Clustering Neighborhoods in Brooklyn based on similarity to Toronto

Krishna Sai

July 7, 2019

1. Introduction

Background:

A company has different kinds of businesses around the different neighborhoods of Toronto. How these business flourish without any losses depends on several other factors in the neighborhood such as restaurants, cinemas, and other such venues in the neighbourhood. If a neighbourhood already has many food places it would be a bit difficult to start a new food business there and turn profitable.

Problem:

Now, the company wants to expand their businesses to neighborhoods around Brooklyn, Newyork. They want to know which neighborhoods are similar to those in Toronto so that they can set up a similar business around a similar neighbourhood. This helps them in divesting their capital and resources at a particular neighborhood based on the particular business that seems more viable.

2. Data Acquisition and Cleaning

Data Sources:

For the information about a particular neighborhood we are going to use Foursquare API to explore the neighbourhood and retrieve at least 100 venues within 500 meters radius of the neighborhood.

Data Cleaning:

For different neighborhoods around Toronto we use wikipedia page having the postal codes and neighborhood information of Toronto. We use BeautifulSoup to scrape the wikipedia page and store the neighbourhoods information in a pandas Dataframe. We then use FourSquare developer API to retrieve venues around each neighborhood in Toronto.
For the Newyork city data we use the json provided in week 3 project lab to get the neighborhoods around Newyork
From the data we receive from FourSquare API we only need the Venue name, Venue Location (that has Latitude and Longitude information) and the category of the Venue, we pick this data and create a dataframe with the information.

Feature Selection:

We then proceed with one hot encoding on venue category column so as to create a dataframe with category information. The resulting data frame has columns for each of the categories for each neighborhood. We then use this dataframe and group the data based on frequency of the category in each neighborhood. This gives us the features for classifying neighborhoods based on the different categories of venues present around each neighborhood.

We also have to make sure that we only take the categories of venues in Newyork that are present in Toronto to predict the similarity of the neighborhoods.

3. Exploring the Data

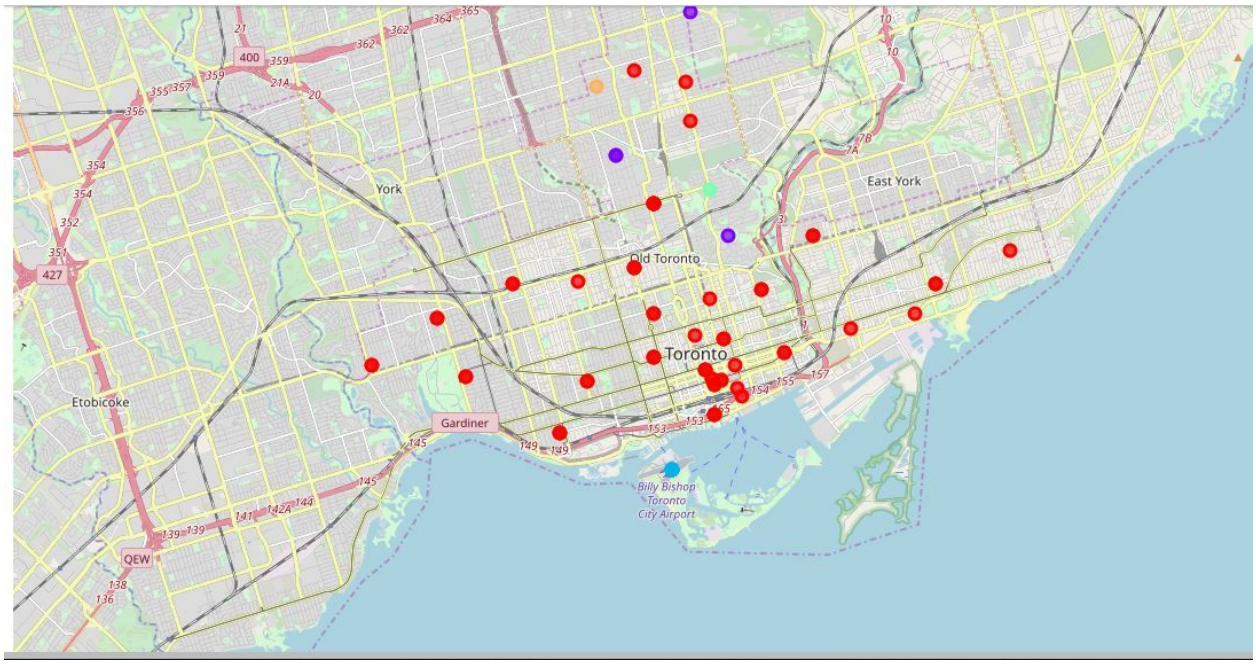
We can use the grouped neighborhood data to then figure out the most common categories of places around a given neighborhood. This gives a vague idea about similar neighborhoods based on the more common venues around the neighborhood.

