IBM

Data Science Project

INTRODUCTION

This is a part of the capstone project for IBM Data Science Professional Certificate. In this project I am assuming that there are not enough Indian Restraunts in Toronto,Canada.A friend of mine wants to expand his family business of hotels in Canada as it might be an opportunity for making a good profit of it facilitating the South Asian community in Toronto.I am designing the project for finding a proper location for the restraunt given that it is quite clost enough to the main city.

BUSINESS PROBLEM

The objective of the project is to fing most suitable loation for the business organisation to open a new Indian Restraunt in Toronto,Canada.By using Machine Learning Algorithms like clustering and a good visulisation tool it is possible to answer the question : Which is a good location in Toronto,Canada to open a Indian Restraunt.

TARGET AUDIENCE

The project targets the organisation which wants to find a location to open a Indian Restraunt given the region.

DATA

To solve the above stated problem, we need the following data :

- Latitude and Longitude of the given location.
- Data related to the Indian restraunts.This will help us to find more suitable location to open the Indian restraunt.
- Neighborhoods in Toronto,Canada.

FETCHING THE DATA

- Scrapping the Wikipedia page of Toronto neighborhoods.
- Identifying the latitude and longitude data of the neighborhoods using GeoCoder package or uisng a predefined set of data pertaing to the same.
- Using Foursquare API to details related to the neighborhoods.

METHODOLOGY

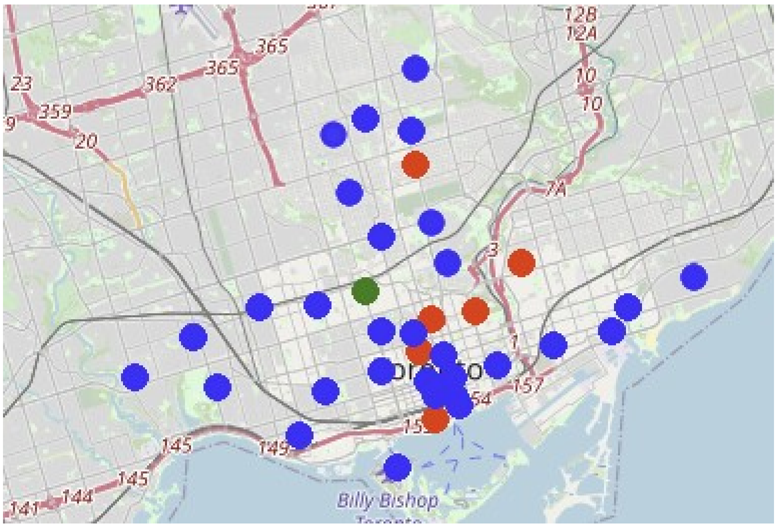
First,I collected the list of neighborhoods in Toronto,Canada.The data was extracted from the list of neighborhoods from Wikepedia : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M . The above mentioned page has a html table with postal code,borough and neighborhoods.I did the web scrapping utilising pandas HTML table scrapping method as it is convinient to pull tabular data directly from a web page into data frame.

However it is not possible to utilise Foursquare API using just neighborhoods and postalcodes from the data. I used the CSV file provided by IBM : http://cocl.us/Geospatial_data to match the co-ordinates of Toronto neighborhoods. Afer gathering these co-ordinates,I used folium pakage to visualise a map of these co-ordinates. Next, I used Foursquare API to pull the list of top 100 venues with 500 meters radius I have created a Foursquare developoer account in order to obtain the account ID and secret key to pull the data. Using Foursquare I was successful to pull out the name,category and geographical coordinates of the venue.I converted the dataframe to one-hot-vector on the basis of category and grouping it on neighborhood taking mean on frrequency of occurance of each venue category. This is to prepare the data for the clustering to be done later.

Since I am only interested in Indian Restaurants I decided to cluster the venues specifically for “Indian Restaurants” and “Neighborhoods” for clustering to avoid noise being generated from other categories for clustering process.I performed K-Means clustering method. This algorithm identifies k number of centroids and then allocates every point to it’s nearest centroid. A new cetroid is formed by taking mean distances between them. I have clustered the neighborhoods in Toronto into 3 clusters based on the frequency of “Indian Restaurants” in respective neighborhoods.Based on the results I am now here with a analytical solutin on location where to open a restauratn.

CONCLUSION

The result of K-Means algorithm shows that Toronto can be classified into 3 regions based on how many Indian restaurants are there in the neighborhood



- There are quite a large number of Indian Restaurants in Davisville, Riverdale, Cabbagetown, Central Bay Street region and all these neighborhoods appear to be in the heart of the city. Shown by Red markers.
- While the outskirts of the city lack Indian Restaurants shown by blue markers.

- There is one peculiar location where there is only one Indian Restaurant nearby The Annex,North Midtown,Yorkville. This is solo winner of Indian Restaurants in this region. Shown by green markers.

RECOMMENDATION

So starting a Indian Restaurant in region belonging to cluster 3 is the best option as it is in the heart of the city as well as have less competition and might yield more income over the outskirts of the city.