Applying Data Science to Compare the Neighborhoods of New York and Toronto

1. Introduction

1.1 Background

New York is one of the largest metropolitan area in the world and most populous in the United States. On the other hand, Toronto is most populous city in Canada. Both cities are very diverse and are the financial capitals of their respective countries.

Larger metropolitan areas such as New York and Toronto are attracted by many people around the world due to their socio-economic values. Those metropolitan areas are excessively large in their geography and therefore usually divided into several boroughs and each borough consists of several neighborhoods. Diverse population in each neighborhood has a significant impact on its culture and various socio-economic aspects such as commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

1.2 Problem

It is a known fact that the venues around a given location add values to that location. It could be useful and meaningful if we could find out similar neighborhoods among each cities by analyzing venues surrounded by them. It would be great if we could simply illustrate group of similar neighborhood on a map so that one could make a valuable decision. However, due to its diversity it is very difficult to create a model or index to compare one neighborhood with another.

Modern data science tools have been evolved such that various unsupervised clustering techniques can be used to solve this kind of problems. Unsupervised clustering algorithms mainly focuses on creating clusters such that distance between points within the clusters are minimized and distance between inter-cluster points are maximized. Leveraging this concept together with data sources provided by popular location data providers we can find a scientific approach to group neighborhoods of different metropolitan areas based on surrounded venues.

1.3 Interest Groups

Anybody who is curious of socio-economic aspects of both New York and Toronto will be interest groups. Few of them would be;

1. Business personals who are interest on investing various ventures in those two cities.
2. Tourists who are interesting to find places which are similar to the places which they visited before in one of the cities.
3. Employees/workers/students who are migrating from one city to the other city
4. Data scientist who are curious about real-world application of data science

2. Data Acquisition and Cleaning

2.1 Data Sources

New York City data containing the neighborhoods, boroughs, latitudes and longitudes can be obtained from the data source https://cocl.us/new_york_dataset whereas neighborhoods and boroughs of Toronto can be find from Wikipedia website https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and latitudes and longitudes of each borough of Toronto can be obtained from http://cocl.us/Geospatial_data

Using Foursquare APIs we can obtain venues surrounded by each neighborhood of each city. The venues can be then used as features to cluster the neighborhoods of both New York and Toronto.