December 29,2019

MACHINE LEARNING TO FIND LOCATIONS TO OPEN AN INDIAN RESTAURANT IN TORONTO (IBM CAPSTONE PROJECT)

This is for my Final capstone project in Coursera

Introduction:

This project is intended to create a concept for a person who wants to explore opening a restaurant which would serve Indian breakfast and brunch in the area. Not many Indian restaurants serve breakfast in Toronto. So it might be a great opportunity to attract more business and he wants to choose a suitable location. He is looking for an area where there are less Indian Restaurants but more Asian Restaurants. This would attract more business since most of them don't serve breakfast.

The purpose of this project is to help him find a suitable location for his restaurant.

Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur to open an Indian restaurant in Toronto, Canada. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open an Indian Breakfast restaurant, where should they consider opening it?

Target Audience

The entrepreneur who wants to find the location to open authentic Indian restaurant which serves Breakfast.

Data

To solve this problem, below data was collected.

- 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 most suitable to open an Indian Breakfast restaurant.

Extracting the data

- Scrapping of Toronto neighborhoods via Wikipedia.
- Getting Latitude and Longitude data of these neighborhoods using csv file which is available.
- 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 page ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M").

- Retrieved the data by using urllib.request and then using BeauitfulSoup. Then used find command to table the values of Postcodes, Burroughs and Neighborhood. The tables were cleaned by removing "Not assigned" values and then grouped the neighborhoods based on Postcode and Burroughs.

- Another csv file which holds postcode, Latitude and Longitude values was used from capstone training project. Sorted the values based on postcode and I combined the values of two tables and created a data frame which has postcode, Burrough, Neighborhood, Latitude and Longitude values.

- After gathering all these coordinates, visualized the map of Scarborough was 1st row using Folium package to verify whether these are correct coordinates.

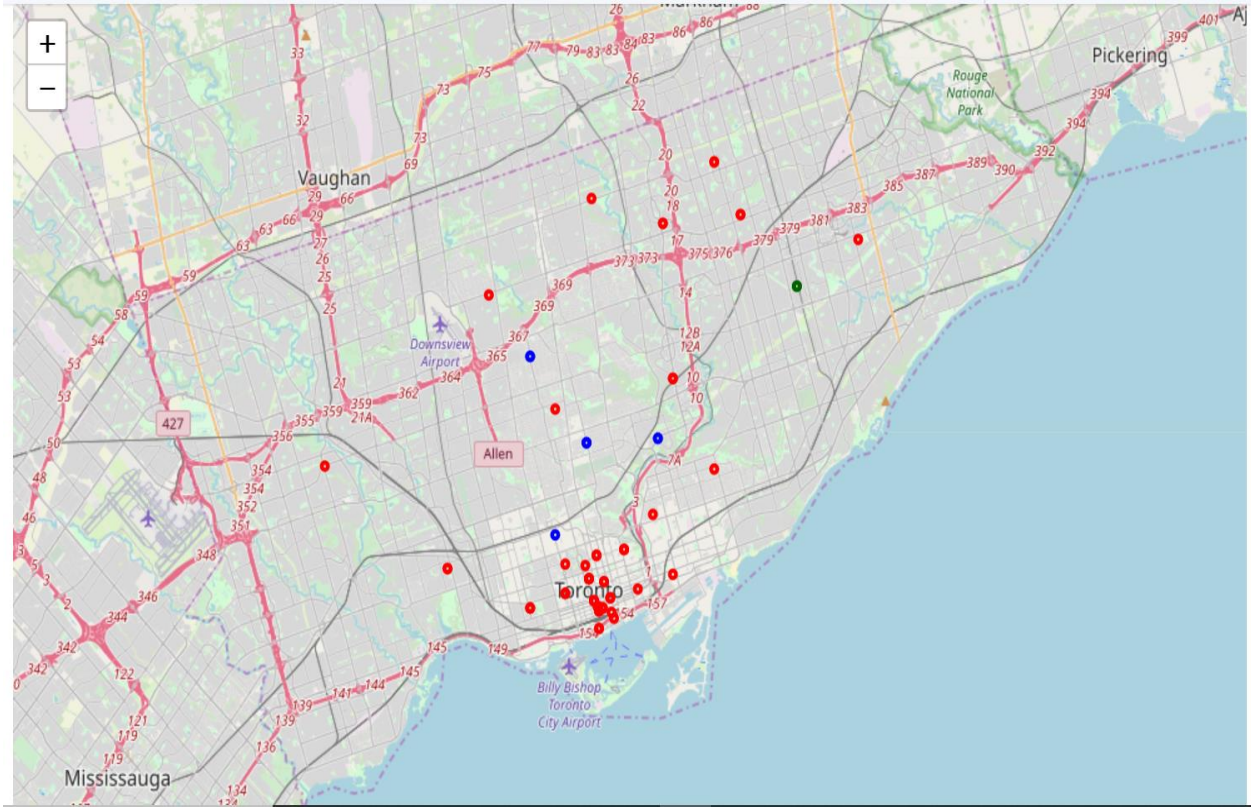
- Next, I used Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyzed 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. There were 276 unique venue categories.

- Here, I made a justification to specifically look for "Indian restaurants and Asian, Chinese and Thai Restaurants" .This would help me identify the neighborhoods where more people would be looking for Indian Food. It provides opportunity for a better business.

- 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". Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Results

Clusters:



The results from K-Means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Indian restaurants are in each neighborhood.

- Cluster 0 has neighborhoods with little Indian restaurants and more Asian Restaurants.
- Cluster 1 has the most Indian restaurants but they are all around 1 neighborhood area which is Dorset Park, Scarborough Town Centre, Wexford Heights.
- Cluster 2 has more Indian Restaurants than Cluster 0.

The results are visualized in map with Cluster 0 is represented by red color, Cluster 1 by dark green and Cluster 2 by blue color.

Limitations and Suggestions for Future Research

In this project my main consideration was to look for one factor: The existence of some restaurants related to Asian like (Asian, Chinese, Thai and Indian).My thought is that this would bring more Asian community into the area. Other factors which can be considered are Asian population density, rent in that area and other costs involved with initial setup. Considering the time for this project it is not feasible to take those additional factors into consideration.

But it can be covered as part of future expansion of this project.

Conclusion:

In this project, we have gone through the process of identifying business problem, specifying the data required, extraction and preparing data using FourSquare, performing machine learning algorithms using K-means clustering.

Finally we have provided recommendation to business users.

References:

List of neighborhoods in Toronto along with Latitude and Longitude

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

http://cocl.us/Geospatial_data

Four Square API

https://developer.foursquare.com/docs