Capstone Project - The Battle of Neighbourhoods

Finding a suitable location to open an Italian restaurant in the city of Toronto - using machine learning (K means)

by

Ashwin Kumar Padur Sampath Kumar

I. Introduction

The purpose of this project is to use machine learning (K means) to analyse a set of localities, and the businesses functioning in that area, how well these businesses do (Data from foursquare), and based on the analysis provide a hypothetical recommendation to an entrepreneur who is willing to establish a business.

Toronto is the second largest financial centre in North America. The economy of Toronto has had a GDP growth rate of 2.4 percent annually since 2009, outpacing the national average.
As the capital of Canada's most populous province, the city has a widely diversified economy.
That gives us a picture of how dense the state could be, leading it to be one of the most competitive of places for one to startup and run a business, especially a restaurant.

And hence, as part of this project we will be exploring the feasibility of an entrepreneur opening and running an Italian restaurant in Toronto, a great business hub with the myriad of choices in cuisine.

II. Business Problem

The business problem is broken down as questions

1. Locate all the good places that have Italian restaurants in Toronto
2. Where exactly is the utmost demand for Italian food?
3. Which places are scarce of Italian restaurants?
4. And finally where in Toronto can a new Italian restaurant do well?

III. Data section: Source and method of extraction

I. **Source**: List of Toronto's Boroughs, neighbourhoods and their postcodes from - h*ttps://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.*
Method: By scrapping information from Wikipedia using the requests function.

II. **Source**: Geographical coordinates of Toronto's Boroughs, neighbourhoods - *http://cocl.us/Geospatial_data.*
Method: By retrieving information from the URL directly and saving it as a .CSV file.

III. **Source***:* Venues data from Foursquare
Method: Retrieving information using functions and developer credentials.

IV. Methodology:

In order to achieve the desired purpose of this project, a number of steps are performed sequentially, beginning with accessing the list of neighbourhoods and boughs of Toronto city, from Wikipedia page (Refer Data section), using the page requests function. This data is in a raw format and hence it is cleaned by removing blanks and NA columns, which may hinder with our analysis. Further the data is also sorted to group neighbourhoods. We use data frames to save/capture this data, for ease of handling, in the next steps.

Since the data retrieved from Wikipedia, only contains names of neighbourhoods and boroughs, we need further data, from another source, in order to locate these areas in a map with their geographical coordinates. To do this we access the geospatial data from a web page (Refer Data section), and capturing in the same in a data frame, again fro ease of handing in the next steps. once we have this data the final step in the process of getting Toronto data is to merge the respective data frames contains names of neighbourhoods, boroughs and their geographical coordinates. The resulting data frame is grouped by boroughs to get the count of number of boroughs in individual neighbourhoods.

Once the data is ready, the next step is to understand how the neighbourhoods are placed geographically, hence we visualise it using the folium library and adding markers for all neighbourhoods. The resulting map gives us an idea of location, necessary for us to proceed further, i.e. finding a places that have Italian restaurants.

In order to achieve this, we interact with Foursquare using the API, the developer credentials and more importantly the radius of a particular search location, around which we want the information retrieved. We create a function to get nearby venues. We then count the neighbourhoods by their nearby venues, and list all unique venues, to check if there are Italian restaurants.

Finally, we look for Italian restaurants and the neighbourhoods they are located in. Further, we assign cluster labels to all line items in the dataframe,

to sort data to suit our requirement of clustering neighbourhoods. Then we use K means clustering algorithm to segment neighbourhoods in order to achieve the prime objective of the project, i.e. to understand which neighbourhood is suitable for a new Italian restaurant.

V. Results:

The results of this project are achieved by clustering the neighbourhoods of Toronto. We chose to segment the neighbourhoods into 3 and hence there are 3 clusters, namely Cluster 0, Cluster 1, Cluster 2.

Analysing the 3 clusters achieved using the K means clustering algorithm, listing the different business in each cluster.

Cluster 0: Restaurants, coffee shops, aquariums, Hotels, Pubs, Bagels shop, Jazz club, Cheese shop and a clothing store.

Cluster 1: Restaurant, Chinese restaurant, Market, Pharmacy, Pizza place, Restaurant, Grocery store, Bakery, Coffee shop.

Cluster 2: Bookstore, Juice bar, Salon/barbershop, Restaurant, Park, Pizza place, coffee shop, seafood restaurant, Hotel, Coffee shop.

Discussion:

Cluster 0 areas - Berczy Park, Harbourfront East, Union Station, Toronto Islands,Garden District, Ryerson etc., - contains a mix of businesses such as Restaurants, coffee shops, aquariums, Hotels, Pubs, Bagels shop, Jazz club, Cheese shop and a clothing store. However, do not have any Italian restaurants.

Cluster 1 areas - St. James Town, Cabbagetown, The Danforth West, Riverdale etc., - contains Restaurant, Chinese restaurant, Market, Pharmacy, Pizza place, Restaurant, Grocery store, Bakery, Coffee shop. There is Pizza place in this area, which could pose a threat/competition to a new Italian resutaurant.

Cluster 2 areas - Church and Wellesley, Richmond, Adelaide, King, Kensington Market, Chinatown, Grange Park etc., contains Bookstore, Juice bar, Salon/barbershop, Restaurant, Park, Pizza place, coffee shop, seafood restaurant, Hotel, Coffee shop. There is Pizza place in this area, which could pose a threat/competition to a new Italian resutaurant.

Conclusion¶

Italian restaurants or Italian cuisines are available in nearly all neighbourhoods. The optimal locations to start one are Church and Wellesley, Berczy Parkwhere, Harbourfront East, Union Station, Toronto Islands as there are no Italian restaurants or Italian cuisines.
On the other hand Richmond, Adelaide, King is also a viable option considering there is Hotel nearby, hosting guests other than just walk in crowd. Also there is a Pizza place in Richmond, Adelaide, King, St. James Town, Cabbagetown, The Danforth West, Riverdale meaning there is demand for Italian already and a new restaurant will present as a good option for consumers.