Clustering Similar Neighborhood in Different Cities

D.Dheeraj

May 19, 2020

1 Introduction

1.1 Background

Different cities in the world are filled with numerous kinds of venues that in turn define the cultures of the cities. A city not only differs from another by means of global positioning, what it to showcase to its inhabitants or tourists has put a significant mark on differentiating it from the rest. Despite of having dissimilarities, it is somewhat possible to group together the similar kind of neighborhoods in different cities. It is possible to segment the different venues in a neighborhood according to venue category, and then to group neighborhoods together that incorporate similar kind of neighborhoods. Having grouped together similar kind of neighborhoods may serve as a variable to help make a decision when people consider moving out of a city to another.

1.2 Problem

Comparing various features and neighborhoods in two cities in order to understand the level of similarity between the two cities.

1.3 Interest

To help people by providing a perception of similar neighborhoods which may provide with a great deal of insights in order to make a decision of choosing a neighborhood that is far away, yet somewhat feels like home.

2 Data Acquisition and Cleaning

2.1 Data Sources

This project works with two datasets. The first one is the data related to New York’s Neighborhood and the data is available in json format. And the second dataset is the data related to Toronto’s Neighborhood and this data is obtained from Wikipedia and the data is Scrapped using Web Scrapping Tool BeautifulSoup.

2.2 Data Cleaning

The first data source as mentioned above is in json format. It initially consisted of many different classes of data. Upon examining them, the data that we are interested in was found under ‘features’ category. Therefore, we extracted the data and stored the data in a dataframe using pandas library. The second data source is a Wikipedia page that contains Postal Code of the city of Toronto in a wiki table. To scrape the data from the URL, BeautifulSoup has been used to extract the table data. After going through a few more steps, the dataframe was obtained which consists of: Postal Code, Borough and Neighborhood. But the problem with this dataframe was, it has some values under the column ‘Borough’ which were not assigned in the first place. So, the rows with no assigned value in the ‘Borough’ column were dropped.

2.3 Feature Selection

Now that we have obtained the different neighborhoods and their respective geometric coordinates for the city of New York and Toronto, it is time to come up with different venues that the different venues have to offer. We use Foursquare API to do this job. Foursquare API provides with an access to an enormous database consisting of venues from all around the world including rich variety of information such as addresses, tips, photos and comments. Having signed up for a Foursquare developer, using the Client ID and Client Secret, it is possible to make API requests in order in order to retrieve venue information. By feeding a function with Neighborhood name and its geometric coordinates, using Foursquare API different venues (Restaurants, Coffee shops, etc) were extracted. After performing One-HotEncoding and grouping together the rows by neighborhoods, the NY dataset and Toronto dataset. Both the dataframes were combined into a single dataframe in order to perform clustering operation.

3 Methodology

The goal of this project is to find a set of similar neighborhoods in two different city, Manhattan and Toronto. And since it requires an unsupervised algorithm for finding similar clusters, we use K-Means in order to get the required solution.

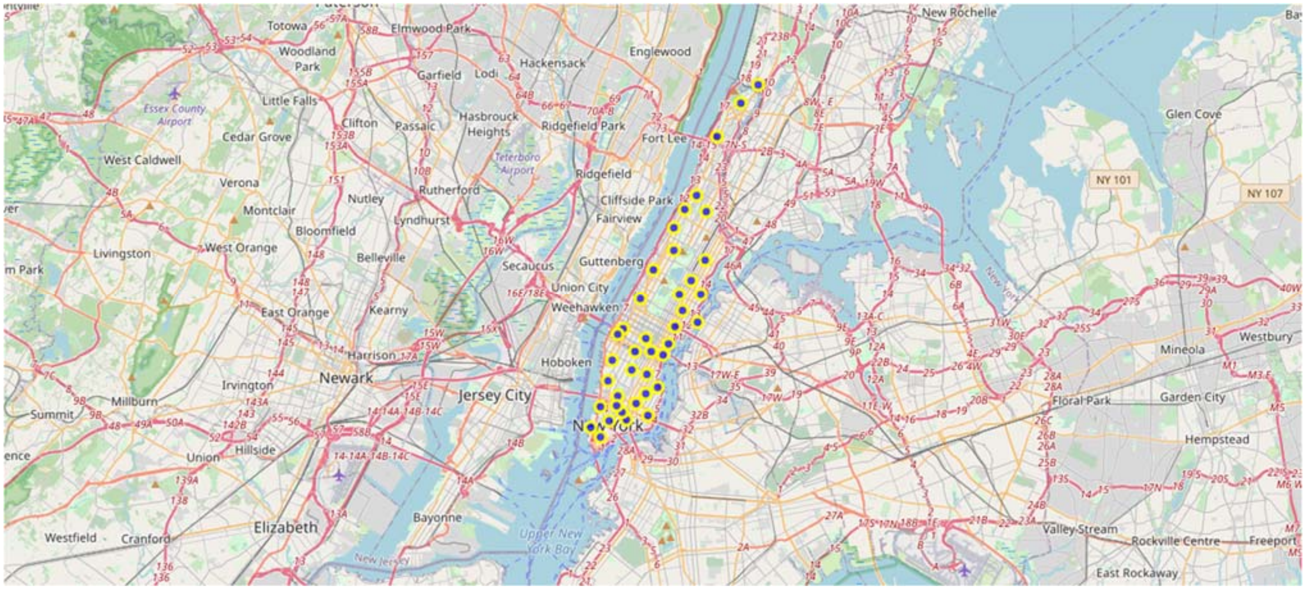
Exploratory Data Analysis (EDA)

