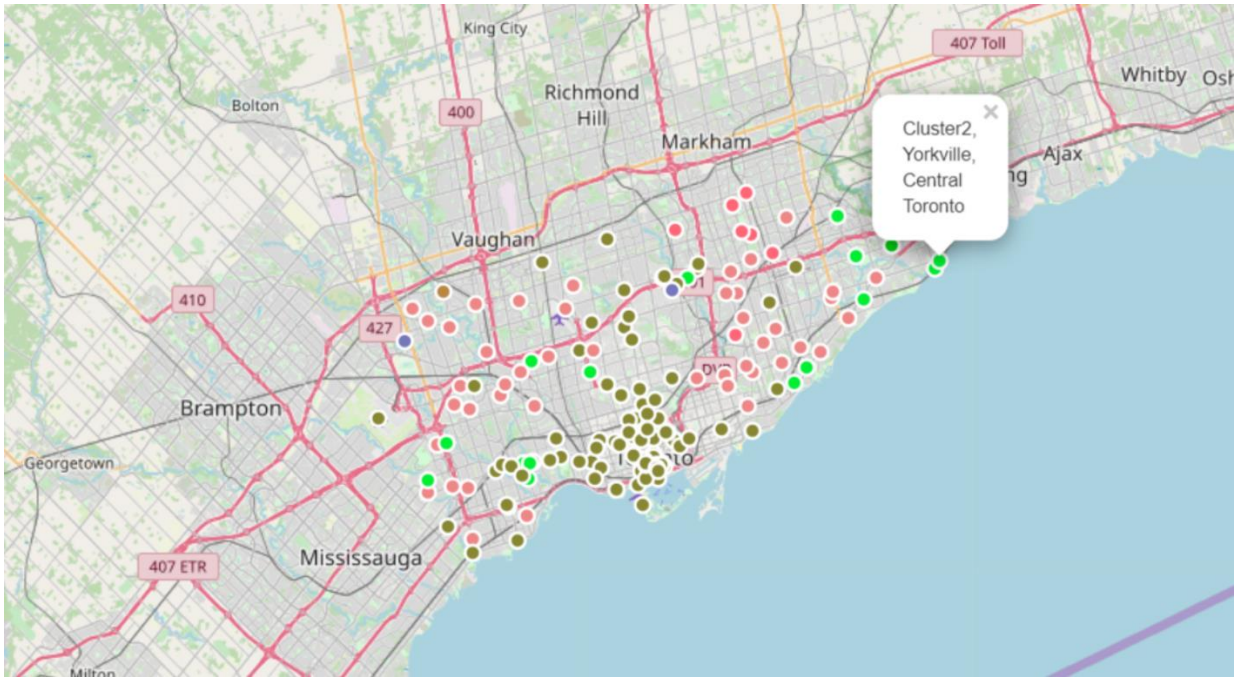


3.4 The least crowded neighborhoods

The least crowded neighborhoods in Toronto are Jamestown, Maple Leaf Park, Humber Summit, Highland Creek each having 4 venues in it.



