Coursera - IBM

Capstone Project

Clustering Neighborhoods of New York and Toronto Cities

Jun 01, 2020

Introduction/Business Problem

Our goal here is to identify similar neighborhoods across 2 different cities New York and Toronto, based on the most popular venue categories of the neighborhood.

We will utilize the Foursquare Location API, to find out the popular venue categories for each of the neighborhoods of New York and Toronto.

This is useful for people, who are likely to move on from New York to Toronto or Toronto to New York, and are looking to move to a neighbourhood with the similar characteristics (the same popular venue categories) as the one they are currently living in.

Data

Foursquare is a location data provider. We need to create a foursquare developer account and get the credentials to extract the data. The Foursquare website link to create the account is: https://developer.foursquare.com/ . The data can be extracted using the API or the foursquare python package.

Using foursquare data, we can search for nearby venues of a specific type, explore a particular venue, and search for trending venues around a location and many more. To find the most popular venue categories, we will be using the explore endpoint of the Foursquare API.

We make use of the data provided by the New York University, to get the list of neighborhoods in New York City along with their lat long values. The link is : https://geo.nyu.edu/catalog/nyu_2451_34572

We scrape the list of Toronto neighborhoods from wikipedia and map the respective lat lng values, using the file Geospatial_coordinates.csv. The wikipedia page is: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Methodology

Both the New York and Toronto Neighborhood's lat long data are concatenated and passed to the Foursquare explore method, with sortByPopularity set to 1, radius set to 1000 m and limit set to 100, so that the venues are within 1 km and not more than 100 venues are fetched for each neighborhood and the venues are sorted by the popularity.

Using this output I got sorted by Popularity, I'm going to assign each venue a popularity order. For example, let's say for a Neighborhood N1, the Foursquare API provides 5 venues as below:

Neighborhood	Venue	Venue Category
N1	V1	VC1
N1	V2	VC1
N1	V3	VC2
N1	V4	VC2

N1	V5	VC3

The Popularity Order will be assigned in descending Order as below:

Neighborhood	Venue	Venue Category	Popularity Order
N1	V1	VC1	5
N1	V2	VC1	4
N1	V3	VC2	3
N1	V4	VC2	2
N1	V5	VC3	1

Then we pick the max popularity order for each venue category, so that we could get the Popularity Order at the Venue Category Level. It will look something like below:

Neighborhood	Venue Category	Popularity Order
N1	VC1	5
N1	VC2	3
N1	VC3	1

Now let's rank it using the percentile method, so that the values range between 0 to 1. And we will name it as the Popularity Index.

Neighborhood	Venue Category	Popularity Index
N1	VC1	1.00
N1	VC2	0.67
N1	VC3	0.33

Now before feeding into the model, we will transform the data into wide format as below:

Neighborhood	VC1	VC2	VC3
N1	1.00	0.67	0.33

Also computed the top 10 popular venues as below

Neighborhood	1st Popular Venue	2nd Popular Venue	3rd Popular Venue
N1	VC1	VC2	VC3

I did the following **exploratory data analysis** over the venue category, before running the model.

1. To find the similar categories among a long list of venue categories, to group similar categories as one category. I found that most of the categories were different types of restaurants.

2. Do a bar plot on Venue Category, so that I could get an idea on how many neighborhoods have such a venue category.

I performed the following **data cleaning** work.

1. I noticed that many categories are available in just less than 50 neighborhoods among the 408 neighborhoods. Hence decided to clean up those categories.

