Report - The Battle of Neighborhoods

IBM Data Science Professional Certificate - Week 2

Introduction

We hypothesize to be a Start Up which has to open his headquarters either in Toronto or in New York. We want to ensure that our employers have the higher standard of life possible. Therefore, we want to assure that our employers live and work in a place which has enough places and facilities for everyone's taste. Another hypothesis is that our enterprise has to deal with financial partners, in order to ease the relations with financial stakeholders we want to locate our Start Up in the financial district of each city. In New York the financial district is in Manhattan, while in Toronto is located in Downtown Toronto.

As mentioned the company wants information about the local businesses in the neighborhoods in order to assure the optimum living standards for its employers. We will explore the similarities and dissimilarities between these two financial districts, and determine which of them could be the best choice.

Data - Part 1

Used dataset:

- Toronto Neighborhoods - Wikipedia
- Toronto Latitude and longitude - https://cocl.us/Geospatial_data
- NY Neighborhoods - NyU Open Data Site
- NY Latitude Longitude - Python Geolibrar

Data - Part 2

As listed in the previous section, we used the Wikipedia web page to list the Neighborhoods in Toronto and the data offered by the New York City University's site.

In addiction, in order to collect neighborhood venue information we used

- Foursquare API

Methodology - Part 1

Python packages Used

- Pandas - for Data Analysis
- NumPy - to handle the data
- JSON - to handle JSON files for lot-lan of Toronto and Ny
- Geopy - To retrieve Location Data
- Requests - to handle http requests
- Sklearn - machine learning library to perform clustering
- Folium - to render the maps

Methodology - Part 2

- To pull location information we made HTTP requests using zip codes of Toronto and NY
- Foursquare API search feature enabled us to collect the nearby places of the neighborhoods
- Folium was used to visualize the neighborhoods cluster distribution of Toronto and New York
- The unsupervised machine learning algorithm K-mean clustering was applied to form the clusters around the neighborhoods

Results - Part 1

Downtown Toronto, Canada

Using K-means, groused the neighborhoods in Downtown toronto into 3 clusters .

- First cluster (Cluster 0) - has 17 neighborhoods and the most common venues are coffee shops, cafes and restaurants, which are optimal for our objective to find places for our younger employers which want meetings point and places to relax
- Second cluster (Cluster 1) - has 1 neighborhood and the most common venues are airports, we are not interested in this kind of places, although it could be interesting for business trip
- Third Cluster (Cluster 2) - has 1 neighborhood are coffee shops and Asian restaurants, which again are optimal results for our research

Results - Part 2

Manhattan, New York City

Again, I used k-means to group Manhattan's Borough into 5 clusters

- First - Has 15 Neighborhoods, and the most common venues are restaurants and cafes
- Second - has 8 neighborhoods, this time we have more gyms, theaters and parks
- Third - has 1 neighborhood, again, the most common venues are coffee shops, together with pizza places
- Fourth - has 8 neighborhoods, the most common venue are Restaurants, Cafes and Coffee shops
- Fifth - has 4 neighborhoods, in which there are more gyms and theaters

Discussion

- In Downtown Toronto we found 1284 venues in 18 neighborhoods. The Neighborhoods with the most venues are Adelaide, Kind, Richmond and Berczy Park. There are 75 distinct venues in 207 categories
- In Manhattan there are 3324 venues in 40 neighborhoods, and the most populated are battery park city and Carnegie Hill. There are 2887 distinct venues in 341 categories.

Many of the neighborhoods are homogeneous and similar to each other.

Conclusion and future development

Based on the quantity and variety of venues, Manhattan would be choose over Downtown Toronto as a choice to locate the Start Up. Manhattan offers more in choises for restaurants, gyms, grocery stores and extracurricular activities.

Limitation

This project doesn't consider employers with family and children, which may find more important have facilities such as schools, kindergartens or sport associations. However, we are considering a Start Up, which probably has employers aged between 19 and 30, more interested in meeting points.

Although cost of living in Toronto and New York is approximately the same, this project doesn't specifically consider it in the analysis. This cost could potentially affect the quality of living for employers.

A further analysis could consider the cost of living and facilities for families