The Battle of Neighborhoods

Clustering similar neighborhood in different cities

Mirzana VM

May, 2020

1. Introduction

1.1 Background

New York (NY), is the most populous city in the United States. The city is the center of the New York metropolitan area. New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. People choose a place which ease their daily living to live on and they like to stick on it as it is difficult to find better or similar place like they living. Budding and flourishing a business or institution in a city like New York depend more on its geographical features and type of venues around them. So when a scenario like relocating a business or institution to another city comes, finding a similar location which promote the growth of the business plays an important role. Segmenting and grouping similar neighbourhoods would be a helping hand to many of them who wish to relocate to another place in different city.

1.2 Problem

'Company_X 'is a leading restaurant located in Manhattan, Lower East side, New York. In a strange scenario they are asked to relocate to some other area in the Toronto City. Company_X needs to find a similar place in Toronto where they can even grow better like their previous place.

2. Data

2.1 Data Source

The problem required two datasets of Neighborhoods, dataset of New York city and dataset of Toronto. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the boroughs and the neighborhoods that exist in each borough of two cities as well as the latitude and longitude coordinates of each neighborhood.

This dataset of New York City exists for free on the web. Here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572 . Download the data and store it in json format on your server.

The neighborhood data of Toronto is not readily available on the internet. For the Toronto neighborhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighborhoods in Toronto.

You will be required to scrape the Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,

Also, it need the Foursquare API to explore neighborhoods in New York City and Toronto. It use the **explore** function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. You will use the *k*-means clustering algorithm to complete this task.

2.2 Data Cleaning

The data set of New York is in json format. All the relevant data is in the *features* key, which is basically a list of the neighborhoods. We require the data to be transformed into data frame of four columns: Borough, Neighborhood, Latitude and Longitude. We can use 'geopy' library to get the latitude and longitude of New York city

The second data source is a Wikipedia page that contains Postcode of the city of Toronto. Wrangle the data, clean it, and then read it into a *pandas* data frame so that it is in a structured format which consists of: PostalCode, Borough and Neighborhood

Combine both data frame for performing clustering.

Data frame of New York data is made from a json file and Data frame of Toronto is made from the Wikipedia scraping.

We use geopy library to get the geographical coordinates of the city. Then we call Foursquare API to get all the neighborhood venues and its category of the city. We use Folium to visualize the city and the neighborhood venues in it. After getting all data, we create a data frame of venues in the city.

3. Methodology

3.1 Exploratory data analysis

When we are ready with the data frame, we analyse the number of venues in each neighborhood by group by function which in turn return the unique venue categories in the city. Analyse each neighborhood using one hot encoding. Rows are grouped by neighborhood and then we find mean of frequency of occurrence of each venues in that neighborhood. From that we retrieve top five or ten common venues in each neighborhood. Then a new data frame is created for storing top ten venues for each neighborhood.

3.2 Statistical testing for finding number of clusters

Before performing clustering we need to decide on the right number of clusters into which the neighborhoods are grouped. Here we use Elbow method to perform the same.

. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 20), and for each value of k calculate the sum of squared errors (SSE). Then, plot a line chart of the SSE for each value of k. If the line chart looks like an arm, then the "elbow" on the arm is the value of k to be used. The idea is that we want a small SSE, but the SSE tends to decrease toward 0 as we increase k (the SSE is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). So our goal is to choose a small value of k that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing k.

3.3 Machine learning approach

The goal of this project is to group together the similar neighborhoods in the city of New York and Toronto. Since the dataset is unlabelled i.e. unsupervised, we use k-means clustering algorithm.

K-means is a simple unsupervised machine learning algorithm that groups a dataset into a user specified number (k) of clusters. It clusters the data into k clusters, even if k is not the right number of clusters to use.

