**WEEK 3 Capstone project: Battle of neighbourhoods**

**A: Introduction/Business Problem**

**Introduction:** Indians are among the largest immigrants to Canada. Generally people migrating to new place will look for a decently paid job , then the locality they live in as to what it offers such as security, frequency of public of transportation, grocery stores, restaurants fitness centres, schools etc.,

So, now this study aims to filter out a place for new migrants moving to Toronto who are strict about their physical routine, including swimming, and are looking to settle near by a grocery store or an Indian restaurant.

**Stake holders**: The study is aimed at new Immigrants moving to Canada from India, who are particular about their fitness and swimming routine as well as Indian groceries and restaurants nearby.

**Leveraging the Foursquare location data:** To locate a suburb nearer to the Greater Toronto Area (GTA) that has a recreation centre or a fitness centre with a public swimming pool that in it.

**Data:** Three datasets are used for the study. Data scrapping from Wikipedia to obtain the neighbourhood and obtaining required information from blogto.com, then the Foursquare API to enable the location data required for the study are employed.

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

<https://www.blogto.com/neighbourhoods/>

<https://foursquare.com/>

**B: Data**

**Data acquisition**

There are different sources from the data was collected for different purpose.

1. List of postal Codes for Canada:-

* Fetched the postal code of the neighbourhoods in Canada from Wikipedia.
* Link — <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

2. Geographical Co-ordinates:-

* A CVS file which contains latitude and longitude of the neighborhoods in Canada, Toronto, instead of geocoder is used, as the data sometimes is inconsistent while using live.
* Link for CVS — [http://cocl.us/Geospatial\_data](https://cocl.us/Geospatial_data)

3. Fetching Details of the venue:

* Foursquare API for fetching the details and location of the venues.
* Venue ratings as a benchmark.
* Visualization using Folium.
* From Foursquare API (<https://developer.foursquare.com/docs>),

4. Following data is retrieved the following for each venue:

* Name: The name of the venue.
* Category: The category type as defined by the API.
* Latitude: The latitude value of the venue.
* Longitude: The longitude value of the venue.
* Likes: Likes of the venue, that the user liked the restaurant.
* Rating: Rating of the venue.
* Tips: Tips given by the users.

**Data Cleaning**

Cleaning the Postal Code data

* The data frame will consist of three columns: Postal Code, Borough, and Neighborhood
* Only process the cells that have an assigned borough. Ignore cells with a borough that is ‘**Not assigned’.**
* More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that **M5A** is listed twice and has two neighborhoods: **Harbourfront**and **Regent Park**. These two rows will be combined into one row with the neighborhoods separated with a comma.
* If a cell has a borough but a ‘**Not assigned’**neighborhood, then the neighborhood will be the same as the borough.

Finally thus obtain data frame using above procedures, latitude and longitude coordinate are added, making it a clean data frame to be worked upon to find out the required locations and then using clustering ML to obtain the final result.