IBM Data Science Capstone Project

Neighborhood Similarity Analyzer Between Multiple Cities

SAMRUDHI KELUSKAR

AUGUST 2019

Introduction & Business Problem

- Many people move to a new city each month with little knowledge about the neighborhood they initially choose to reside in
- Being able to compare neighborhoods with those in cities one has lived in should make the task of choosing a new neighborhood easier
- The objective of this project is to analyze and cluster neighborhoods in two cities to make comparisons possible
- We will test our methodology on Toronto and New York City
- This analysis can easily be extended to more than two cities for more accurate comparisons between neighborhoods in the new city vs. the ones in familiar cities

Data

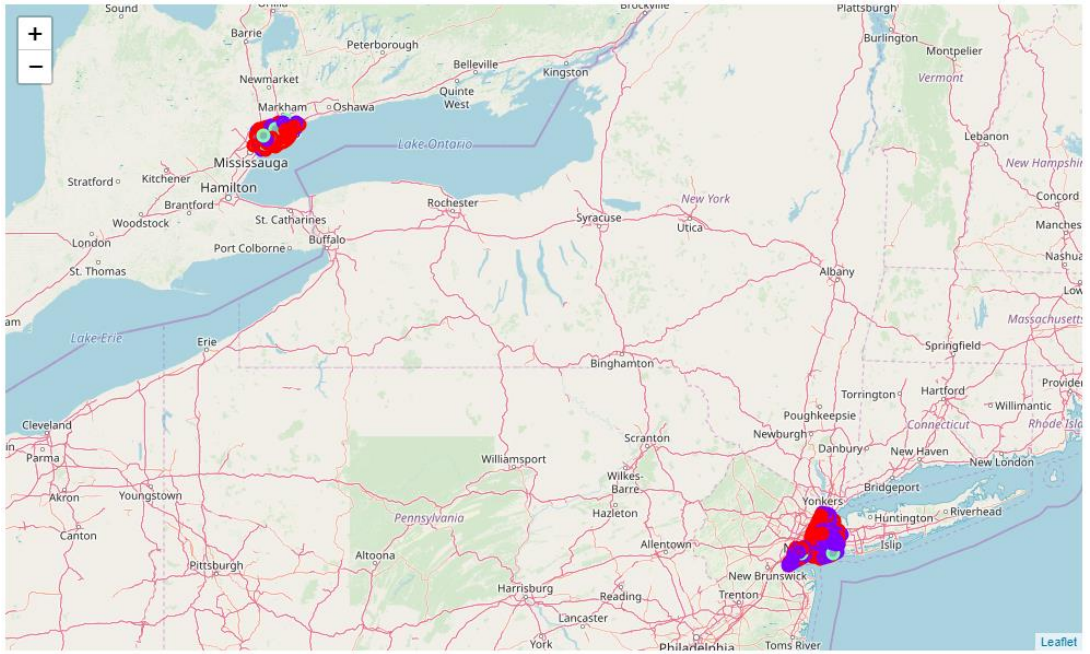
- Data Required:
  - List of Neighborhoods in all the Cities to be analyzed
  - Latitude and Longitudes of these Neighborhoods
  - Venue information for these Neighborhoods
- Data Sources:
  - Wikipedia or other publicly available data sources (we will use Wikipedia for data on Toronto neighborhoods and NYU's website for the list of neighborhoods in NYC)
  - GeoPy package for Latitudes and Longitudes
  - FourSquare API for Venue information

Methodology

- Scrape Wikipedia (and other sources) for list of Neighborhoods in each City
- Get Latitude and Longitude values for these Neighborhoods using GeoPy
- Use FourSquare to get Venue information for these Neighborhoods
- Summarize venue data per Neighborhood by frequency of each Venue Category
- Pick the top 25 Categories per City and combine them to form a single set of Categories
- Perform Clustering on the data but only use this set of Top Categories as clustering inputs
- Analyze the clusters to build descriptions of the clustered neighborhood groups
- Visualize the clusters on a Map

Results

- The neighborhoods in Toronto and NYC are grouped into 3 main clusters:
  - Cluster_0: Mainly consists of places to hangout like Cafés, Coffee Shops and Bars
  - Cluster_1: Mainly consists of places to eat like Pizza Places, Restaurants and Delis
  - Cluster_2: Mainly consists of places surrounded by Parks
- Zooming into the Map on the Jupyter Notebook shows how the three clusters are spread out in the two cities



Limitations of this Project and Conclusion

- This project only considered Venue types in a neighborhood as a defining element of the area
- However in real life, other factors such as population density, demographic distribution and crime rate (amongst others) influence the characteristics of a neighborhood
- These data points were not easily available at the neighborhood level hence weren't included in the analysis
- Keeping in mind the above limitations, we conclude by stating that If the user knows of a neighborhood of their liking in a present or a past city, they could easily find out which neighborhoods in the new city would be similar to the ones they're familiar with
- This project should help the user to at least start shortlisting neighborhoods that might interest them

Thank You!