Comparing the neighborhoods of the two New York City and Toronto

Denny Thomas

27th November 2020

Contents

Comparing the neighborhoods of the two New York City and Toronto

Denny Thomas

27th November 2020

1. Introduction

1.1. Background

Homo sapiens evolved from Africa and spread all over world over the time of millions of years. Earlier men were hunter gatherers slowly they start living in societies and by the introduction of agriculture and farming humans create villages, then slowly into different countries with fixed boundaries. Today we are in world having different countries which is having different cities of different cultures as per the demography.

Now migration going on with high pace as compared to the earlier times with the advancement of science technology and advanced logistics facilities. Earlier humans were migrated in search of food, climatic condition etc. In Present scenario humans' needs are changed, the main governing factors are, living conditions, facility, social security, opportunities etc. There for it is advantageous to analysis how similar and dissimilar are different cities in world.

1.2. Problem

Compare the neighborhoods of the two cities, New York City and Toronto and determine how similar or dissimilar based on the various factors such a facility, living conditions. The project aims to decided how similar are there and cluster the similar neighborhood based on the similar venues.

1.3. Target Audience

Developed for the people who wants to know the different Neighborhoods similarity around the Toronto and New York Cities. Based on this one can take the business decision, sales/marketing strategies for these regions.

2. Data acquisition and cleaning

Neighborhood details can be found in the Wikipedia and foursquare location data is used to explore or compare neighborhoods.

2.1. Description of Data

- Wikipedia: - Used for identifying the neighborhoods of the cities and analyzing the demographic and climatic data, this data will be used for clustering the neighbor hood
- Foursquare API- Used for exploring the venue and venue type, this data will be used for clustering the neighbor hood

2.2. Data cleaning

Data downloaded or scraped from multiple sources were combined into one table.

- Toronto Neighborhood & Borough downloaded from the Wikipedia table
- New York neighborhood and Borough downloaded from the Cousera JSON file

There were a lot of missing, I ignored the cell where borough is not assigned and split the row into multiple rows where one borough has multiple neighborhoods. After preparing the data frame created a function for extracting the longitudes and latitudes of the neighborhood using the Goopy

After completing the above activity for the New York and Toronto Data frame. I combined both Data frames into single Data frame for further analysis

2.3. Feature selection

After cleaning the data, I have extracted the different venues with venue category using the foursquare API. Total 14,549 venues grouped in to 463 categories are extracted. These categories are considered as feature for analyzing the similarity between the neighborhoods. All these features are discrete and converted into numerical form for the exploratory data analysis.

3. Methodology /Exploratory Data Analysis

3.1. Relationship with the Neighborhoods

Euclidean distance between the neighborhood are calculated using the venue categories as the variables. The below table is the distance matrix. where we can compare the different neighborhoods

Eg: - Distance from the Baychester to Deer park is 0.2877 and to Forest Hill Road Park is 0.5104 .The conclusion is Baychester is more similar to Deer Park than the Forest Hill Road Park

Borough	City	Neighborhood	Allerton	Baychester	Bedford Park	Deer Park	Forest Hill Road Park	Forest Hill SE
Bronx	New York City, NY	Allerton	0.0000	0.2593	0.2264	0.2809	0.5270	0.6055
Bronx	New York City, NY	Baychester	0.2593	0.0000	0.2631	0.2877	0.5104	0.5958
Bronx	New York City, NY	Bedford Park	0.2264	0.2631	0.0000	0.2794	0.5212	0.6111
Central Toronto	Toronto, Canada	Deer Park	0.2809	0.2877	0.2794	0.0000	0.5000	0.5963
Central Toronto	Toronto, Canada	Forest Hill Road Park	0.5270	0.5104	0.5212	0.5000	0.0000	0.6667
Central Toronto	Toronto, Canada	Forest Hill SE	0.6055	0.5958	0.6111	0.5963	0.6667	0.0000

Table: -1 extract of Distance matrix b/w the different Neighborhoods

3.2.Clustering (K means)

The neighborhoods of Toronto and New York are grouped in K clusters based on the similarities.

The value of K is iterated based on the mean distance centroid and corresponding K. The K which has minimum mean distance is considered for the clustering. The below Graph is plotted with for different K values and mean distance from the centroid. From the graph it is understood that for 'K'=3, mean is minimum, hence 3 is considered for the further analysis.

