Capstone Project - How similar is New York to Toronto?

Introduction

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016. Current to 2016, the Toronto census metropolitan area, of which the majority is within the Greater Toronto Area, held a population of 5,928,040, making it Canada's most populous CMA. Toronto is the fastest growing city in North America, and is the anchor of an urban agglomeration, known as the Golden Horseshoe in Southern Ontario, located on the northwestern shore of Lake Ontario. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. [1]

The City of New York, usually called either New York City (NYC) or simply New York (NY), is the most populous city in the United States. Located at the southern tip of the state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass and one of the world's most populous megacities, with an estimated 19,979,477 people in its 2018 Metropolitan Statistical Area and 22,679,948 residents in its Combined Statistical Area. A global power city, New York City has been described as the cultural, financial, and media capital of the world, and exerts a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. The city's fast pace has inspired the term New York minute. Home to the headquarters of the United Nations, New York is an important centre for international diplomacy. [1]

From the description above, both of these two megacities are centres of business, finance, arts, culture, etc. Except for the fact that the population of New York is almost quadruple of that of Toronto, one may expect these two financial centres of North America to bear much similarities between them. In this project, we will use the techniques of Data Science to find out if these two cities are similar and how similar they are. In this process we may find unexpected answers that are only possible to obtain from data.

Business investor/people interested in moving to these two cities might be interested in this question. The differences in these two cities will provide insights of make financial decisions and relocation.

Data required

The venue exploration data from the FourSquare API, as well as Geolocation data for these two cities would be used for this project. From the distribution of different venues, analysis with be carried out to classify the neighbourhoods and measure the similarities between two cities.

1**) FourSquare API data** would be used to obtain venue information from Foursquare database for Toronto and New York.

2**)** Python package **geopy geocoder** provides the geometric data for these two cities which will then be referred to in calling the Foursquare database.

Methodology

Data preprocessing
Geolocation of Toronto and New York was obtained from previous Capstone projects. It was saved as csv files. We will use these files and pass them to FourSquare API calls. Venue information will be obtained from FourSquare Databases and incorporated into two pandas data frames.

Since New York is a much larger and consequently has more neighbourhoods and venues, we will randomly choose the same amount of neighbourhoods in city New York as that of Toronto. We will also confine the types of venues to those where both cities have.

Data analysis
K-means clustering algorithm will be used to classify all combined neighbourhoods from both cities into 10 clusters. These clusters will provide insights to our question.

Scikit-Learn package in python provides k-means clustering function, which is implemented in sklearn.cluster.KMeans.

Results

As is expected, out of 412 total neighbourhoods of both cities, 177 from city New York and 162 from city Toronto share much similarities. They account for 82.3% of total number of neighbourhoods. They belong to cluster 1.

However, there are 19 Neighbourhoods in Toronto are classified as cluster 2, 3, 5, and 9. They share no similarity to any neighbourhoods of New York. These Neighbourhoods in Toronto are: Cloverdale, Islington, Martin Grove, Princess Gardens, West Deane Park, Emery, Humber Bay, Humberlea, King's Mill Park, Kingsway Park South East, Mimico NE, Old Mill South, Royal York South East, Sunnylea, The Queensway East, Malvern, Rouge, Silver Hills, York Mills.

Likewise there are 24 Neighbourhoods in New York city are classified as cluster 7 and 10.They share no similarity to any neighbourhoods of Toronto. These Neighbourhoods in New York are: Arlington, Arrochar, Bellaire, Briarwood, Brookville, Clifton, Country Club, Elm Park, Fox Hills, Grasmere, Lighthouse Hill, Mariner's Harbor, Midland Beach, Oakwook, Ocean Hill, Randall Manor, South Ozone Park, Throgs Neck, Tottenville, Willowbrook, Breezy Point, Neponsit, Sea Gate, South Beach.

In the remaining cluster 4, 6 and 8, there are 6 New York Neighbourhoods and 24 Toronto Neighbourhoods. We observe that the ratio between them is 25%.

Discussion

The results also precisely identifies some special neighbourhoods in New York that belong to cluster 10, These places share the most common venue as beach.

	Cluster Labels	City	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
224	9	New York	Breezy Point	Beach	Board Shop	Monument / Landmark	Yoga Studio	Food Court	Financial or Legal Service	Fish & Chips Shop	Fish Market	Flea Market	Flower Shop
326	9	New York	Neponsit	Beach	Yoga Studio	Farmers Market	Filipino Restaurant	Financial or Legal Service	Fish & Chips Shop	Fish Market	Flea Market	Flower Shop	Food
370	9	New York	Sea Gate	Beach	Spa	Yoga Studio	Food	Fast Food Restaurant	Filipino Restaurant	Financial or Legal Service	Fish & Chips Shop	Fish Market	Flea Market
373	9	New York	South Beach	Beach	Athletics & Sports	Yoga Studio	Food & Drink Shop	Filipino Restaurant	Financial or Legal Service	Fish & Chips Shop	Fish Market	Flea Market	Flower Shop

While in Toronto, there are 5 neighbourhoods belonging to cluster 2 share the most common venue as bank. More surprisingly, they share the exact same 10 most common venue. One might be curious to know what are the 11th to 20th most common venue in these neighbourhoods.

	Cluster Labels	City	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
30	1	Toronto	Cloverdale	Bank	Yoga Studio	Drugstore	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	Dessert Shop
81	1	Toronto	Islington	Bank	Yoga Studio	Drugstore	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	Dessert Shop
103	1	Toronto	Martin Grove	Bank	Yoga Studio	Drugstore	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	Dessert Shop
131	1	Toronto	Princess Gardens	Bank	Yoga Studio	Drugstore	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	Dessert Shop
189	1	Toronto	West Deane Park	Bank	Yoga Studio	Drugstore	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	Dessert Shop

Similar pattern occurred for other clusters as well.
Even though the majority of the neighbourhoods share much similarities, however the data combined with a simple k-means clustering technique still spotted neighbourhoods of these two cities that are very different.

This result opens up avenues for further research and study, such as why in cluster 2, these neighbourhoods are almost identical to each other.

Conclusion

In this project we have demonstrated that a simple data analysis algorithm using k-means clustering combined with the comprehensive FourSquare database indeed are capable to differentiate geolocations of different cities.

This methods also provides insights are to why these locations are different and in what ways they are different.

This is only a naive application of Data science. Its potential is signifiant and vast.

References:

- **[1]** Toronto, New York - Wikipedia