Capstone Project - The Battle of Neighborhoods

Coursera

Introduction

- This project is a kind of summary of the learned material covered in IBM course of data science. In this work, we used the algorithm – K-means clustering - applied to the neighborhoods in Toronto.

- The objective of this Capstone is to learn about segmenting and k-means clustering, to understand how to use the Github platform, to learn the Foursquare API and to try to show the results of studying by publishing the final project on Github. The main goal is to understand the practical side of the application of techniques and algorithms of data science in real life tasks.

Data

- The data has been collected from Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M .
- This html table was converted to Pandas DataFrame for cleaning and preprocessing. We have dropped the rows where Borough is 'Not assigned' and combined the neighbourhoods with same Postal code.
- Then we have imported the csv file conatining the latitudes and longitudes for various neighbourhoods in Canada and merged two tables.

- We will analyze all the rows from the data frame which contains Toronto in their Borough.

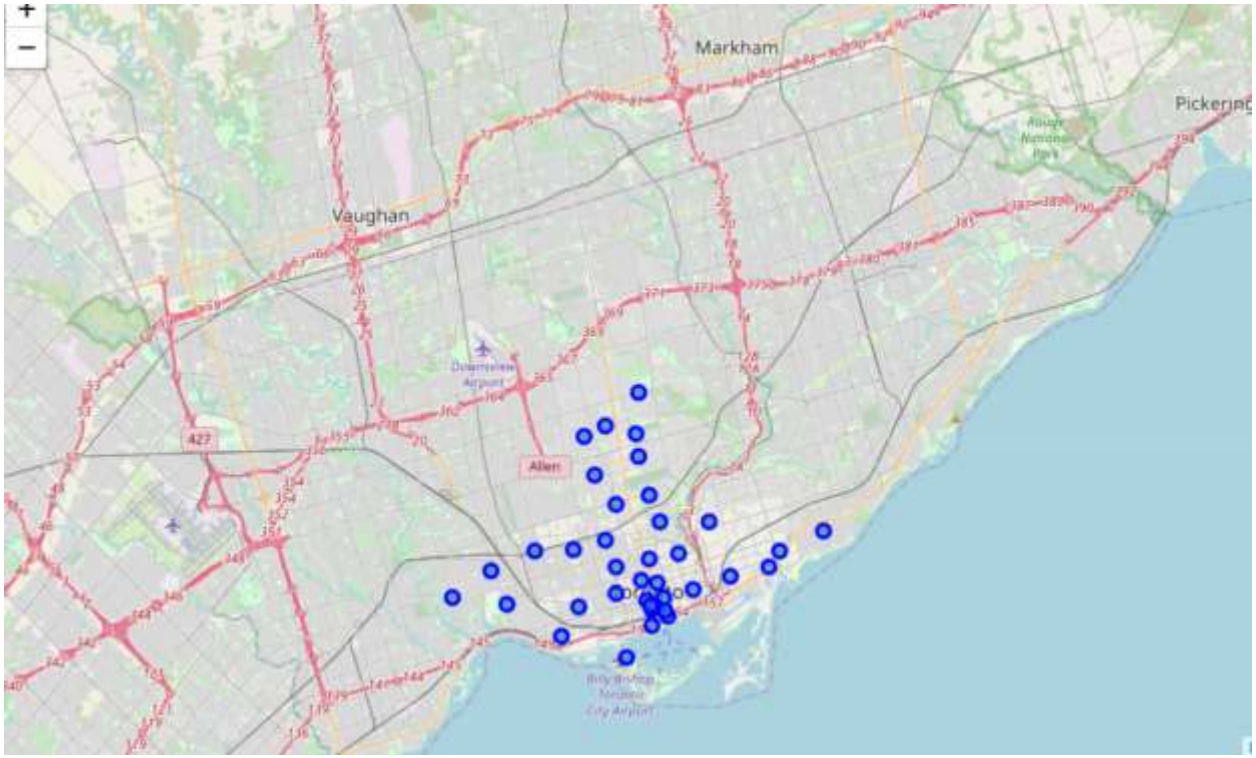
Segmenting and Clustering

- Wiki page contains 103 rows × 3 columns.

- According to the merged table (Wiki + latitudes and longitudes for various neighbourhoods in Canada) we have found several neighbourhoods which contains Toronto.