In order to obtain the required clusters for our project, first we need to process our data such that it is ready to be used as an input for the algorithm. In case of Manhattan Data after extracting the required data from the json file, we obtain data of various boroughs of New York, thus we need to extract only the Manhattan's Data. Similarly, in case of Toronto data, after extracting all the data from Wikipedia using web scrapping tool called BeautifulSoup, we filter the data further and extract Toronto data from it and merge it with a dataframe containing its locations. Later we obtain nearby venues of both data separately and obtain mean of it. And later we join both data processed and feed it as an input for K-Means Algorithm with 8 clusters.

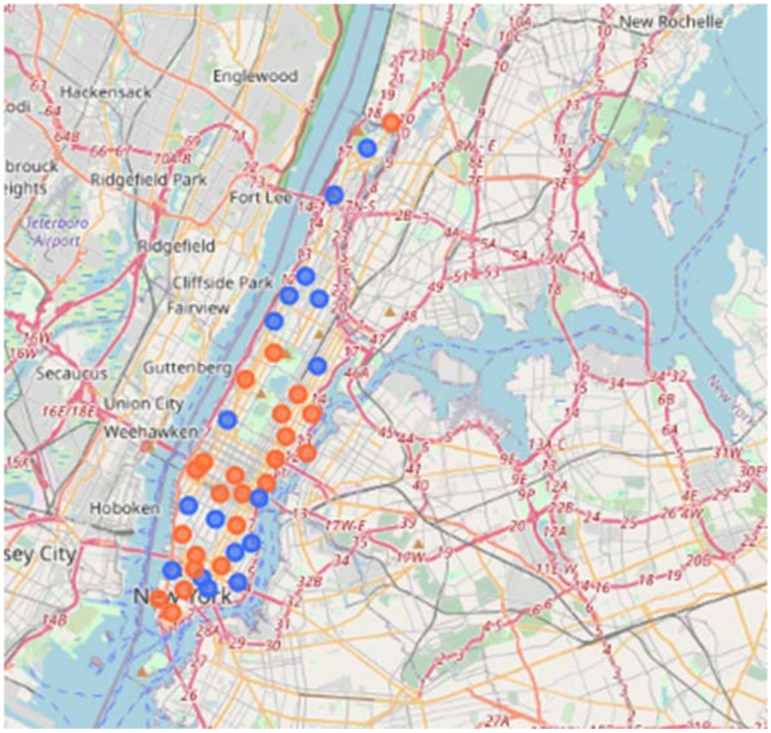
Apart from finding the clusters, we also find the most commonly visited places for each neighborhood and append the cluster labels obtained for each neighborhood and get ready to visualize it using Folium.

4 Result

After performing all the steps as mentioned above, we feed the data for visualization using Folium library and color code each cluster in order to visualize it better.



