Capstone Project - The Battle of Neighborhoods (Week 1)

Objective:

1. A description of the problem and a discussion of the background.
2. A description of the data and how it will be used to solve the problem.

Capstone Project – Comparing Metropolitan Neighborhoods

Problem Statement:

An international realtor has identified an opportunity to assist customers in relocating to large cities in North America. The realtor is seeking to a way provide a greater level of detail to their customers by developing information regarding the quantity and variety of venues within prospective cities. This information will allow customers to select the neighborhoods that best suit their interests. The research identifies the type and frequency of venues which can assist in defining a city's culture and allow customers to align their living preferences with specific neighborhoods.

The realtor's clients most common destinations are Toronto and New York City. To illustrate the available data and provide a comparative analysis, a report describes each city's respective neighborhoods and most common venues. This data is arrayed in illustrative maps that allow clients examine and explore wide array of neighborhoods which comprise each city.

Problem Description:

This project will analyze and compare the two most populous cities in Canada and the United States. Toronto is the capital city of the Canadian province of Ontario and the most populous city in Canada. It is comprised of various districts including East York, Etobicoke, North York, Old Toronto, Scarborough, York. New York City is the most populous city in the United States. The constituent counties, also known as boroughs include Bronx (The Bronx), Kings (Brooklyn), New York (Manhattan), Queens (Queens), Richmond (Staten Island).

Resources and data:

The websites listed below provided data in support of the research:

-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

The Python packages listed below were used to analyze the data:

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

Data Analysis Process:

1. Obtain Postal Code, Borough, and Neighborhood information from Wikipedia

2. Obtain Latitude and Longitude (lat/long) data from http://cocl.us/Geospatial data. Once collected, combine (lat/long) with previous data obtained from Wikipedia.

3. Use Folium to create maps that depict each city's neighborhoods. Once the city data is obtained, identify a more specific location (Borough) to identify neighborhoods.

4. Utilize HTTP requests to the Foursquare API server using (lat/long) of Toronto's and New York City's neighborhoods to pull venue and venue category information.

5. The Foursquare API search enabled the collection venue proximity within specific neighborhoods. Due to http request restrictions the number of places per neighborhood parameter is set to 100 and the radius parameter would be set to 500.

6. Folium- Python visualization library would be used to visualize the neighborhood clusters distribution of Toronto and New York City over an interactive leaflet map.

7. Extensive comparative analysis of two randomly picked neighborhoods (Scarborough and Flushing) is carried out to derive the desirable insights from the outcomes using python's scientific libraries Pandas, NumPy and Scikit-learn.'

8. Unsupervised machine learning algorithm K-mean clustering is 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 are analyzed individually and comparatively to derive the results.

9. In accordance with our problem statement, the data obtained would allow a potential client to review and consider the most frequently occurring venues. By comparing specific neighborhoods and illustrating K-means clustering on a map, clients would be able to select the best neighborhood to fit their needs.