4. Results

4.1 Optimal Number of Clusters: Elbow method

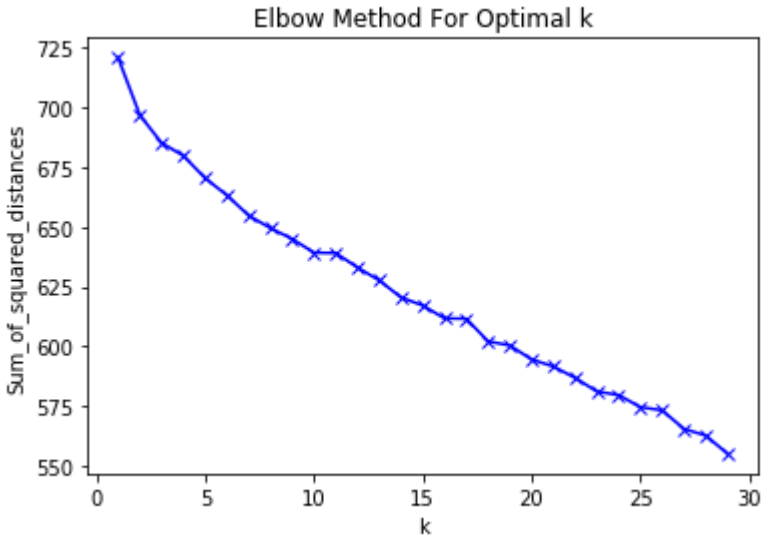

Figure 4.1

Elbow method to determine the number of K .Here we can't see a distinctive elbow

4.2 Visualizing the Clusters on the Map

We use Folium to visualize the clusters on the map

- Map of New York city neighborhood venues before clustering.

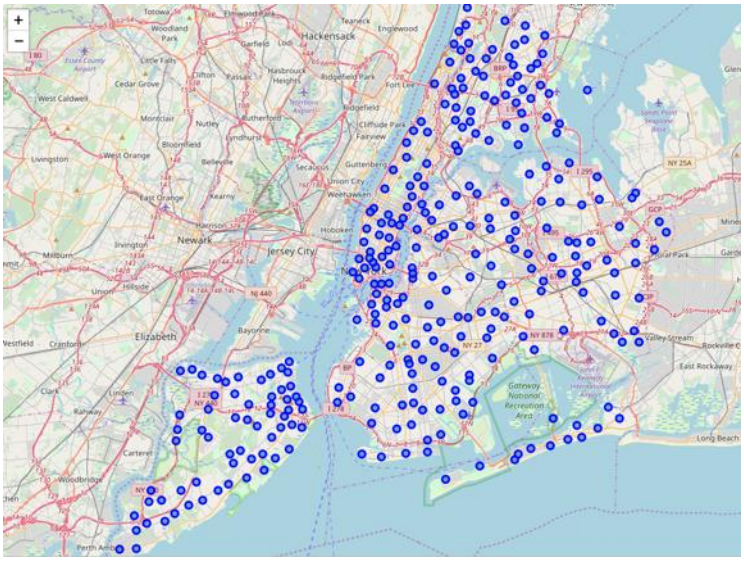

Figure 4.2(a)

- Map of New York city neighborhood venues before clustering

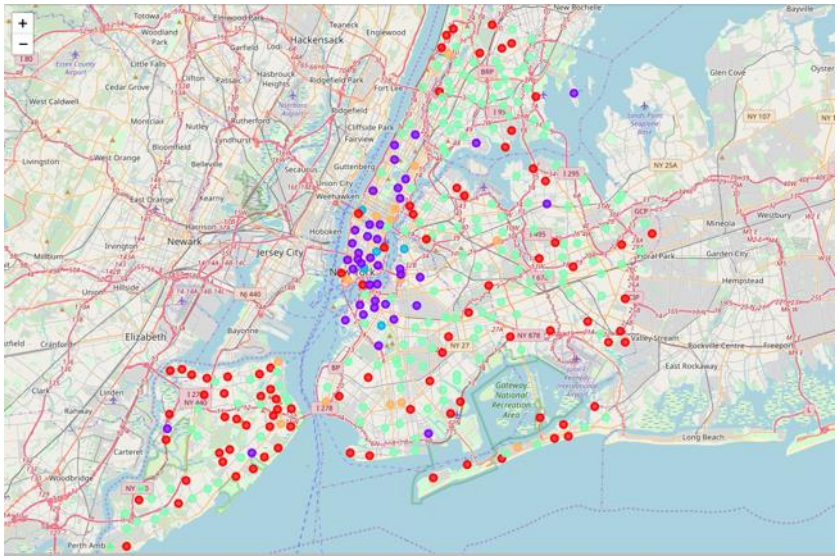

Figure 4.2 (b)

