**Clustering similar neighborhood in different cities**

Isha Choubey

June,2020

**Data Acquisition and cleaning**

## Data Sources

This project works with two sets of data. The first dataset consists of New York’s different neighborhoods and their respective geometric coordinates. The second dataset consists of Toronto’s different borough and their respective postcodes.

## Data Cleaning

The first data source in the described link is in json format. It initially consisted of many different classes of data. Upon examining them, the data that we are interested in was found under ‘features’

category. Further formatting of the json data finally resulted in a dataframe that consists of 4 columns, namely: Borough, Neighborhood, Latitude and Longitude.

The second data source is a Wikipedia page that contains Postcode of the city of Toronto in a wikitable. To scrape the data from the URL, BeautifulSoup has been used to extract the table data. After going through a few more steps, the dataframe was obtained which consists of: PostalCode, Borough and Neighborhood.

But the problem with this dataframe was, it has some values under the column ‘Borough’ which were not assigned in the first place. So, the rows with no assigned value in the ‘Borough’ column were dropped. Another problem was there were a few rows in the ‘Neighborhood’ column that too had no values assigned to it. As a solution, the value from the ‘Borough’ column of the respective row was copied into the ‘Neighborhood’ column.