Manhattan Data Before Clustering

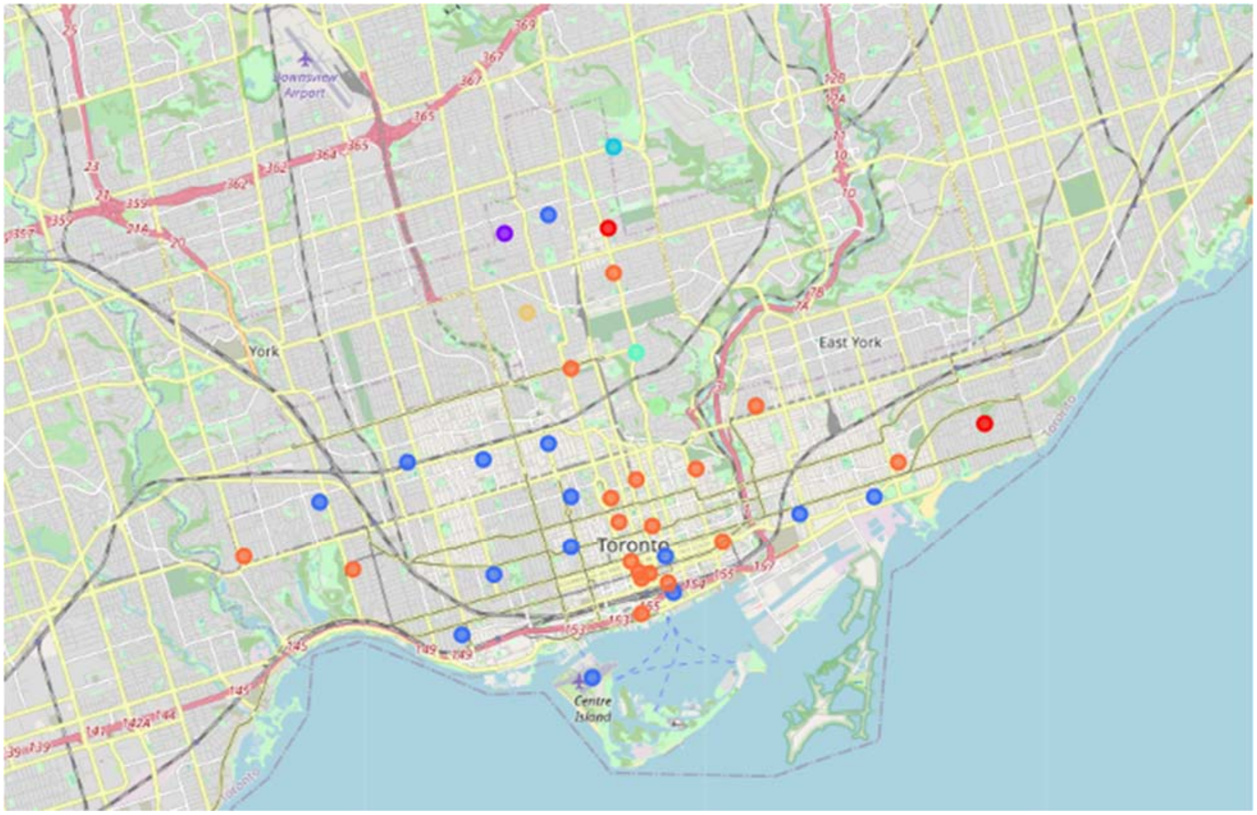


Manhattan Data After Clustering

Similarly, we visualize the other city, Toronto's Data and get the required clusters.

York
East York
Toronto
Centre Island

Toronto Data before Clustering



Toronto Data After Clustering

Therefore, from the above figures we obtain various neighborhoods in two different cities having similar locality. The number of clusters used were 8 and every clusters were color coded in order to differentiate.

5 Discussion

Since this is an unsupervised clustering work, many different approaches can be adopted in order to achieve better results. The project was only done on the zip codes of New York and Toronto, more samples may result in a better clustering For instance, for the outliers that are being observed on the maps could be defined by using DBSCAN algorithm.

Having dealt with location data on a deeper level, for instance at neighborhood level may result in better grouping of similar data points which eventually may result is better clustering. The study here is being ended by visualizing the data and clustering information on the map of the City of New York and Toronto.

6 Conclusion

People are frequently moving into new cities. And in this ever-growing world filled with technology, having a neighborhood recommendation based on location data is something to be considered basic now-a-days. And the application of neighborhood segmentation lies beyond this application too. This can serve to be an impressive tool to better organize a city resource. Furthermore, it can be used as a tool for security measurement if combined with crime data.