Segmenting and Clustering

- Now we can visualize all the Neighbourhoods of the above data frame “Toronto” using Folium and the result we can see on the map:



Segmenting and Clustering

- The first neighborhood's name in our data frame is 'The Beaches'.

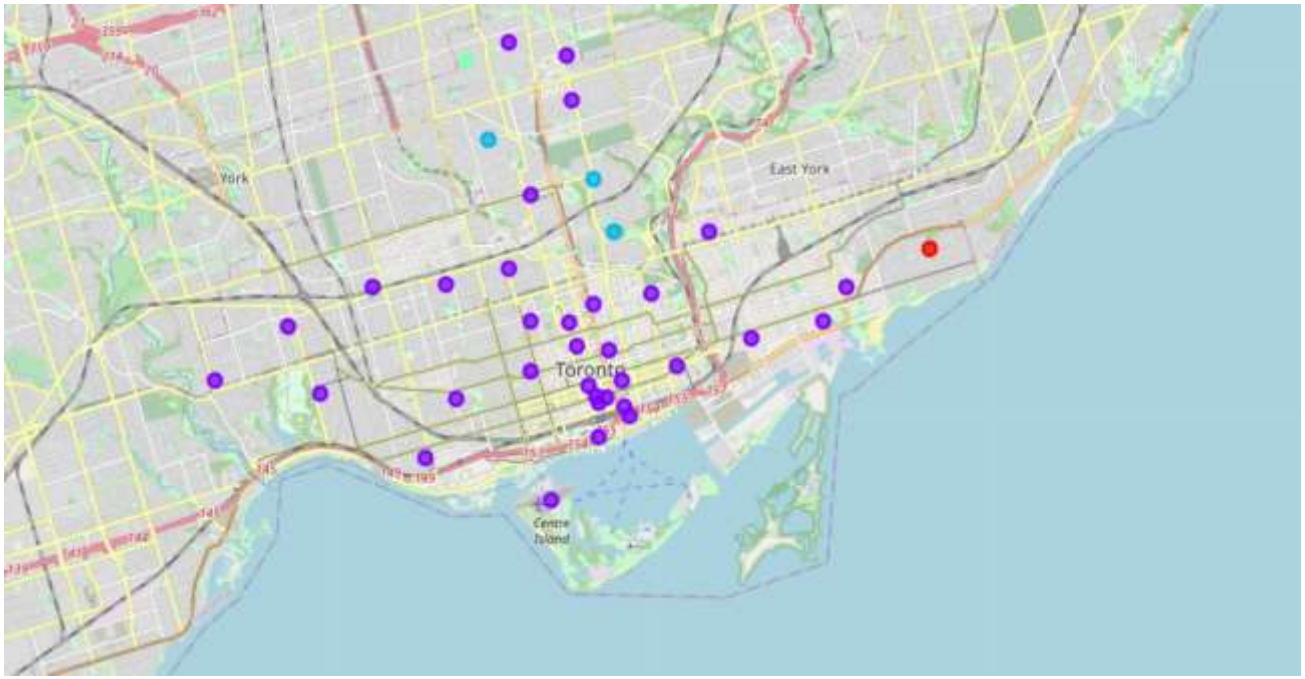
Name	Latitude	Longitude
The Beaches	43.67635739999999	-79.2930312

- Using Foursquare API we could get the top 100 venues that are in The Beaches within a radius of 500 meters.

	name	categories	lat	lng
0	Glen Manor Ravine	Trail	43.676821	-79.293942
1	The Big Carrot Natural Food Market	Health Food Store	43.678879	-79.297734
2	Grover Pub and Grub	Pub	43.679181	-79.297215
3	Upper Beaches	Neighborhood	43.680563	-79.292869

Cluster neighborhoods

- For clustering we will use K—Means clustering method. The set number of clusters = 5.
- By the next step we will create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood. Visualization of the resulting clusters is on the map below:



Cluster neighborhoods

- For clustering we will use K—Means clustering method. The set number of clusters = 5.

Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
1	Pub	Trail	Health Food Store
2	Coffee Shop	Restaurant	Café
3	Park	Trail	Playground
4	Home Service	Garden	Wine Bar
5	Park	Bus Line	Swim School

Conclusion

- During these 5 weeks we have understand the basics of K-Means Clustering, using of Foursquare API and repository of Github.