

Clustering the Toronto Neighborhoods

Husam Alhwadi

19th December 2020

2. Data

2.1 Data Source

1- Data from Foursquare system (<https://foursquare.com/>) will be used to build proper data science model to address this problem, Foursquare free user account will be used to retrieve the available and relevant data from Foursquare system via using Foursquare API's.

2- List of Toronto boroughs and neighborhoods from Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

2.2 Data Type

There are two types of data will be retrieved from Foursquare system:

Type1 The available venues for each neighborhoods

Type2 The category for each venue

2.3 Data Sample:

	Borough_Neighbourhood	Neighborhood_Latitude	Neighborhood_Longitude	Venue_Count	Variety
0	Central Toronto/Davisville	43.7020	-79.3853	19	11
1	Central Toronto/Davisville North	43.7135	-79.3887	18	12
2	Central Toronto/Forest Hill North & West, Fore...	43.6966	-79.4120	16	9
3	Central Toronto/Lawrence Park	43.7301	-79.3935	23	14
4	Central Toronto/Moore Park, Summerhill East	43.6899	-79.3853	20	12
...
98	York/Caledonia-Fairbanks	43.6889	-79.4507	22	13
99	York/Del Ray, Mount Dennis, Keelsdale and Silv...	43.6934	-79.4857	29	13
100	York/Humewood-Cedarvale	43.6915	-79.4307	25	15
101	York/Runnymede, The Junction North	43.6748	-79.4839	22	12
102	York/Weston	43.7068	-79.5170	20	11

2.4 Data Processing

1- Adding the latitude and longitude for each neighborhood by using foursquare end point:

<https://api.foursquare.com/v2/venues/explore>

2- Adding the category for each venue by using foursquare end point:

<https://api.foursquare.com/v2/venues/explore>

3- Preparing DataFrame with columns for Toronto neighborhoods, neighborhood latitude, neighborhoods longitude, Venue Id, Venue Category.

4- Grouping the dataframe records by using neighborhoods column.

- 5-** Process the column of venue category to split the categories for restaurants for each neighborhoods.
- 6-** Count the number of unique restaurant categories for each neighborhood.

2.5 Data Cleaning

- 1-** Removing the duplicated categories within each neighborhood, duplicated categories occurs when dataframe record is grouped so to avoid multi-counting for same restaurant category removing duplicated categories has been proceed.
- 2-** Continuous checking for dataframe to assure its freeness of null, unknown cells after manipulation its records.

2.6 Features Selections

Number of restaurants (Venue_Count) and the number of unique restaurants categories (Variety) will be selected as features for clustering model in this report.

These two fields will be used to segment the Toronto neighborhoods

2.7 Data Limitations

dataset used in this report is so restricted and don't represent the actual number of venues and their categories in Toronto neighborhoods Due to the limitation of free foursquare account which is used in this report as this user is eligible for 950 regular call type and 50 premium calls type per day and to retrieve up to 100 items when explore venue end point API's is used.