

## Data Description

Toronto neighbourhoods were chosen for this project due to the abundance of information given in the prior assignments within the Capstone module. The data that will be used for this project and report are from:

- Website scrapings from [https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) will be used to identify Toronto neighbourhoods
- Geospatial data will be given both by the geocode python package: <https://geocoder.readthedocs.io/index.html> and the csv file of Toronto postcode coordinates which can be downloaded from here: [http://cocl.us/Geospatial\\_data](http://cocl.us/Geospatial_data)
- Foursquare API, which provides the venues which surround the postcode coordinates and can be mined for information

To collect and clean the data

- The Toronto neighbourhood data will be scraped from the website and tidied, combining with the postcode coordinates from the geospatial data to create a dataframe of information
- For each neighbourhood in this dataframe, the coordinates will be passed to the Foursquare API using the explore endpoints to find the venues in a defined radius
- The neighbourhoods will then have the occurrence of each venue counted, with the dataframe transformed using one hot encoding to turn each venue type into a column with occurrence as the value

This will result in a dataframe where each row represents a neighbourhood in Toronto and each column is the occurrence of the venue in that neighbourhood. The data can then be processed for the rest of the project, finding the neighbourhoods where coffee shops are most prevalent, where other venues which may be related to coffee shops are also prevalent (e.g. parks, sports venues), and also to find the correlation between coffee shops and other venues through statistical analysis. Machine learning techniques will then be used to find the neighbourhoods which fit the criteria of suited for a coffee shop but not overcrowded already.