Applied Data Science Capstone Project

The Battle of Neighbourhoods:
using an Unsupervised Machine Learning Algorithm: KMean clustering

April 2020

Introduction

Toronto and New York City are the financial, entertainment and cultural centres of North America. They are less than 90-mintue flight apart. People sometimes refer Toronto to the “New York North”. However, with location-based data being more accessible to public nowadays, it would be interested in knowing the following questions:

- Whether it’s possible to qualify how similar (or dissimilar) these cities are by utilizing this data?
- Are we able to build a cluster model that captures the city dynamic and characterizes the urban neighbourhood?
- Is there any clear pattern in the model that will spark urban phenomena insights?

This report is a preliminary effort to analyze the dynamic in both cities and compare their similarity. It can be a reading material to people who are planning a move from one city to another but not sure about the uncertainty in changing of environment, or to people who are simply curious about these two famous North American cities. It can also be beneficial for readers who want to have some tourism guidances or would like to gain a sense on neighbourhood planning strategies in these two famous cities. By the end of the report, we will be able to disclose whether Toronto is the “New York North” or not.

Data

For our analysis, three main data sources will be used:

1. New York City Neighbourhood Dataset:

This dataset that contains the 5 boroughs in NYC and the neighbourhoods that exist in each borough as well as the the latitude and longitude coordinates of each neighbourhood.

Link: https://geo.nyu.edu/catalog/nyu_2451_34572

Out[3]:

	Borough	Neighborhood	Latitude	Longitude	City
0	Bronx	Wakefield	40.894705	-73.847201	New York
1	Bronx	Co-op City	40.874294	-73.829939	New York

2. Toronto Neighbourhood Dataset:

Similar to the New York City Dataset, this dataset also contains boroughs and the neighbourhoods that exist in each borough in Toronto, as well as the the latitude and longitude coordinates.

Link: *https://en.wikipedia.org/w/index.php title=List_of_postal_codes_of_Canada:_M&oldid=862527922*

Out[2]:

	Borough	Neighborhood	Latitude	Longitude	City
0	North York	Parkwoods	43.753259	-79.329656	Toronto
1	North York	Victoria Village	43.725882	-79.315572	Toronto

3. Foursquare API:

Foursquare is a location technology platform dedicated to collect trusted location data. In this result, it is used to get the top 100 venues within a radius of 500 meters of a neighbourhood. Data was retrieved using API calls.

Link: https://api.foursquare.com/v2/venues

Out[33]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Parkwoods	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy

Methodology, Results, Discussion and Conclusion

[to be continued in Capstone Project - The Battle of Neighborhoods (Week2)]