

Capstone Project: The Battle of Neighbourhoods

03/23/2020

Data Acquisition and Cleaning

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.

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

Data Cleaning process

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.

Selecting Features

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