IBM Applied Data Science Capstone

JULY 7

Coursera Data Science Project
Authored by: Pranesh Kulkarni

Logo Name

The Battle of Neighborhoods

Toronto, Canada

The main objective of this Project is to help people in exploring better facilities around their neighborhood. It will help people making smart and efficient decision on selecting great neighborhood out of numbers of other neighborhoods in Scarborough, Toronto. Lots of people are migrating to various states of Canada and needed lots of research for good housing prices and reputed schools for their children. This project is for those people who are looking for better neighborhoods. For ease of accessing to Cafe, School, Super market, medical shops, grocery shops, mall, theatre, hospital, likeminded people, etc. This Project aim to create an analysis of features for a people migrating to Scarborough to search a best neighborhood as a comparative analysis between neighborhoods. The features include median housing price and better school according to ratings, crime rates of that particular area, road connectivity, weather conditions, good management for emergency, water resources both fresh and waste water and excrement conveyed in sewers and recreational facilities. It will help people to get awareness of the area and neighborhood before moving to a new city, state, country or place for their work or to start a new fresh life.

The Location:
Scarborough is a popular destination for new immigrants in Canada to reside. As a result, it is one of the most diverse and multicultural areas in the Greater Toronto Area, being home to various religious groups and places of worship. Although immigration has become a hot topic over the past few years with more governments seeking more restrictions on immigrants and refugees, the general trend of immigration into Canada has been one of on the rise.

Foursquare API:
This project would use Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

Work Flow:
Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

Clustering Approach:
To compare the similarities of two cities, we decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods in a big city like New York and Toronto. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm

Libraries Which Are Used to Develop the Project:
Pandas: For creating and manipulating data frames.
Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.
Scikit Learn: For importing k-means clustering.
JSON: Library to handle JSON files.
XML: To separate data from presentation and XML stores data in plain text format.
Geocoder: To retrieve Location Data.

Beautiful Soup and Requests: To scrap and library to handle http requests.
Matplotlib: Python Plotting Module.

Dataset

Raw Data Link:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The neighborhood data is available on the Wikipedia article. We have used beautiful soap a python library to get useful insights from raw html pages this data consists of zip postal codes. We have obtained latitude and longitude from geodecoder library
Will use Scarborough dataset which we scrapped from wikipedia on Week 3. Dataset consisting of latitude and longitude, zip codes.

Foursquare API Data:
We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API. After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meter.
The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude

4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category

Dataset link:
https://github.com/Pranesh6767/Coursera_Capstone/blob/master/toronto_part2.csv
This dataset is publicly available under MIT licence

Sample data

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.81153	-79.19552
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.78564	-79.15871
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.76575	-79.17520
3	M1G	Scarborough	Woburn	43.76820	-79.21761
4	M1H	Scarborough	Cedarbrae	43.76969	-79.23944
...	...	...	...	...	...
98	M9N	York	Weston	43.70357	-79.51645
99	M9P	Etobicoke	Westmount	43.69623	-79.52926
100	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.68674	-79.55729
101	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.74453	-79.58624
102	M9W	Etobicoke	Northwest, West Humber - Clairville	43.71174	-79.57918

103 rows × 5 columns