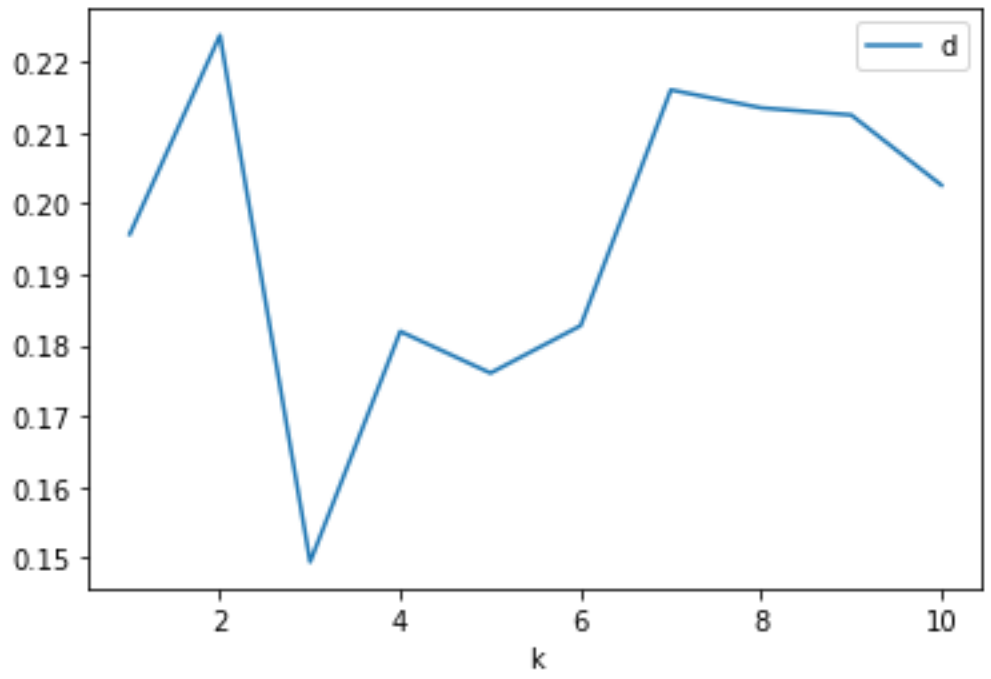

Fig:-1 Line plot , for the K Versus Mean distance from the centroid

The complete neighborhood is grouped in 3 clusters

1. Cluster 0: -455 no's of Neighborhoods
2. Cluster1: - 27 Nos of Neighborhoods
3. Cluster2:- 3 Nos of Neighborhoods

3.3.Visualizing the Neighborhoods in Folium

	Cluster Labels	Borough	City	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	0	Bronx	New York City, NY	Allerton	40.865788	-73.859319	Pizza Place	Deli / Bodega	Spa	Supermarket	Chinese Restaurant
1	0	Bronx	New York City, NY	Baychester	40.866858	-73.835798	Donut Shop	Fried Chicken Joint	Pizza Place	Shopping Mall	Men's Store
2	0	Bronx	New York City, NY	Bedford Park	40.870185	-73.885512	Diner	Pizza Place	Mexican Restaurant	Chinese Restaurant	Deli / Bodega
3	0	Bronx	New York City, NY	Belmont	40.857277	-73.888452	Italian Restaurant	Pizza Place	Deli / Bodega	Bakery	Bank
4	0	Bronx	New York City, NY	Bronxdale	40.852723	-73.861726	Italian Restaurant	Pizza Place	Spanish Restaurant	Breakfast Spot	Gym

