

Data

Data are the crucial part of any project. We need data that has four dimensions: volume, variety, velocity, and veracity. These four dimensions are the foundation of data science. For this project, we will need a data with the list of neighborhoods of the city, then we will need data with latitude and longitude coordinates of those neighborhood and finally venue data related to the coffee shops in the city.

Sources of data and methods to extract

This Wikipedia page ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M") has the list of neighborhoods in Canada. We will use web scraping techniques to extract the data from this page with the help of built in function of python pandas library. Then we will use this website (http://cocl.us/Geospatial_data) which was given by coursera to get the geographical coordinates of the neighborhoods of the Toronto. We will use geopy library to convert an address into latitude and longitude values. After that, we will use Foursquare API to get the venue data for those neighborhoods of the Toronto city. Foursquare has one of the largest databases of places of 105 million places and has multiple well-known partners such as Samsung, TouchTunes, Airbnb, uber. The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which you can send requests, so there is really nothing to download onto your server. In order to extract data from Foursquare API, we will need to open an account in Foursquare API then use credentials given by Foursquare API to request the venue data. In this way we can extract venue data of interested location. Foursquare API will provide many categories of the venue data, but we will focus more on coffee shops or shops related to coffee for this project. For the New York city we will download the data with borough, neighborhood, latitudes and longitudes for the website given by

coursera (https://cocl.us/new_york_dataset). We will use the web scraping technique to extract the data. For New York city, we will use wget library to download the file first and convert it to data frame. After that, we will use Foursquare API to get the venue data of the city focusing on coffee shop