We explore the data further by creating another data frame with the top 10 most commonly occuring category of venues around each neighborhood around toronto. We can compare this with brooklyn neighborhood data to get a fair idea about the similarity of most common venues.

4. Predictive Modeling

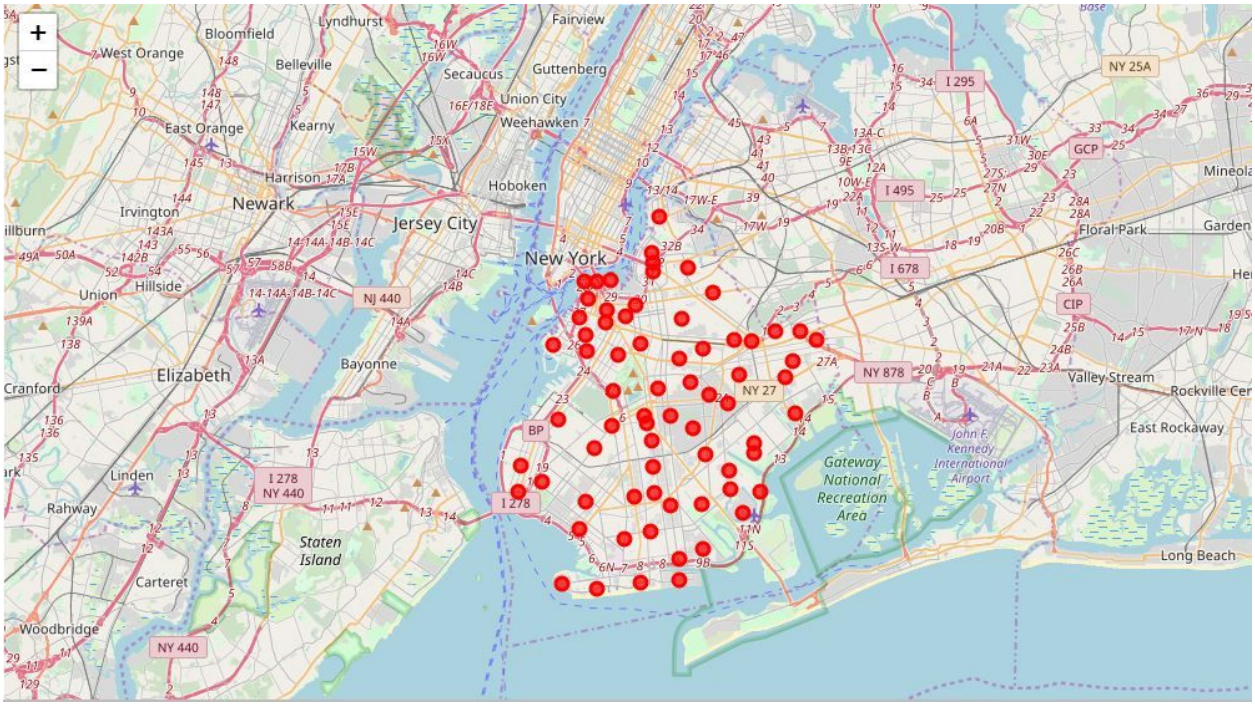
Since we do not have a labelled data to classify the neighborhoods, we use KMeans clustering to cluster the neighborhoods around toronto with the categories of different venues as Features for the modeling. As a start we use five clusters to cluster the neighborhoods around toronto.
Once the model is fit, we use the cluster labels to visualize the clusters on the map using Folium package.

The result is as below:



The different clusters of neighborhoods around Toronto are shown using different color maps.

We then use the KMeans model used to fit the Toronto neighborhoods data to predict the similar clusters around Brooklyn New york. This gives us the labels for clustering the neighborhoods around Brooklyn.
The visualization of clusters around brooklyn is as below:



5. Conclusion

From the modeling I was able to see that most of the neighborhoods around Brooklyn are falling under cluster one of Toronto neighborhood clusters, this shows that the businesses in that particular cluster of neighborhoods in Toronto will do well in the neighborhoods around Brooklyn