The Mover's Guide: Toronto to New York

2. Data

2.1 Data Source

Information about neighborhoods and their geographical coordinates are obtained using two different methods.

For the New York boroughs of Brooklyn, Manhattan and Queens, the data is available online and we use the same file as in the Week 2 assignment from the course.

For Toronto, since there is no widely available database, we make use of the wikipedia page - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M - to extract information about the geographical coordinates of the neighborhoods.
We use the BeautifulSoup package to import the wiki page as a table into the Jupyter notebook.

The data required for comparing the neighborhoods based on its attributes is achieved using data from Foursquare about these locations.

2.2 Data Cleaning

The data extracted for Toronto does not have the same format as the data extracted the New York.
For example, there are certain neighborhoods which are not assigned a name.
For these neighborhoods, we replace the neighborhood name with the name of the borough it is part of.
Additionally, while importing the table, the line breaker '\n' appears as part of the string, which is cumbersome to work with.
Hence, we manually remove '\n's from the dataframe.
Through these methods, we are able to generate a dataframe for Toronto containing the neighborhood names and geographical coordinates, consistent with the format for New York.
This is essential, because during the clustering step, these two datasets will be combined.

2.3 Feature Selection

The primary features required for comparison include attributes of the neighborhood such as restaurants, educational and sports facilities, access to various modes of transportation, recreational areas, etc.
These attributes are imported using freely available data from Foursquare.
Importing Foursquare data is the crucial step because it lists the features as a function of the geographical location, which is key in order to compare neighborhoods.

The features are chosen in a 500 metre radius from the center of the particular neighborhood as given by its latitude and longitude.