Coursera Capstone Project : Applied Data Science

Abhinav Choudhary

Dayananda Sagar College of Engineering, India

abhinavchoudhary100@gmail.com

Introduction

- In this undertaking, I am making a speculative situation for the idea that there may not be sufficient Indian Restaurants in Toronto Area. In this manner, it may be an incredible open door for a business person who is situated in Canada. As Indian food is well known among the Asian people group, so this a business person may consider starting its business in zones where Asian network lives. In view of the reason, finding the area to open such a café is one of the most significant choices for this business visionary and I am structuring this task to assist him with finding the most reasonable area

Business Problem

- The target of this capstone venture is to locate the most reasonable area for the business person to open another Indian Restaurant in Toronto, Canada. By utilizing information science strategies and instruments alongside AI calculations, for example, bunching, this task intends to give answers to answer the business question: In Toronto, if an a business person needs to open an Indian Restaurant, where should they think about opening it?

Dat
a

To take care of this issue, we will require underneath information:

- List of neighborhoods in Toronto, Canada
- Latitude and Longitude of these areas
- Venue information identified with Indian eateries. This will assist us in finding the neighborhoods that are progressively appropriate to open an Indian Restaurant.

Getting Data

- Scrapping of Toronto neighborhoods through Wikipedia
- Getting Latitude and Longitude information of these areas through Geocoder bundle
- Using Foursquare API to get scene information identified with these neighborhoods

Methodology

Accuracy of the Geocoding API

In the initial development phase with OpenCage Geocoder API, the number of erroneous results were of an appreciable amount, which led to the development of an algorithm to analyze the accuracy of the Geocoding API used.

In the algorithm developed, Geocoding API from various providers were tested, and in the end, Google Maps Geocoder API turned out to have the least number of collisions (errors) in our analysis.

Folium

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the leaflet.js library. All cluster visualization are done with help of Folium which in turn generates a Leaflet map made using OpenStreetMap technology.

Methodology

One hot encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the K-means Clustering Algorithm, all unique items under Venue Category are one-hot encoded.

Top 10 most common venues

Due to high variety in the venues, only the top 10 common venues are selected and a new DataFrame is made, which is used to train the K-means Clustering Algorithm.

Methodology

K-means clustering

The venue data is then trained using K-means Clustering Algorithm to get the desired clusters to base the analysis on.

K-means was chosen as the variables (Venue Categories) are huge, and in such situations K-means will be computationally faster than other clustering algorithms.

Results

The neighbourhoods are divided into n clusters where n is the num- ber of clusters found using the optimal approach. The clustered neighbourhoods are visualized using different colours so as to make them distinguishable.

Figure:Neighbourhoods (Clustered).

Conclusion

- Most of the Indian restaurants are in around Central Bay Street, Church and Wellesley, Berczy Park, Union Station, Richmond, 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 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.