Capstone Project – The Battle of the Neighborhoods
Report

Applied Data Science Capstone by IBM/Coursera

Author: Maia Ludmila Budziñski

Date: July 01, 2020

Introduction

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2.731.571 in 2016. Current to 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA. Toronto is the fastest growing city in North America. and is the anchor of an urban agglomeration, known as the Golden Horseshoe in Southern Ontario, located on the northwestern shore of Lake Ontario.

Toronto encompasses a geographical area formerly administered by many separate municipalities. These municipalities have each developed a distinct history and identity over the years, and their names remain in common use among Torontonians. Former municipalities include East York, Etobicoke, Forest Hill, Mimico, North York, Parkdale, Scarborough, Swansea, Weston and York. Throughout the city there exist hundreds of small neighborhoods and some larger neighborhoods covering a few square kilometers.

Having such vast population and immense geographical area, there also exists big rivalries between different businesses. Therefore it became very challenging for stakeholder to decide in which area should their start a business to get higher revenue with the lowest possible competition.

Business Problem

In this project we will try to find an optimal location for a restaurant and identify the ideal type of food it should serve keeping in mind the competitors. Specifically, this report will be targeting stakeholders interested in opening a restaurant in Toronto, Canada.

Since there are lots of restaurants in Toronto, we will try to detect locations that are not too crowded already with restaurants but also show a prosperity for this kind of business.

By using data science methods and tools along with machine learning algorithms such as clustering we will identify the most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Target audience

Stakeholders who wants to open a restaurant in Toronto, Canada.

Data

Regularly spaced grid of locations, surrounding city center, was used to define the different neighborhoods.

Data sources used:

- List of neighborhoods in Toronto, Ontario, Canada

- Latitude and Longitude of these neighborhoods

- Number of restaurants and their type and location in every neighborhood will be obtained using Foursquare API

Extract/generation of required information was performed as follow:

- Scrapping of Toronto neighborhoods via Wikipedia

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

- Getting Latitude and Longitude data of these neighborhoods via Geocoder package

- Using Foursquare API to get venue data related to these neighborhoods

Methodology

The aim of this project was to find best location to open a new restaurant and determine the type of food it should serve based on competition of different locality in Toronto.

First, a list of neighborhood names and postal codes of Toronto was obtained from Wikipedia:

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

Since coordinates were needed to utilize Foursquare to pull the list of venues near these neighborhoods, Geocoder was used to obtain them. After gathering these coordinates, Foursquare API was used to pull the list of top 100 venues within 500 meters radius. For this propose, a Foursquare developer account was created in order to obtain a Client ID and Client Secret to pull the data. From Foursquare, I was able to pull the names, categories, latitude, and longitude of the venues. With this data, I could also check how many unique categories contained 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 last was to prepare clustering to be done later.

Lastly, I performed a clustering analysis using K-means clustering. K-means clustering algorithm is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are the centroids of the K clusters, which can be used to label data.

With an assumption of 7 clusters, K-means clustering algorithm come up with 7 different clusters for Toronto neighborhoods, with similar set of Venues. Each cluster was explored and the discriminating venue categories that distinguish each one was determined. Clusters and Boroughs/Neighborhoods with Maximum number of restaurants and their types were identify. Based on these project results, recommendations about the ideal location to open a restaurant in Toronto were made.

Results and Discussion

We have 4 boroughs and 72 neighborhoods inside geographical coordinate of **43.6534817, -79.3839347.** The results from K-means clustering show that we can categorize Toronto neighborhoods into 7 clusters based on similar set of Venues.