3.5 Map plot of all 8 Clusters

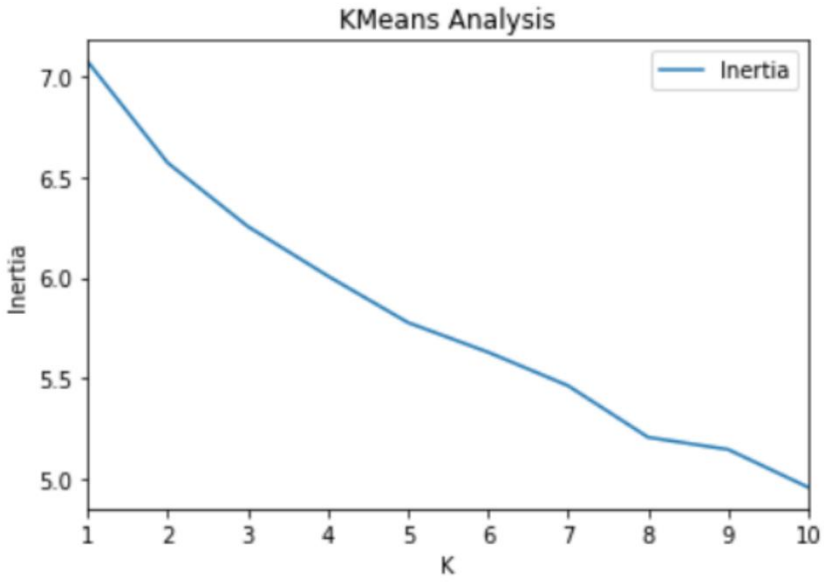


4. Models

4.2 K-Means Clustering

I used a clustering model (K-Means clustering) 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 352 features.

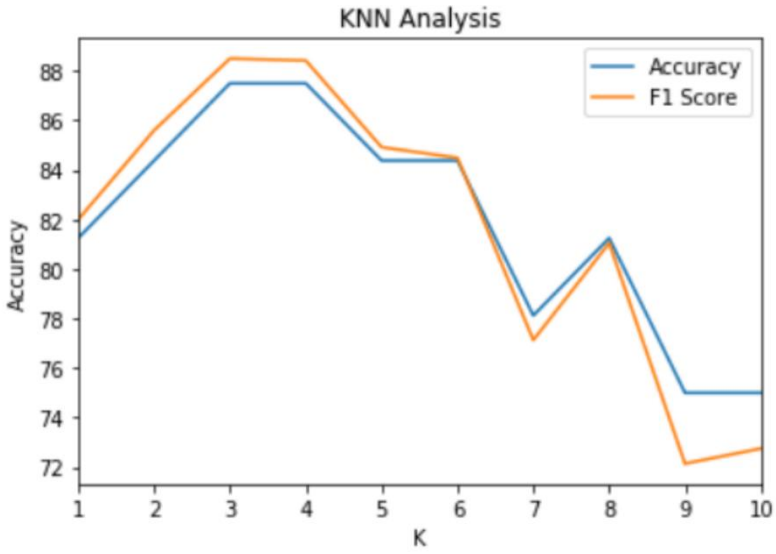
K-Means clustering work by assigning K number of centroids, and calculating the Euclidean distance of a data point to the nearest centroid. Thus the centroid shifts to the first data point. Then for a next data point the nearest centroid is again found and the centroid shifts to the mean of the current and previous data points. This process is repeated for each data point till all data points have been assigned a cluster.



4.2 K-Nearest neighbors Classification

I used a classification model (KNN classification) to classify the current neighborhood (given by user) to a cluster. KNN Classification works by assigning the new data the same group as its K nearest neighbors. The dataset was split into testing and training dataset (test size = 0.2). The model was evaluated using different values of K and the best K was found to be 3. The Jacquard similarity index for the model (K = 3) was 87.5 and the F1 score was 88.4. The cross validation score for the model was found to be 0.78

	K	Accuracy	F1 Score
0	1	81.25	82.00
1	2	84.38	85.60
2	3	87.50	88.50
3	4	87.50	88.42
4	5	84.38	84.92
5	6	84.38	84.49
6	7	78.12	77.13
7	8	81.25	81.05
8	9	75.00	72.14
9	10	75.00	72.75




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

           1       1.00      1.00      1.00         1
           2       1.00      0.67      0.80         6
           3       1.00      0.80      0.89         5
           4       0.83      1.00      0.90        19
           5       0.00      0.00      0.00         1

   micro avg       0.88      0.88      0.88        32
   macro avg       0.77      0.69      0.72        32
weighted avg       0.87      0.88      0.86        32

The cross validation score is :  0.78
```

5. Conclusion

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. 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 72 % using Jacquard similarity score, and 78% 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.