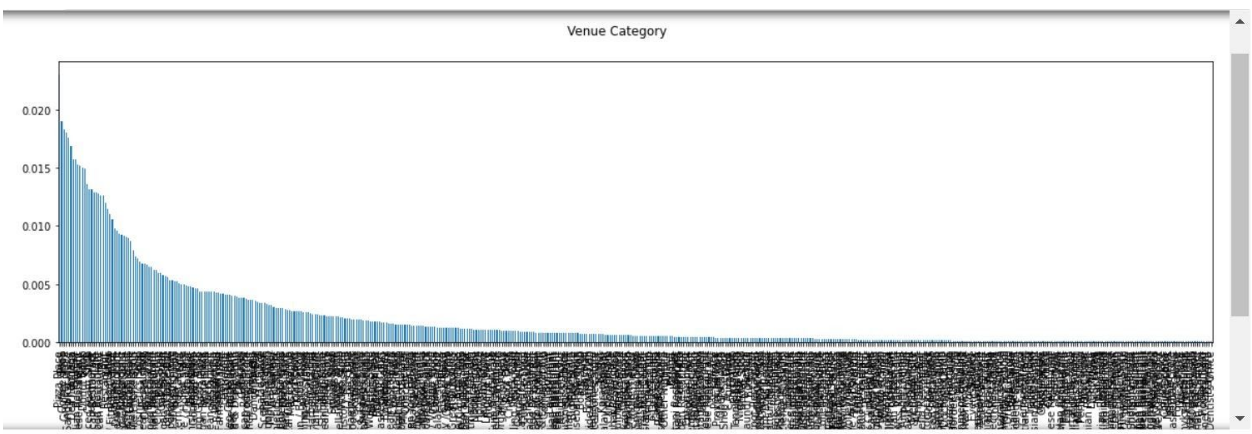
2. Grouped all the categories that have restaurant in its name and the neighborhood count is less than or equal to 50 as a common Restaurant Category

3. Removed all the rest of the categories that have neighborhood count less than 50

I selected the **KMeans** Algorithm to cluster the Neighborhoods. To find out the optimal number of clusters I'm using the Elbow method and doing the Silhouette Analysis. And I selected k=6 as the optimal number of clusters and ran the model using it.

