Data Science Capstone Project

By Aayush Kamath

Problem Statement in brief:

This project intends to analyze suburbs/neighboorhoods in Mumbai, and come up with suburbs/neighboorhoods which might be ideal for opening a café.

Data Required to solve the business problem:

1. The list of suburbs/neighborhoods in Mumbai, India.

2. The geographical co-ordinates, that is, the latitude and longitude of the neighborhoods in Mumbai.

3. Data of venues near all neighborhoods, especially data of cafes.

Data Sources:

- The list of suburbs/neighborhoods in Mumbai is scraped from the wikipedia page of suburbs in Mumbai.

  (Link: https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai)

  The scraping of data is done using the Beautiful Soup library in Python.

- Once we have the list of all neighborhoods, we find the latitude and longitude for each of them using the Geocoder package in Python. The location data is necessary as it is used for creating a map of the neighborhoods of Mumbai. While the data remains stored in the dataframe, a quick glance at the map gives us an idea of the distribution of neighborhoods in Mumbai.

- The venue data for all neighborhoods are obtained by using the Foursquare API. By writing appropriate queries, we can access the necessary venue data through this API. Once the data is obtained, we sift through it to find the data that is necessary for arriving at a solution for our business problem. In this case, the necessary data being talked about is the data about cafes.

Data being used:

- The scraped data from the wikipedia page of Suburbs of mumbai for creation of an intial dataframe that displays the names of all the suburbs/neighborhoods in Mumbai.
- The latitude and longitude of every neighborhood, which is found using the Geocoder package
- Data of the top venues (limit 100) in a 1km radius of every neighborhood accessed using foursquare API.
- Finally, working on the top venues data obtained through the Foursquare API leads us to necessary data regarding cafes.