- Map of Toronto city neighborhood venues before clustering

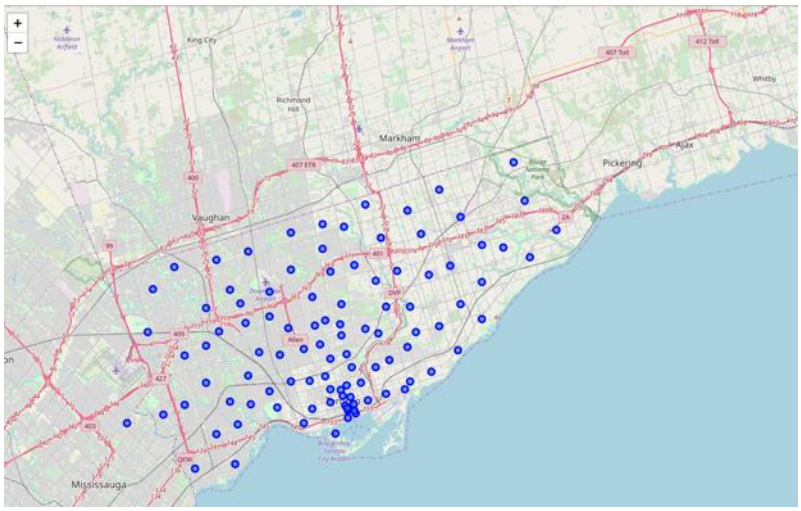

Figure 4.3(a)

- Map of Toronto city neighborhood venues after clustering

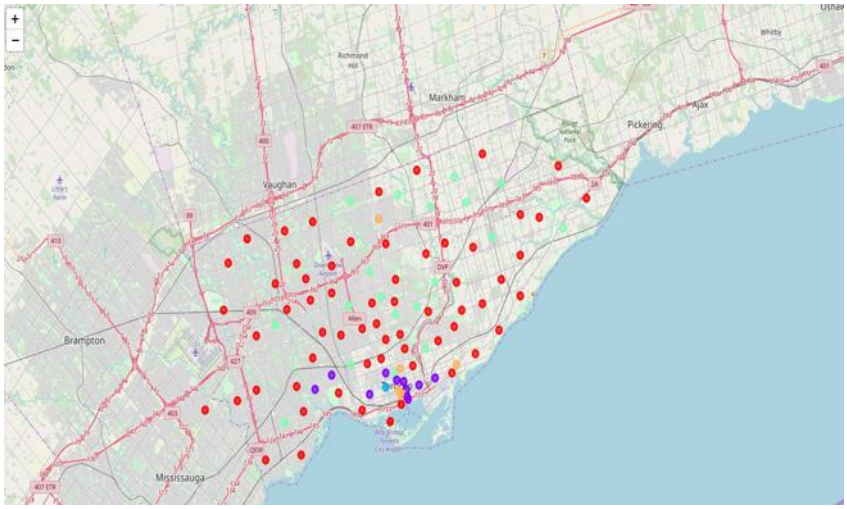

Figure 4.3(b)

Venues of both New York and Toronto have been grouped into five different clusters. Similar venues are pinned on the map with same colour i.e. same colour venues are belonged to a same class and those are assumed to be similar in features.

5 Discussion

In this project grouping of similar neighborhoods are performed with k-means clustering algorithm where k is the number of clusters which we attain by any one of the statistical testing method. Here we used Elbow method to find k, but it did not plot elbow of any good. Here we assigned k with an assumed value of five. We can use one or two method to find k for more accurate number of clustering. Here in this project outliers are not defined and if you want to deal with it, DBSCAN could be used.

6 Conclusion

Segmentation and clustering of neighborhoods among cities benefits many people who wished to relocate to another city. It also provide neighborhood recommendation to business. Besides that it can help in many other fields like national security enforcement, city structuring, understating geographical features and so on.