**Coursera IBM Data Science Capstone Project : Opening a new**

**Bermese Restaurant in Toronto**

 

**Introduction :**

For the Capstone project, we here are creating an imaginary scenario for a concept Burmese

restaurant which wishes to explore opening an authentic Burmese restaurant in Toronto area.

The concept behind this project is that there might not be enough Burmese restaurants in Toronto

and it may present a great deal of opportunity for the entrepreneur who is based in Canada.

As

Burmese food is very similar to other Asian cuisines, the entrepreneur here is thinking of starting

this restaurant in locations where Asian food is well known (aka many Asian restaurants in the

neighbourhood).

**Business Problem:**

The goal of this capstone project is to figure out the most suitable location for the entrepreneur

to start a new Burmese restaurant in Toronto, Canada. By using data science methods and

machine learning methodology such as clustering, this project aims to provide solutions to analysis

the business question: In Toronto, if an entrepreneur wants to open a Burmese restaurant,

where should they consider opening it?

**Target Audience:**

The entrepreneur who wants to find the location to open authentic Burmese restaurant

**Data :**

To solve this problem, I will need below data:

● List of neighbourhoods in Toronto, Canada.

● Latitude and Longitude of these neighbourhoods.

● Venue data related to Asian restaurants. This will help us find the neighbourhoods that

are most suitable to open a Burmese restaurant.

**Extracting the Data :**

● Scrapping of Toronto neighbourhoods via Wikipedia

● Getting Latitude and Longitude data of these neighbourhoods via Geocoder package

● Using Foursquare API to get venue data related to these neighbourhoods

**Methodology :**

First, I need to get the list of neighbourhoods 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 ”) 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 dataframe.

However, it is only a list of neighborhood names and postal codes. I will 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

all these coordinates, I visualized 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, we analyse each neighbourhood by grouping the rows by neighbourhood and

taking the mean on the number of occurrence of each venue category. This is to prepare

clustering to be done on a later date.

Here, I made a justification to specifically look for “Thai restaurants”. Previously, when I ran

the model, I was seeing for “Asian restaurants” but there are very few results (maybe due to

Foursquare categorization) so I looked for the restaurants closest to Burmese cuisine taste

(side note: Burmese food and Thai food are very similar in taste, so my justification is that if

there are people who enjoyed Thai food, they likely are going to enjoy Burmese food too!)

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

algorithm identify k number of centeriods, 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 “Thai 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 Thai restaurants are in each neighborhood:

● Cluster 0: Neighborhoods with little or no Thai restaurants

● Cluster 1: Neighborhoods with no Thai restaurants

● Cluster 2: Neighborhoods with high number of Thai restaurants

The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple

color and Cluster 2 in light green color.

**Recommendations :**

Most of Thai restaurants are in Cluster 2 which is around Adelaide, King, Richmond areas

and lowest (close to zero) in Cluster 1 areas which are North Toronto West and Parkdale

areas.

Also, there are good opportunities to open near Chinatown, St James town as the competition

seems to be low. Looking at nearby venues, it seems Cluster 1 might be a good location as

there are not a lot of Asian restaurants in these areas. Therefore, this project recommends the

entrepreneur to open an authentic Burmese restaurant in these locations with little to no

competition. Nonetheless, if the food is authentic, affordable and good taste, I am confident

that it will have great following everywhere :)

**Limitations and Suggestions for Future Research :**

In this project, I only take into consideration of one factor: the occurrence / existence of Thai

restaurants in each neighborhood. There are many factors that can be taken into consideration

such as population density, income of residents, rent that could influence the decision to open

a new restaurant. However, to put all these data into this project is not possible to do within a

short time frame for this capstone project. Future research can take into consideration of these

factors. In addition, I am relying on the existence of Thai restaurants only for this project but

future research can take into consideration of other variables such as existence of Asian

restaurants, Asian population level in each neighbourhood etc.

**Conclusion :**

In this project, we have gone through the process of identifying the business problem,

specifying the data required, extracting and preparing the data, performing the machine

learning by utilizing k-means clustering and providing recommendation to the stakeholder.

**References :**

List of neighbourhoods in Toronto:

https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M

Foursquare Developer Documentation: https://developer.foursquare.com/docs