IBM Data Science Capstone Project

Opening a Multi-cuisine Restaurant in Toronto

Submitted By: Zafreena Althaf

20th February 2020

Table of Contents

1.Introduction

Toronto is the provincial capital of Ontario and the most populous city in Canada. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. More than 50 percent of residents belong to a visible minority population group, and over 200 distinct ethnic origins are represented among its inhabitants. While the majority of Torontonians speak English as their primary language, over 160 languages are spoken in the city.

Such a diverse city would be the best place to open a Multi-cuisine restaurant.

1.1 Business Problem

The objective of this capstone project is to analyze and recommend the best neighborhoods in the city of Toronto to open a new Multi-cuisine restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: Where would you recommend a new investor to open a Multi-cuisine restaurant in the city of Toronto?

In this project we would be mainly looking at the below:

- Areas for large scope of shopping- Because common most people shop at least once every week and that is usually the same day they decide to dine out!
- Areas having restaurant venues- So that we can perform clustering on them and accordingly check for the concentration of restaurants in these areas.

1.2 Target Audience

This project is useful for any investors who are willing to open a new Multi-cuisine restaurant in the city of Toronto. All potential investors in the food industry, particularly in Toronto can draw insights from this project.

2. Data acquisition and cleaning

Here we will see what data we need, what sources we got the data from and how data was extracted and cleaned and made ready for analysis**.**

2.1 Data Required

To solve the problem, we will need the following data:

- List of neighborhoods in Toronto.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to areas with restaurants and shopping places.

2.2 Data sources

The Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of neighborhoods in and around Toronto, along with its corresponding boroughs and postal codes. We will use the web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and 'beautifulsoup' packages. Then we will get the geographical coordinates of the neighborhoods using https://cocl.us/Geospatial_data csv file which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Restaurants and shops category in order to help us to solve the business problem put forward.

2.3 Data extraction and cleaning

Web scraping of wiki page is done using the BeautifulSoup package and dataframe of three columns: PostalCode, Borough, and Neighborhood is created.

We then only process the cells that have an assigned borough. Ignore cells with a borough that is 'Not assigned**'.**

Since more than one neighborhood can exist in one postal code area we do a groupby of Postal code and Borough and group neighborhoods in same borough.

If a neighborhood is 'Not assigned' neighborhood, then the neighborhood will be the same as the borough.

We will then get the geographical coordinates of the neighborhoods using https://cocl.us/Geospatial_data csv file which will give us the latitude and longitude coordinates of the neighborhoods. Which we will then merge with our original dataframe based on postalcode and now our data is ready!

3. Methodology

Now that we have our data containing the neighborhoods and the coordinates ready we can plot the Map of Toronto with the neighborhoods superimposed on top. In this project we are looking mainly at boroughs having 'Toronto' and we use 'folium' to visualize our maps.

Using Foursquare API, captured a list of top 100 venues that are within a radius of 500 meters. To do so, I first registered a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. Then added made API calls to Foursquare passing in the geographical coordinates of the neighbourhoods and the Foursquare credentials in a Python loop. Foursquare returned the venue data in JSON format and extracted the venue name, venue category, venue latitude and longitude. With the data, checked how many venues were returned for each

neighbourhood and examine how many unique categories can be curated from all the returned venues.

First the data related to shopping venues in each neighborhood was extracted and we plotted a pie chart of the top 8 neighborhoods with highest number of shopping places.

Next, we analyzed each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. Since we are analyzing the "Restaurants" data, we will filter the "Restaurants" as venue category for the neighbourhoods.

Lastly, we performed machine learning technique of clustering on the data 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 is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Restaurants". The results will allow us to identify the concentration of restaurants in different neighborhoods.

4. Results

The results from the k-means clustering shows that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence of Restaurants:

Cluster 0: Neighbourhoods with low concentration of Restaurants
Cluster 1: Neighbourhoods with High concentration of Restaurants
Cluster 2: Neighbourhoods with medium concentration of Restaurants

Also according to our pie chart the neighborhoods: (Ryerson,Garden District) , (Adelaide,King,Richmond) and (Harbourfront East,Toronto Islands,Union Station) all come under the top 8 largest shopping neighborhoods and these three also fall under cluster 2)

The neighborhoods: (Design Exchange,Toronto Dominion Centre),(Commerce Court,Victoria Hotel) and (Harbord,University of Toronto) are also under the top 8 largest shopping neighborhoods and belong to cluster 1

5. Discussion

Cluster 0 has low concentration of restaurants meaning it might not be a good place for profitable restaurants.
Cluster 1 has high concentration of restaurants which means competition will be very high
Cluster 2 has a medium concentration so it would be ideal for restaurants.

Also families spent their weekdays and especially weekends shopping and that is also usually the same day they dine out! So we should look for neighborhoods with lot of shopping possibilities.

As per our observation the top 3 ideal neighborhoods to start a multicuisine restaurant would be (Ryerson,Garden District) , (Adelaide,King,Richmond) and (Harbourfront East,Toronto Islands,Union Station) as they are one of the top largest shopping neighborhood and also fall under cluster 2.

6. Conclusion

To start a multicuisine restaurant we should consider the neighborhoods (Ryerson,Garden District) , (Adelaide,King,Richmond) and (Harbourfront East,Toronto Islands,Union Station). The final decision will depend on in depth analysis of these 3 neighborhoods as well as many other factors such as financing, operational agreements & business-related terms and conditions agreed upon by all the stakeholders involved.

7. References

- Beautiful Soup documentation - https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- Pandas documentation - https://pandas.pydata.org/pandas-docs/stable/
- Folium documentation - https://python-visualization.github.io/folium/
- Wiki: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M