Capstone Project - Battle of the Neighborhoods

Introduction

This project will analyze neighborhoods between Toronto, Canada and New York City, New York. A Fortune 500 company is looking to move its headquarters to either Toronto or New York City. The company wants insight into the neighborhoods and local businesses in the cities so that its employees may have the optimum living standards and quality of life. This project will explore the similarities and dissimilarities between certain neighborhoods in the two cities, and determine which neighborhoods best fit the culture of the Fortune 500 company's employees.

Data

The data used for this project will be acquired from the respective cities Wikipedia website pages. The datasets consists of the postal codes, neighborhood names, latitude, and longitude information for each neighborhood. Foursquare API search feature will be used to collect neighborhood venue information. Details about local venues and locality will be provide insight into the qualities of a neighborhood. In addition to Foursquare, various python packages will be used to create maps and machine learning models to further provide insights into our neighborhood battle project.

I used the following datasets from these websites:

Toronto Neighborhoods - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
Toronto Latitude and Longitude - http://cocl.us/Geospatial_data
New York City neighborhoods - https://geo.nyu.edu/catalog/nyu_2451_34572
New York City Latitude and Longitude = Python Geolibrar

Methodology

Work Flow:

1. HTTP requests would be made to this Foursquare API server using zip codes of the Seattle city neighborhoods to pull the location information (Latitude and Longitude).
2. Foursquare API search feature would be enabled to collect the nearby places of the neighborhoods. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 700.
3. Folium- Python visualization library would be used to visualize the neighborhoods cluster distribution of Seattle city over an interactive leaflet map.
4. Extensive comparative analysis of two randomly picked neighborhoods world be carried out to derive the desirable insights from the outcomes using python's scientific libraries Pandas, NumPy and Scikit-learn.
5. Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters of different categories of places residing in and around the neighborhoods. These clusters from each of those two chosen neighborhoods would be analyzed individually collectively and comparatively to derive the conclusions.

The following are the Python packages I used:

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library