IBM Data Science Capstone Project

Final Report

Neighborhood Similarity Analyzer Between Multiple Cities

Samrudhi Keluskar

August 2019

Contents

Introduction

Thousands of people move to a new city each month. Many of them have little to no idea about the kind of neighborhood they will be moving into, at least initially. It would be nice if we could compare neighborhoods in the new city with a neighborhood from a city that we're already familiar with, like our current or past cities. This would help us make an informed decision about which neighborhood to choose so that the shock due to a new environment is minimalized.

Business Problem

The objective of this capstone project is to analyze and cluster neighborhoods of two cities (Toronto and NYC in this case). Using Data Science Methodology and Machine Learning techniques like Clustering, we aim to find out which Neighborhoods of any two Cities are Similar to each other? This analysis can be extended to more than two cities for more accurate comparisons between neighborhoods in the new city vs. the ones in familiar cities.

Target Audience

This analysis is particularly useful for people who will be moving to a new city in the near future (Toronto to NYC or vice versa in our case). Toronto and New York City are the biggest cities by population of their respective countries, thus comparisons between them will be more meaningful than comparing a large and a small city. As the same analysis can be extended to more than two cities, frequent movers will find it useful to compare the neighborhoods in the new city with the ones they've already lived in, as more accurate comparisons can be drawn if we have multiple cities to compare to.

Data

To solve the problem, we will use the following data:

- List of neighborhoods in two or more cities, Toronto and NYC in this case
- Latitude and Longitude values of these cities, required to obtain venue information from Four Square and also to plot maps
- Venue data from FourSquare, will be used for clustering similar neighborhoods together

Sources of Data and Methods to Extract them

Neighborhood information for NYC is provided by New York University in the form of a GeoJSON file, while that for Toronto can be extracted from the city's Wikipedia page. We will extract that information using web scraping using BeautifulSoup. We will use GeoPy to obtain lat-long co-ordinates of these neighborhoods. FourSquare API will be used to obtain venue information for each of these lat-long pairs for all cities.

This project will make use of many data science skills, from web scraping, to querying APIs, wrangling and cleaning data, to machine learning and map visualization.

In the next few sections, we will discuss Methodology and Results as well as Limitations for Future Research.

