IBM
Coursera Capstone Project

Introduction

The objective of this project is to help entrepreneurs open up a restaurant in Toronto, Canada. As Indian cuisine is one of the most popular cuisine among Asians and considering that there are not enough Indian restaurants in Toronto, this project will help entrepreneurs to open up restaurants in places where there are very few or none.

By using data science methodology and various machine learning algorithms, the project intends to provide solution to the business problem which is: If an entrepreneur wants to open up an Indian restaurant, what will be the most considerable location?

Data Used

The data that will be used are as follows

1) List of all the neighbourhoods in Toronto
2) Locations of the neighbourhoods on the map
3) Data related to Indian restaurants.

Extracting the Data

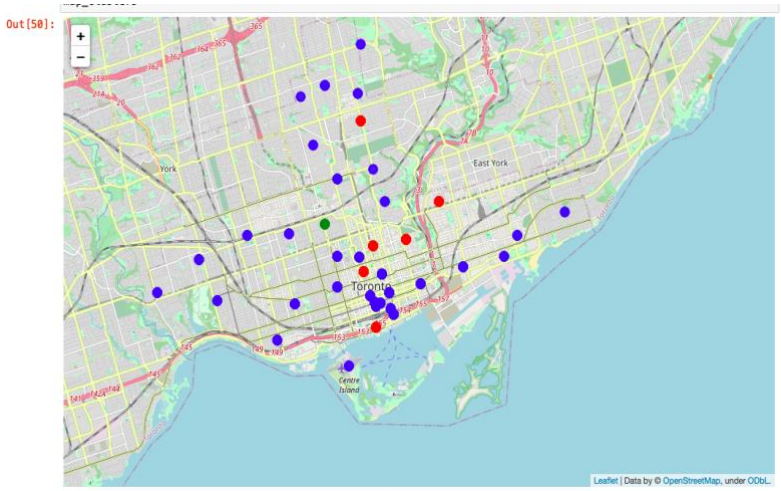
1) The details of the neighbourhoods in Toronto along with the codes from Wikipedia
2) Finding out the latitude and longitude coordinates of these neighbourhoods via Geocoder package
3) Using Foursquare API to get venue data related to the neighbourhoods.

Methodology

1) The list of neighbourhoods In Toronto is extracted from Wikipedia https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2) Pandas HTML table scrapping method is utilized to pull tabular data directly from a web page to a data frame.
3) Gathering of the coordinates of the neighbourhoods in Toronto and matching it with them

4) Creation of Foursquare account to get the data required. Foursqaure API is used to pull the Top 100 venues in 500 metre radius. Now the names, categories and coordinates of the venues can be found.
5) From this data unique categories can be found out in these neighbourhoods.
6) In order to perform clustering, the mean on the frequency of occurance of each venue is taken out by grouping rows in neighbourhood.
7) K Means clustering algorithm is used which identifies k number of centroids and then allocates each data point to the nearest cluster while keeping the centroids as small as possible. Clustering of neighboods in Toronto is done based on the frequency of Indian food.

Result



The result of K means clustering show us the neighbourhoods in 3 clusters:

1) Cluster 0: Neighbourhood with most number of Indian restaurants
2) Cluster 1: Neighbourhood with least number of Indian restaurants
3) Cluster 2: Neighbourhood with moderate number of Indian restaurants

Most of the Indian restaurants are in **Cluster 0** which is around Davisville, Church and Wellesley, Central Bay Street etc.

The lowest number of restaurants are present in **Cluster 1** which is near North Toronto west and Parkade area.

Cluster 2 has very limited number of Indian restaurants and might be a good location for an entrepreneur to start up an Indian restaurant.

Concluding the project would recommend an entrepreneur to start up an Indian restaurant in **Cluster 1** or **Cluster 2**