IBM

Applied Data Science Capstone Project

Week 4 Report

Introduction:

The Capstone project is for the Data Science certificate. From this project I am creating the scenario for the concept that there may not be enough in Indian restaurants in Toronto area. So, it might be a good opportunity for the enterprises with the Canada base. As the Indian food is popular enough amoung the world, so this enterprise might think of openings its business in are where Asian food community resides. with the use of open mind for finding the location to open a restaurant is open of the important decisions for the enterprise and I am designing this data science project to help them to find the perfect location using this project.

Bussiness Problem:

The main goal of this project is to find the perfect accurate location from the enterprise or the entreprener

To open a new Indian restaurant in Canada in Toronto Area.by the use of the data science tools with the K - means clustering algorithm this aims to provide the accurate solution to answer the question:

In Toronto, if an entrepreneur wants to open an Indian Restaurant, where should they consider opening it?

Target Audience:

Enterprice or enterperner who want an accurate location to open a Indian restaurant

DATA

To solve this problem, we will need below data:

- **List of neighborhoods in Toronto, Canada**
- **Latitude and Longitude of these neighborhoods**
- **Venue data related to Indian restaurants. This will help us find the neighborhoods that are more suitable to open an Indian Restaurant.**

EXTRACTING THE DATA

- **Scrapping of Toronto neighborhoods via Wikipedia**
- **Getting Latitude and Longitude data of these neighborhoods via**

Geocoder package
● Using Foursquare API to get venue data related to these
neighborhoods

Methodology:

First, I need to get the list of neighborhoods in Toronto, Canada. This is
possible by extracting the list of neighborhoods from Wikipedia:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
I did the web scraping by utilizing pandas HTML table scraping method as
it is easier and more convenient to pull tabular data directly from a web
page into the data frame.
However, it is only a list of neighborhood names and postal codes. I need
to get their coordinates to utilize Foursquare to pull the list of venues
near these neighborhoods. To get the coordinates, I tried using Geocoder
Package but it was not working so I used the CSV file provided by IBM

team to match the coordinates of Toronto neighborhoods. After gathering these coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues.

Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for "Indian restaurants".

Lastly, I performed the clustering method by using k-means clustering.

K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and

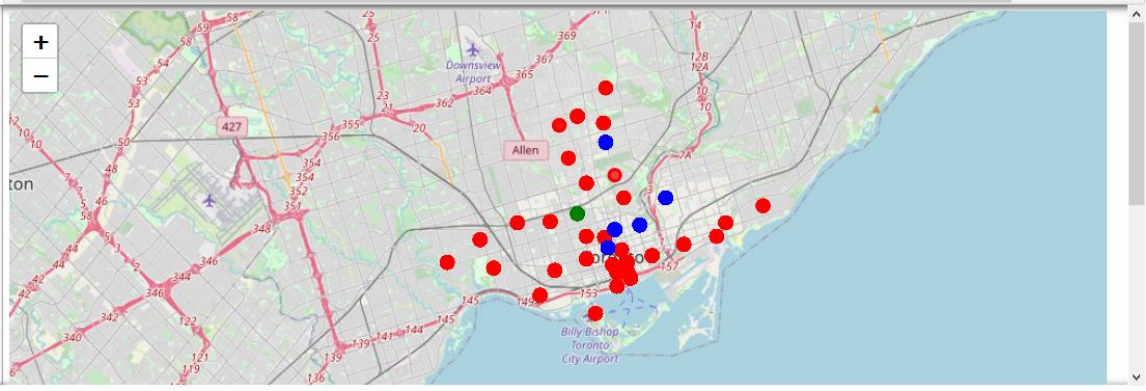
popular
unsupervised machine learning algorithms and it is highly suited for this
project as well. I have clustered the neighborhoods in Toronto into 3
clusters based on their frequency of occurrence for “Indian food”.

Result:

Cluster:

```
In [59]: map_clusters = folium.Map(location=[lat_toronto, lon_toronto],zoom_start=14)

         # set color scheme for the clusters


         # add markers to the map
         markers_colors={}
         markers_colors[0] = 'red'
         markers_colors[1] = 'blue'
         markers_colors[2] = 'green'
         markers_colors[3] = 'yellow'
         markers_colors[4] = 'cyan'
         markers_colors[5] = 'black'
         for lat, lon, cluster in zip(to_merged['Neighborhood Latitude'], to_merged['Neighborhood Longitude'], to_merged['Cluste

             folium.CircleMarker(
                 [lat, lon],
                 radius=5,

                 color =markers_colors[cluster],
                 fill_color=markers_colors[cluster],
                 fill_opacity=0.7).add_to(map_clusters)

         map_clusters
```

Out[59]:



Cluster0(green)-neighbourhood with less no of Indian restaurents

Cluster1(blue)-neighbour hood with no Indian restsaurents

Cluster 2(red) -neighbour hood with more o of Indian restaurents.

Recomondations:

Idian restaurants are in cluster 2 which is around
Central Bay Street, Church and Wellesley, Berczy Park, Union Station, Richmond,
lowest in Cluster 1 areas which are in North Toronto West and Parkade areas. Also, there are good opportunities to open near St James Town,Cabbagetown
Looking at nearby venues it seems cluster 0 might be a good location as there are not a lot of Indian restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Indian restaurant in these locations.