Venue Category Word Cloud

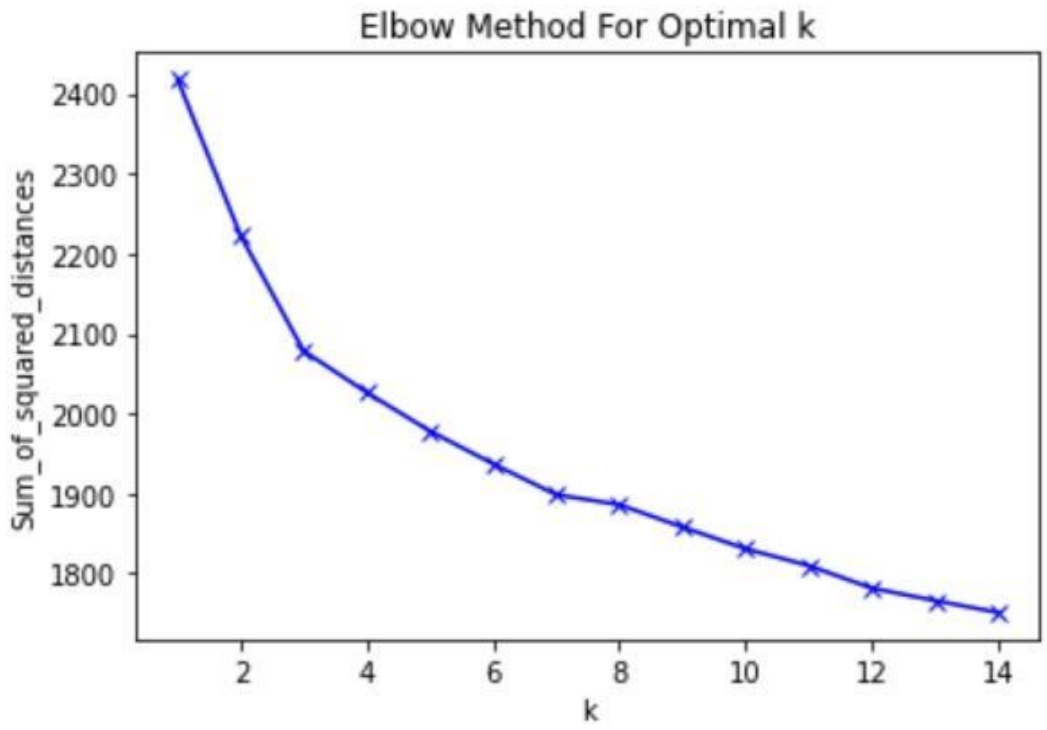
Venue Category Bar Plot

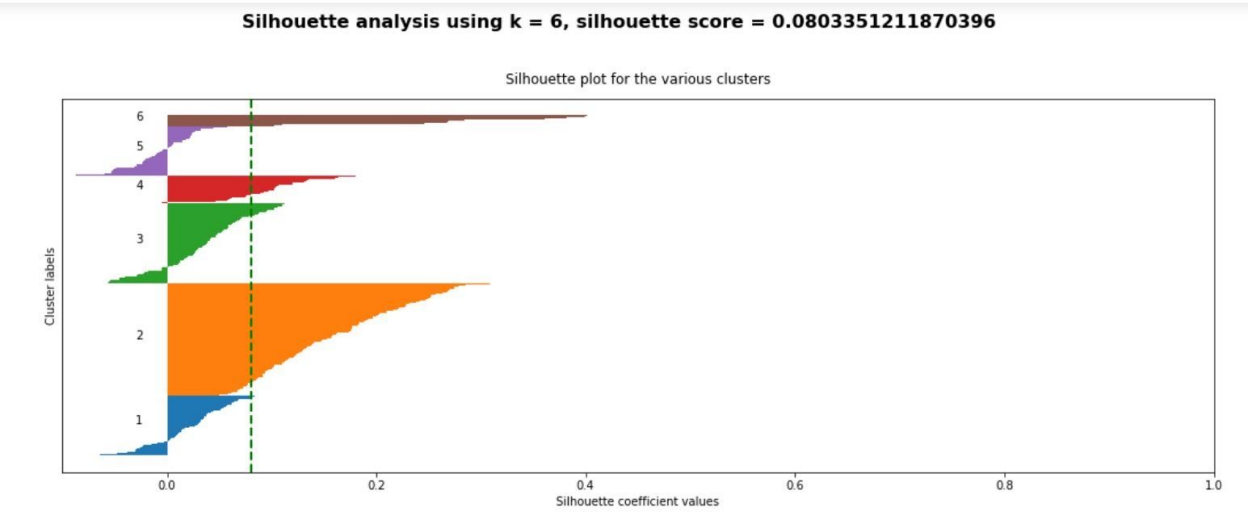


Final DF to be fed into the model

```
(408, 93)
```

	Neighborhood	American Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Bookstore	Breakfast Spot	Brewery	Bubble Tea Shop	Burger Joint	Bus Station	Bus Stop	C
0	Allerton, Bronx, New York	0.266667	0.0	0.0	0.0	0.066667	0.6000	0.2	0.0	0.0	0.633333	0.666667	0.0	0.0	0.033333	0.0000	
1	Annadale, Staten Island, New York	0.000000	0.0	0.0	0.0	0.000000	0.0000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.0000	
2	Arden Heights, Staten Island, New York	0.000000	0.0	0.0	0.0	0.000000	0.8125	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.1875	
3	Arlington, Staten Island, New York	0.000000	0.0	0.0	0.0	0.000000	0.0000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.8750	





Results

Found the top 10 popular venues for each cluster.

Cluster 1: Pharmacies are more popular

Cluster 2: Parks are more popular

Cluster 3: Donut Shops are more popular

Cluster 4: Grocery Stores are more popular

Cluster 5: Coffee Shops are more popular

Cluster 6: Shopping Malls are more popular

When someone enters the neighborhood they live in New York, it will output the suggested neighborhoods in Toronto, if any, and also display the top 5 popular venue categories in their cluster.

