Capstone Project: The Battle of Neighbourhoods

Mojdeh Soltani

November 2020

1 Data Acquisition and Cleaning

1.1 Data Sources

The following Wikipedia page is used to get information about neighborhoods in Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. This defines the scope of this project, which is the city of Toronto in Canada.

Also, we use the following CSV file to extract the geographical coordinates of different postal codes (neighborhoods): http://cocl.us/Geospatial_data. This is required to get the venue data and plot the map.

Finally, we request the venue data for each neighborhood from the Foursquare API. This data is used to execute clustering on the neighborhoods.

1.2 Data Cleaning

We combine the data downloaded from multiple sources into one table. After transforming the data into the Pandas data frame, we ignore the rows with ‘Not assigned’ label in the Borough column. Then we merge the neighborhoods with the same postal code. Finally, if a neighborhood has ‘Not assigned’ name, we consider the name of their borough as their neighborhood’s name.

1.3 Feature Selection

After all the merging and cleaning data that we mentioned above, we consider postal code, borough, neighborhood's name, latitude, and longitude of each neighborhood as shown in the following table (there are 103 rows and five columns). Note that in the methodology section, we will discuss how to consider and insert different events for each neighborhood as a new data frame.

	Postalcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

2 Methodology

2.1 Preparing the Primary Data

First, we use the BeautifoulSoup package to read the data about Toronto neighborhoods on the Wikipedia page, and then we transform it into the Pandas data frame as below.

	Postalcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront

Second, we use the CSV file to extract the geographical coordinates of different neighborhoods. Finally, after doing some data cleaning mentioned in section 2.2, we combine the data as follows.

	Postalcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476