Data

There are two different datasets used for retrieving the neighbourhoods and location data; one for Toronto and another for New York. From here, we use the FourSquare API to gather the venues data for the respective neighbourhoods in these cities.

Toronto

The dataset used for Toronto consists of a list of postal code, followed by the Borough and Neighbourhood and their respective location in terms of longitude and latitude. Data for Canada location details were scraped from the Wikipedia page and input into a pandas dataframe. However, longitude and latitude data were taken from the excel file from the week 3 assignment. Data cleaning (filtering) was also performed on the pandas dataframe with the purpose of removing the neighbourhoods that do not belong to Toronto.

Next, we are able to obtain the surrounding venues (and their details) of the neighbourhoods in Toronto via FourSquare API, by providing the following inputs to FourSquare:

- List of Neighbourhoods in Toronto
- List of Longitude
- List of Latitude

The results returned contains the various venues near the list of neighbourhoods within a radius of 500 metres and a limit of up to 100 venues per neighbourhood.

New York

The dataset used for New York is retrieved from a JSON file provided in week 3 lab. We proceed to extract the neighbourhoods and location data in New York from the JSON file and input the data into a pandas dataframe.

Next, we are able to obtain the surrounding venues (and their details) of the neighbourhoods in New York via FourSquare API, by providing the following inputs to FourSquare:

- List of Neighbourhoods in New York
- List of Longitude
- List of Latitude

The results returned contains the various venues near the list of neighbourhoods within a radius of 500 metres and a limit of up to 100 venues per neighbourhood.

With both the neighbourhood and nearby venues data for Toronto and New York on hand, we can perform “clustering” separately for both cities.

Last but not least, we can perform a comparison between Toronto and New York based on the clustering results and analyse the similarities and differences between both cities.