Also noticed that Cluster 1 & 4 have only neighborhoods of New York and Cluster 6 have only neighborhoods of Toronto. So we get to know, neighborhoods in these clusters are dissimilar with the neighborhoods of their respective other city

Discussion

I would like to discuss the further improvements that could be made to the model.

1. The data can be analyzed further, for any opportunity available for further data cleaning work.

2. We can try to visualize the data, and get to know about the shape of data, and try to select a most suitable clustering algorithm based on the shape of the data.

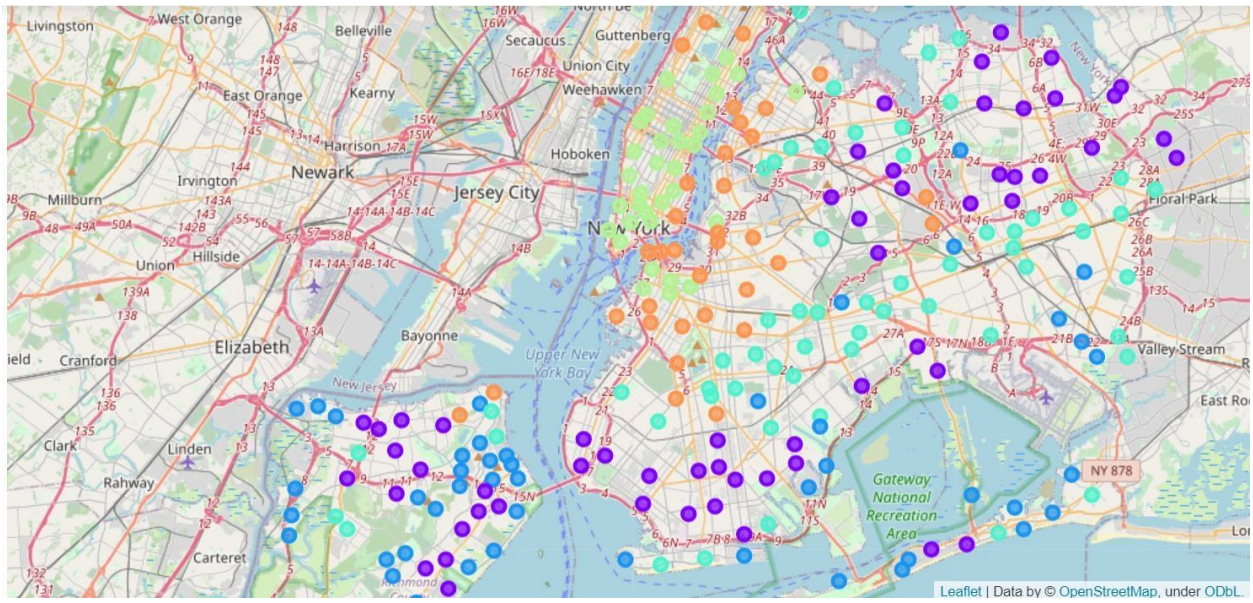
Top 10 Popular Venue Categories for each Cluster

	Cluster Labels	1st Most Popular Venue	2nd Most Popular Venue	3rd Most Popular Venue	4th Most Popular Venue	5th Most Popular Venue	6th Most Popular Venue	7th Most Popular Venue	8th Most Popular Venue	9th Most Popular Venue	10th Most Popular Venue
0	1	Pharmacy	Donut Shop	Supermarket	Pizza Place	Bank	Coffee Shop	Bagel Shop	Italian Restaurant	Convenience Store	Deli / Bodega
1	2	Park	Grocery Store	Coffee Shop	Pizza Place	Pharmacy	Restaurant	Bank	Fast Food Restaurant	Italian Restaurant	Bakery
2	3	Pharmacy	Donut Shop	Supermarket	Fast Food Restaurant	Grocery Store	Pizza Place	Discount Store	Park	Fried Chicken Joint	Bank
3	4	Grocery Store	Park	Gym	Hotel	Burger Joint	Coffee Shop	Gym / Fitness Center	Bakery	Pizza Place	Restaurant
4	5	Park	Coffee Shop	Grocery Store	Restaurant	Bar	Café	Gym	Pizza Place	Bakery	Supermarket
5	6	Shopping Mall	Hotel	Park	Supermarket	Plaza	Coffee Shop	Grocery Store	Bookstore	Gym	Gastropub

