Comparing Neighborhoods in Toronto and Mumbai

IBM DATA SCIENCE SPECIALIZATION CAPSTONE PROJECT

Introduction:

Toronto is the capital city of Canadian province of Ontario. It is the most populous city in Canada, a multicultural city, and Canada's financial and commercial centre. It is home to more than 2,731,571 people spread over an area of 630.2 sq. km.

Mumbai is the capital city of Maharashtra state of India. It is the largest city in India, being a cultural and financial centre for the country. It is home to more than 12.5 million people spread over 603.4 sq. km.

Despite being so far apart from each other, Mumbai and Toronto share a few characteristics:

- Similar sizes.
- Both cities are financial centres of their respective country.
- Both cities have a vibrant culture and strong commerce sector.

The goal is to analyse how different the neighborhoods in both these cities are by utilizing the location data of their neighborhoods and different venues present in them.

Methodology:

The goal is to analyse the difference between the neighborhoods of Mumbai and Toronto. The data required for this problem includes neighborhood names and their respective location data followed by the venue data for each neighborhood.

Location Data:

This project utilizes neighborhood names and location data scrapped from Wikipedia articles of Toronto Neighborhoods and Mumbai Neighborhoods read through **BeautifulSoup4** and **Pandas** libraries in python. This data is then used to get appropriate location encodings, if absent for respective neighborhoods.

There are 103 records in Toronto neighborhood data each containing multiple neighborhoods while there are a total of 93 records in Mumbai neighborhood data each containing single neighborhood.

Venues Data:

The data for recommended venues in each neighborhood is procured using **Foursquare API**. This data is further modified by getting the main category for each venue instead of the sub category gotten through api request. The data is then filtered to remove venue categories with less than 10 instances in all neighborhoods which further reduces the no of features we will work with while clustering neighborhoods.

Final Data preparation step consists of one-hot encoding the category column to get proper features for each venue. Further grouping by neighborhood is done with aggregate function being **sum.**

Lastly the data is utilized in clustering algorithm to cluster similar neighborhoods and for further analysis.

Results:

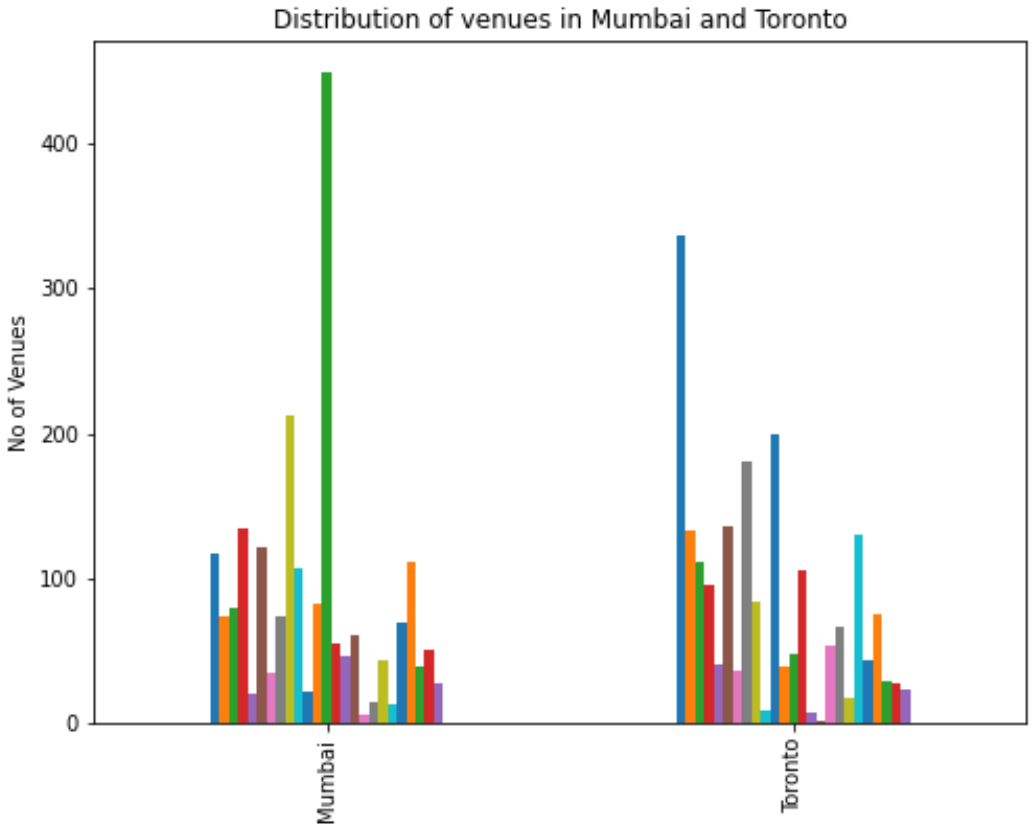
The Location Encodings for 103 records in Toronto dataframe were utilized to plot the distribution of the neighborhoods in Toronto.

The Location Encodings for 93 records in Mumbai dataframe was utilized to plot the distribution of neighborhoods in Mumbai.

Most of the neighborhoods in both cities are situated near the water source (Lake Ontario for Toronto and Arabian Sea for Mumbai). This signifies the importance of water route for transportation for both cities as they are commercial hubs.

A total of **5851** venues were requested succesfully and stored in a dataframe. There were 187 categories for the venues, which were reduced to 25 by removing venues of categories which had a frequency of less than 50 (total 162 categories removed).

The data upon grouping and plotting produced the following visualization indicating the difference in distribution of venue categories in Toronto and Mumbai as a whole.



After One hot encoding and Clustering operations, a total of 38 clusters were formed with average size around 5 neighborhoods per cluster. The largest cluster had 13 neighborhoods, 5 from Mumbai and 8 from Toronto. The rest of the larger clusters displayed similar proportionate distribution.

Conclusion:

- Mumbai and Toronto despite having their similarities have a very different distribution of venues categories.
- Despite the difference in between cities, many of the neighborhoods have similar venues present in them.
- Majority of largest clusters were composed of almost equal number of neighborhoods from both cities.
- The final conclusion of the project is that there is high degree of difference between neighborhoods within a city but, there are similarities among various neighborhoods irrespective of their location.