Table: -2 sample table with first 5 common venues

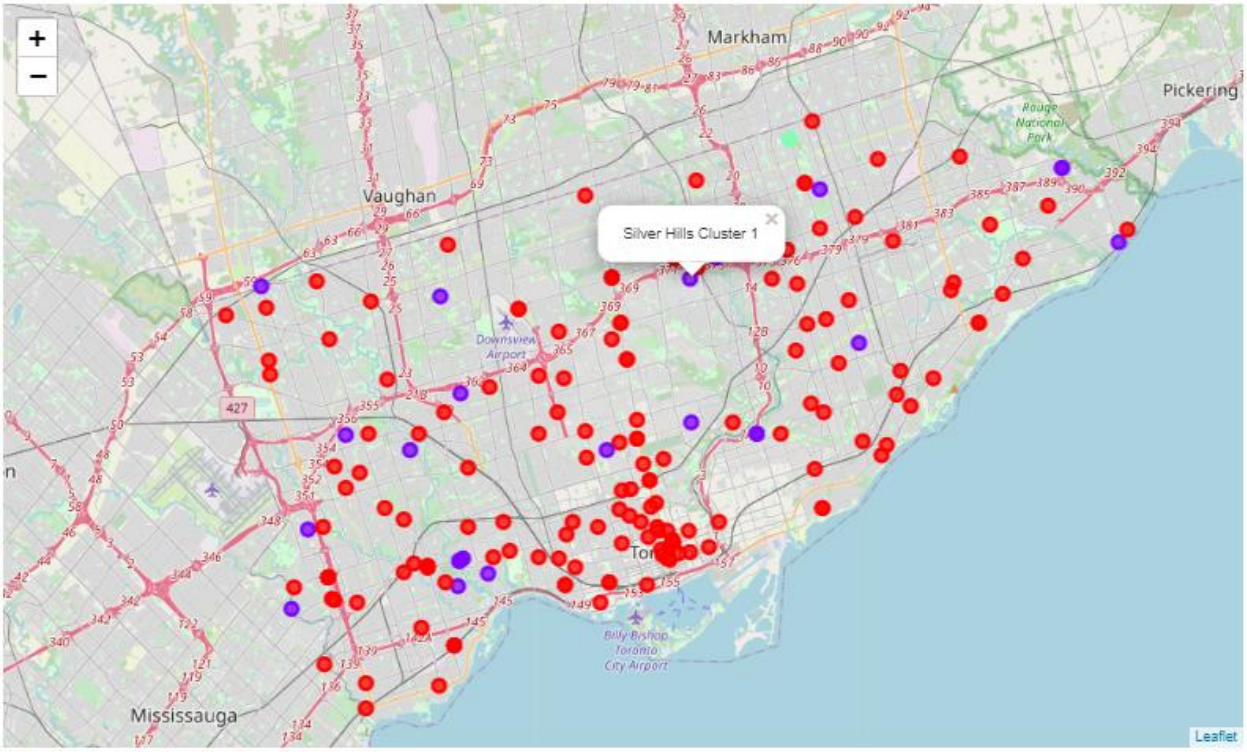

Fig:-2 Toronto Neighborhoods

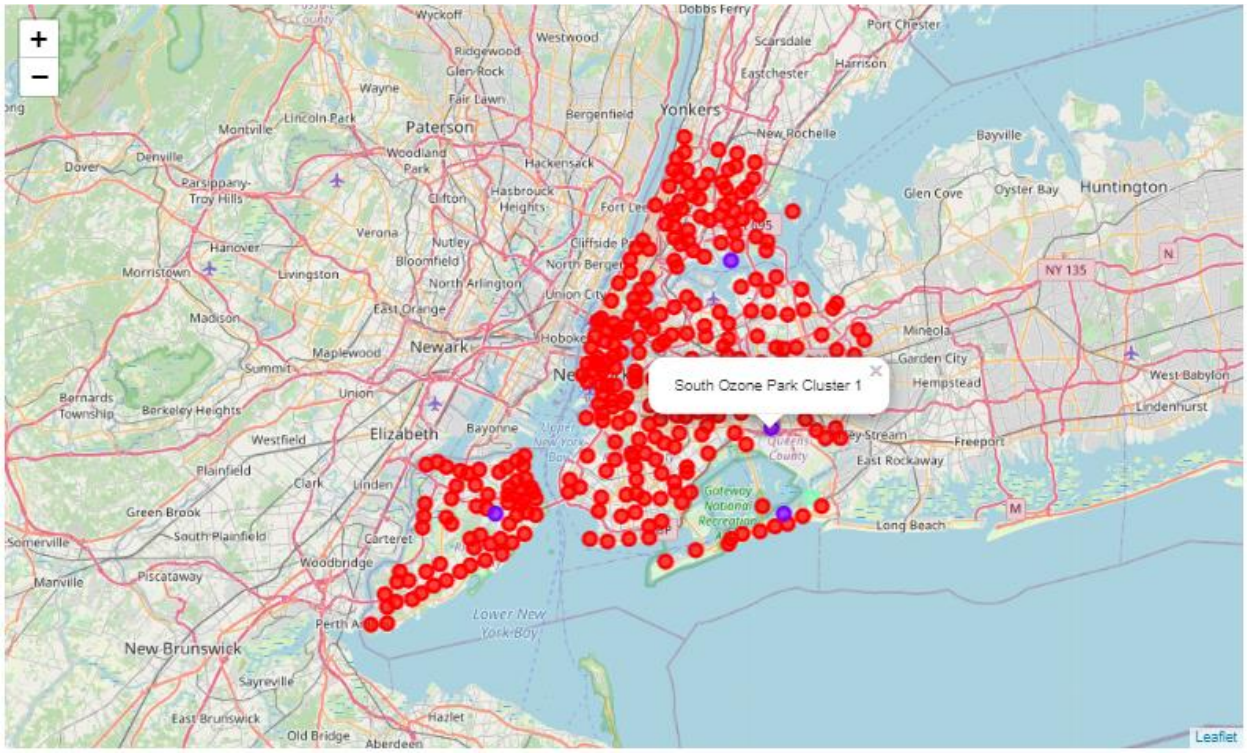

Fig3: New York City Neighborhoods

4. Results

In this Study I analyzed the similarity neighborhoods Toronto and New York City based on the data extracted from the foursquare. I developed the distance matrix between the neighborhoods and grouped the neighborhoods in to 3 clusters. This distance matrix is very useful, and this will help to give idea how neighborhood in particular city is like another neighborhood of other city. For example, a person living in New York and recently planning to migrate to Toronto can compare the similarity of different neighborhood in the Toronto with the help of distance matrix and can visualize the same in folium map

5. Future directions

I was able to cluster the neighborhoods only in the basis of venue category, I think model could use more variable such climate, population, per capita income. These data are obviously more difficult to extract and quantify, but if optimized, could bring significant improvements to the models