Methodology

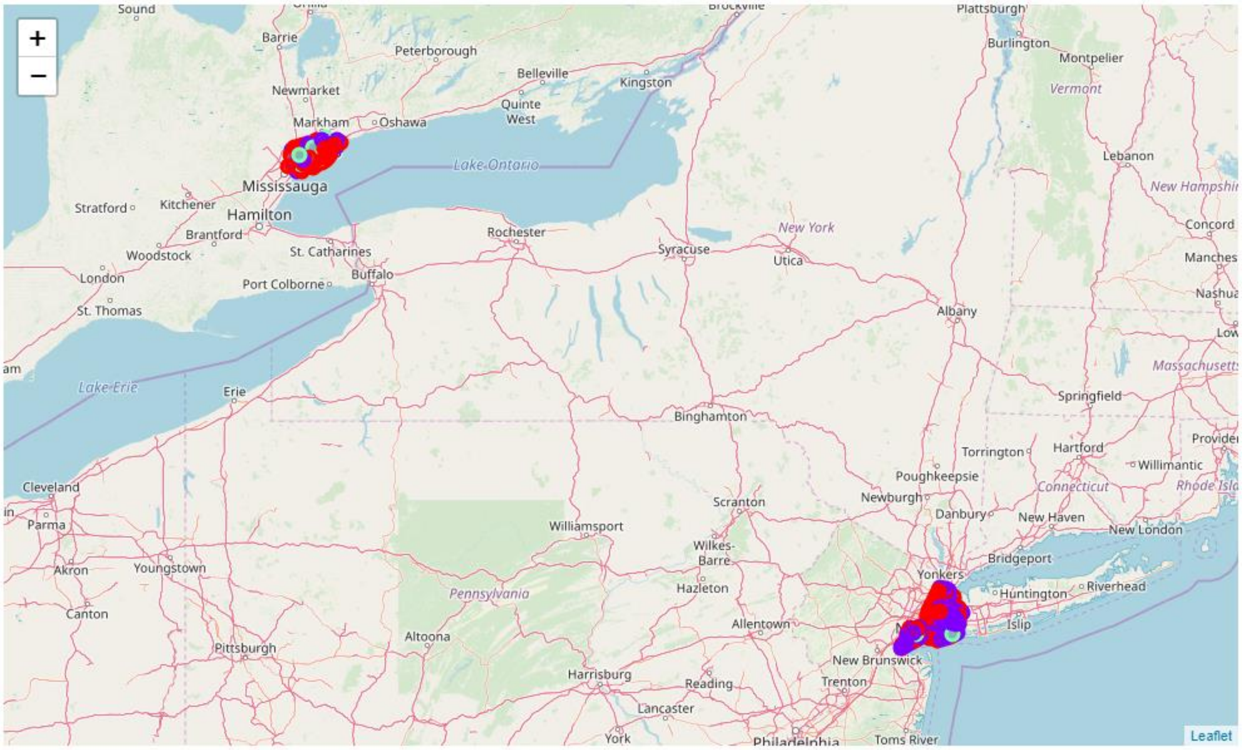
Firstly, we need to get the list of neighborhoods in the cities of interest. Fortunately, we have that provided to us for NYC by NYU and available for Toronto on Wikipedia. We will perform web scraping using Python requests and the BeautifulSoup package to extract the list of neighborhoods. However, these are just lists of neighborhood names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use the FourSquare API. To do so, we will use the GeoPy package that will allow us to convert addresses into geographical coordinates. After gathering the data, we will populate the data into a Pandas DataFrames and then visualize the neighborhoods on a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate data returned by GeoPy are correct. Next, we will use FourSquare API to get the top 500 venues that are within a radius of 500 meters of each neighborhood. We need to register for a FoursSuare Developer Account in order to obtain the credentials needed to access the API. We then make API calls to FourSquare passing in the geographical coordinates of the neighborhoods in a Python loop. FourSquare will return the venue data in JSON format and we will extract the venue's name, category, latitude and longitude. With this data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. We will pick top 25 venue categories in each of our cities (Toronto and NYC in our case) and combine them into a single set of categories. We will only use this new set of categories to cluster similar neighborhoods. We will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as away as possible from each other. We will divide the neighborhoods into 3 clusters based on their frequency of occurrence of the top venue categories in both (or all) cities. Neighborhoods falling in the same cluster will be considered similar and hence we can populate a list of neighborhoods in the new city similar to the neighborhoods in cities we're familiar with.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence of various Venue Categories:

- 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

Different values of 'k' in the clustering algorithm were tested, but anything more than k=3 resulted in sparse clusters (with less than 5 neighborhoods assigned to each cluster), hence the optimal value of k was set to 3.



Discussion

The three clusters coming out of the clustering algorithm represent the three broad types of environments represented by the different neighborhoods. The first cluster consists of lively areas filled with places for people to hangout, like Cafes and Bars. These neighborhoods are particularly great for young professionals and college students as they are usually more social. The next cluster consists of neighborhoods with a lot places to eat around. These kinds of areas would be best suited for people that enjoy eating out or want to keep food options close-by. The last cluster consists of neighborhoods with public recreation areas like parks around them. These areas are suitable for people that enjoy spending time outside around nature or people with young kids that would love to spend evenings outside playing.

If you already live in a neighborhood of your liking, you could easily find out which neighborhoods in the new city would be similar to the one you're currently in.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of various venue categories, but there are other factors such as population density, demographic distribution and crime rate (amongst others) that could influence a person's liking for a particular neighborhood. However, to the best knowledge of this researcher, these data points are not easily available at the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred location for a person to live in.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant users i.e. people looking for recommendations on selecting a suitable neighborhood in a new city. The analyses of this project will help the user to atleast start shortlisting neighborhoods that might interest them.

References

- Neighborhoods of Toronto Wikipedia page:
  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- Neighborhoods of New York City:
  https://geo.nyu.edu/catalog/nyu_2451_34572