Battle of the Neighborhoods: Final Project

-Manasa Sriram

Introduction

This project is a part of the capstone project under the IBM Data Science Professional Certificate. It involves a hypothetical scenario where we determine where to start a hospital branch using the data on the Toronto Area. From the previous parts of the capstone project it was found that according to the map, the commercial neighborhoods of Toronto have a multitude of eateries and recreational spots. The number of pharmacies and general healthcare providers is very less. Hence, this project focuses on finding ideal areas or neighborhoods for setting up of new hospitals

Business Problem

Toronto being a very fast moving city with a lot of commercial spots, it becomes important to have adequate healthcare facilities. Suppose our stakeholders are a leading hospital chain who want to start a hospital branch in Toronto. Their main aim will be to find locations where there are no hospitals or very few hospitals so that they can cover a larger area of people to provide healthcare facilities to.

Data

To solve this problem, the following data will be required:

- List of neighgborhoods in Toronto, Canada
- Latitude and longitude of the neighborhoods
- Data on the venues present in the neighborhoods

We would be using the data to make a dataframe consiting of the venues along with their neighborhoods and the respective latitudes and longitudes. This dataframe will help us zero down on the location of hospitals which will help us proceed with our project.

Data Collection

- Scraping of Toronto Neighborhoods from the wikipedia page for the list of neighborhoods: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

(Note: Since it is a capstone project at a beginner level, only a small fraction of data related to neighborhoods of Toronto is being used)

- Getting the latitude and longitude of these neighborhoods from the Geocoder Package
- Using Foursquare API to get venue data related to these neighborhoods.
  (Prerequisites for the same include setting up a Foursquare account and generating a client ID and client secret)

Methodology

First, the data on the neighborhoods in Toronto, Canada was extracted from the above given wikipedia URL using BeautifulSoup. The data was then converted into a dataframe consisting of the columns

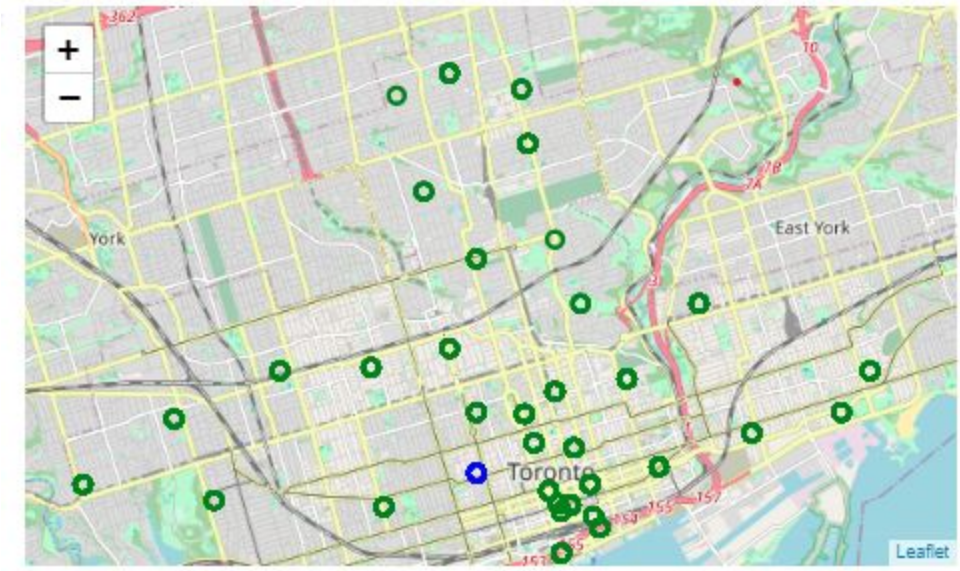
- PostalCode - The postal code of the area
- Borough - The bigger cluster/group the area belongs to
- Neighborhood - The name of the neighborhood

All the rows that had a 'Not assigned' value under both the Borough and Neighborhood columns was dropped. The resulting dataframe was then grouped by the Boroughs.

Using https://cocl.us/Geospatial_data, the latitude and longitude for each postal code was found. This dataframe was then merged with the previously created dataframe with the Neighborhoods and Boroughs. Now we have a dataframe with the postal code, neighborhood, borough, latitude and longitude of the areas in Toronto.

Using the folium and geocoders libraries, we visualise the dataframe in the form of an interactive map. Following this, data on venues was taken from Foursquare and merged with our dataframe using hot encoding. Using this new dataframe the most common venues in each neighborhood was found.

The machine learning concept used was KMeans clustering algorithm with 5 as the mean. It is one of the simplest and popular unsupervised technique and is also highly relevant for this project. The following map was generated.



Results

From the above map and the results it was seen that there was only one hospital, Toronto Western Hospital located in Kensington Market / Chinatown / Grange Park. This is depicted by the blue marking on the map. Thus the recommendation from this finding would be to start a hospital preferably in the northern or outskits of the city.

Discussion

From the project, the following observation was made

As the data used was very limited, there was only one hospital located by Foursquare. This means that out of the localities that have a postal code starting with M, there is only one hospital located. Had we taken the entire list of postal codes, there would have been a better and clearer understanding of the problem and a more accurate solution can be provided.

Conclusion

To conclude, this project is a working prototype for a larger problem. Using the same codes, we can use data of all the postal codes of Toronto, or any city for that matter to find out the distribution of venues and an accurate result can be obtained. The same procedure can be followed to make observations on any kind of venue in any location provided we have an updated map of the area.