**TORONTO And NEW YORK - How Similar Are They?!**

**Problem In Hand**

Toronto is one of the main/major cities of Canada, in fact it is the largest and most populous city of Canada. Similarly, New York is one of the main/major cities of USA, on an interesting note it is also the largest and most populous city of USA. Both are multicultural as well as the financial hubs of their respective countries. Seeing these similarities it would be interesting to see how similar or diverse they are to visit. As a tourist, I would like to understand how similar or dissimilar these 2 cities are with regards to restaurants, accommodation, places to visit and so on.

Tourism is one of the major pillars for the growth of an economy. Every city is unique in it's own way - culture, tradition, history and so on. So, as a tourist it would be great to have information on similarities or dissimilarities between 2 cities which would allow one to plan accordingly - like where to stay, places of interest and so on. For Instance - say I have visited New York before and like the restaurants I visited and now I am planning to visit Toronto and would like to stay in a similar place like New York.

**Data Sources**

For the above comparison to take place would be using the below sources -

For Toronto -

1. Wikipedia Page to get Borough, Post codes and Neighbourhood information - <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
2. Geospatial Coordinates csv file used in Week 3 of the assignment to get Latitudes and Longitudes of the Boroughs of Downtown Toronto
3. Foursquare API for getting nearby venues of Downtown Toronto

For New York -

1. JSON file provided as part of the capstone project containing Borough, Neighbourhood, Lat and Longitudes of Manhattan
2. Foursquare API for getting nearby venues of Manhattan

**Methodology**

The below approach has been followed as part of this project and has been detailed in this section of the report.

1. Data Acquisition
2. Data Wrangling
3. Data Exploration
4. Feature Engineering
5. Data Modelling / Clustering
6. Observations
7. **Data Acquisition**

For the above data sources, there are two major considerations while extracting/acquiring data.

1. Data from a web page – Wikipedia – Data for Toronto containing details of Boroughs, Post Code and Neighbourhood had to be extracted from a wiki table. To achieve the same BeautifulSoup package from Python has been used to scrape the Wikipedia page and extract the required information.
2. Data from Foursquare API – Foursquare is a local search and discovery engine which provides personal recommendations of places to go near a user’s current location. They provide an API call for accessing their data which has been used to extract information of nearby venues like restaurants, parks and so on based on the location details which we have provided (viz, Manhattan and Downtown Toronto)

Apart from the above the Longitudes and Latitude information has been taken from the csv and json files provided by Coursera as part of the project.

1. **Data Wrangling**

For both Toronto and Manhattan, similar process has been used and is described in this section as follows:

Postal Code, Borough and Neighbourhood information has been scraped from the web using Beautiful Soup package from Python. The wikitable and tags ‘td’ were identified using Find and FindAll function respectively within the package to segregate the table from the web page and to extract the required information from the table. A pandas dataframe was created and if there was no Neighbourhood assigned to a Borough then the Borough name was assigned to the Neighbourhood.

Latitude and Longitude information was taken from the csv file provided by Coursera and was merged with the data scraped from the web using Postal Codes.

The latitude and longitude information was used to extract the nearby venues of the Neighbourhood using Foursquare API.

1. **Data Exploration**

The nearby venues were grouped by Neighbourhood and details of how many venues per neighbourhood was extracted.

<>

Number of unique venues were identified and was found to be 206 for Toronto and 334 for Manhattan.

1. **Feature Engineering**

For each unique venue identified dummy variables are created indicating binary values where 1 indicates the venue exists within the neighbourhood and 0 indicating that it doesn’t.

The top 10 most common venues per neighbourhood was identified to give a view of most commonly visited places in the neighbourhood.

<>

1. **Data Modelling**

To cluster all the unique venues – K Means Clustering Algorithm with k value of 5 has been used to divide the venues into 5 different clusters according to similarities.

<>