2. Data

2.1 Data Source

The problem required two datasets of Neighborhoods, dataset of New York city and dataset of Toronto. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the boroughs and the neighborhoods that exist in each borough of two cities as well as the latitude and longitude coordinates of each neighborhood.

This dataset of New York City exists for free on the web. Here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572 . Download the data and store it in json format on your server.

The neighborhood data of Toronto is not readily available on the internet. For the Toronto neighborhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighborhoods in Toronto.

You will be required to scrape the Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,

Also, it need the Foursquare API to explore neighborhoods in New York City and Toronto. It use the **explore** function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. You will use the *k*-means clustering algorithm to complete this task.

2.2 Data Cleaning

The data set of New York is in json format. All the relevant data is in the *features* key, which is basically a list of the neighborhoods. We require the data to be transformed into data frame of four columns: Borough, Neighborhood, Latitude and Longitude. We can use 'geopy' library to get the latitude and longitude of New York city

The second data source is a Wikipedia page that contains Postcode of the city of Toronto. Wrangle the data, clean it, and then read it into a *pandas* data frame so that it is in a structured format which consists of: PostalCode, Borough and Neighborhood

Combine both data frame for performing clustering.