Capstone Project-The Battle of Neighborhoods

(Choosing the best neighborhood for a Japanese customer)

2. Data Acquisition and Cleaning

2.1 Data sources

The following data sources were used to obtain the necessary information for the analysis:

- The information of Toronto neighborhoods and boroughs were obtained from Wikipedia https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
- The coordinates of Toronto neighborhoods were obtained using Geospatial dataset which can be obtained from http://cocl.us/Geospatial_data .
- Foursquare API was used to explore the available venues in each neighborhood in Toronto.

2.2 Data cleaning

Data was downloaded and scrapped from multiple sources to obtain the required information for the analysis and applying the clustering algorithm.

Beautiful Soap library was used for performing web scraping for Wikipedia site to obtain Toronto neighborhood and borough information. Data cleaning was done to convert the scrapped data into tabular DataFrame format.

The neighborhood and borough information obtained from Wikipedia was combined with the coordinate information to be to explore the venues available within 2 KM radius in each neighborhood using Foursquare API.