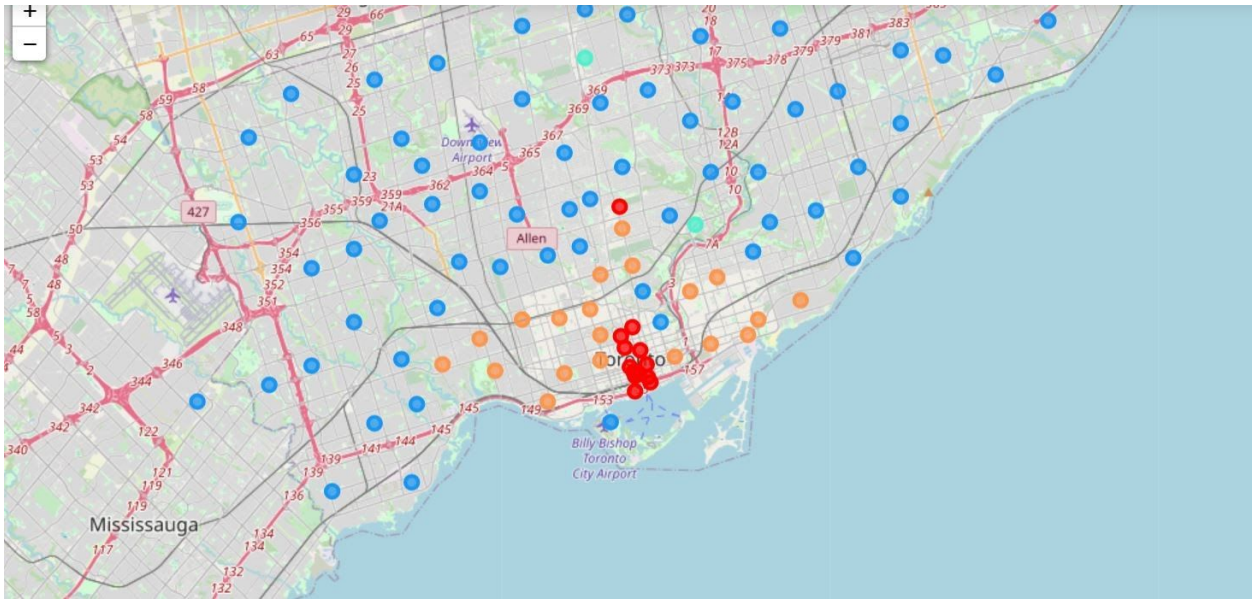
Suggested similar neighborhoods in the other city

```
Please enter the Neighborhood Name: Wakefield, Bronx, New York

 Cluster: 3

The top 5 popular venues for the cluster are:
Pharmacy
Donut Shop
Supermarket
Fast Food Restaurant
Grocery Store

The other similar neighborhoods you might be interested in Toronto are:
M4H, Thorncliffe Park, East York, Toronto
M2N, Willowdale, Willowdale East, North York, Toronto
```

New York Map



Toronto Map



Conclusion

Hence a model has been prepared using the KMeans algorithm on the New York and Toronto Neighborhood Data, with the motive to cluster the neighbourhoods based on the popularity of the venue categories that are provided by the Foursquare. And hence find out similar neighborhoods among both the cities.