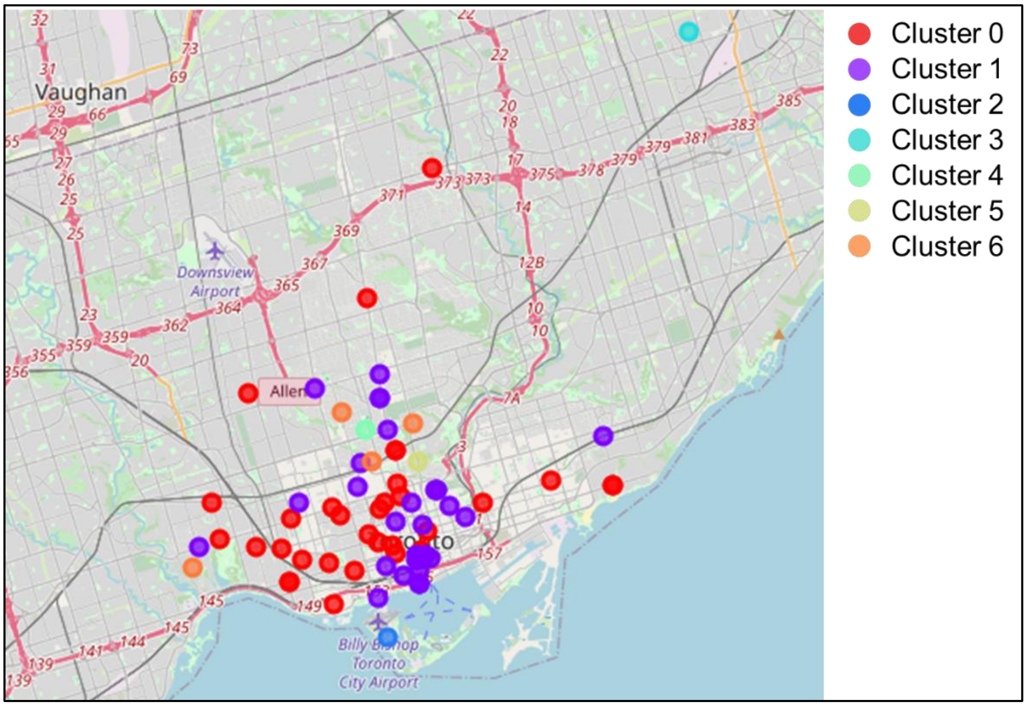

Figure1: Map of Toronto, Ontario, Canada visualizing K-means clustering results.

These 7 clusters were grouped into 3 different Sets, based on the number of restaurants as Popular Venues in neighborhoods:

- Cluster 0-1: Neighborhoods with more number of restaurants.

- Cluster 2-5: Neighborhoods with no restaurants.

- Cluster 6: Neighborhoods with less number of restaurants.

If we assume that the cluster with maximum number of restaurants will have the best possibility to have a new restaurant due to the need in the area, neighborhoods contained in Cluster 0 and Cluster 1 would be our best choices.

A more deep analyze shows that the number of restaurant in Cluster 0 is higher than the one in Cluster 1. Also, while the 1st Popular Venues in Cluster 0 include Indian, Sushi, Tibetan, Vietnamese, Italian and Korean restaurants Cluster 1 only report Sushi restaurants. Exploiting this lack of variety in popular restaurants could be convenient in Cluster 1 neighborhoods. Since restaurants in 2nd and 3rd Popular Venues of Cluster 1 include Italian, Sushi and Thai restaurants, might be also convenient to invest in a restaurant that serve Vietnamese, Korean or Japanese food, since would encounter less competition and would serve plates with ingredients already popular among Cluster 1 neighborhoods. Suggested neighborhoods from Cluster 1 include: Toronto Dominion Centre, Runnymede, Church and Wellesley, Harbourfront, St. James Town, etc.

Never the less, recommended zones should be considered only as a starting point for more detailed analysis which could eventually result in a location which has not only no nearby competition but also possess other convenient conditions.

Conclusion

The purpose of this project was to identify areas in Toronto with prosper conditions to set restaurants and identify the ideal type of food it should serve in order to aid stakeholders in narrowing down the search for optimal location for a new business. By calculating restaurant density distribution from Foursquare data we have first identified general boroughs that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants, that might be seen as competition. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking also into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.