Capstone Project - The Battle of Neighborhoods

Coursera IBM Data Science Certification

RISHI K S

July 2020

Report Contents

INTRODUCTION

In this capstone project as a part of IBM Certification, I am creating a hypothetical scenario for a concept when there are a lot of American Restaurants in Toronto Area. It will be of high competitive market, for a new entrepreneur who is based in Canada to find the best location to open a New American restaurant or a chain of restaurants.

As American food is the most popular in Canada, the level of competition is very high. I am hereby designing this project to help the entrepreneur to find the most suitable location in Toronto.

BUSINESS PROBLEM

The objective of this capstone project is to find the most suitable location for the entrepreneur to open an American Restaurant in Toronto, Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open an American Restaurant, where should they consider opening it?

TARGET AUDIENCE

The entrepreneurs who wants to find the location to open an Exclusive American restaurants.

DATA

To solve this problem, we will need below data:

- List of neighborhoods in Toronto, Canada
- Latitude and Longitude of these neighborhoods
- Venue data related to American restaurants. This will help us find neighborhoods that are more suitable to open an American Restaurant.

EXTRACTING THE DATA

- The scrapping of Toronto neighborhoods via Wikipedia
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

METHODOLOGY

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia: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 webpage into the data frame.

However, it is only a list of neighborhood names and postal codes. I 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 these coordinates, I visualize 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, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

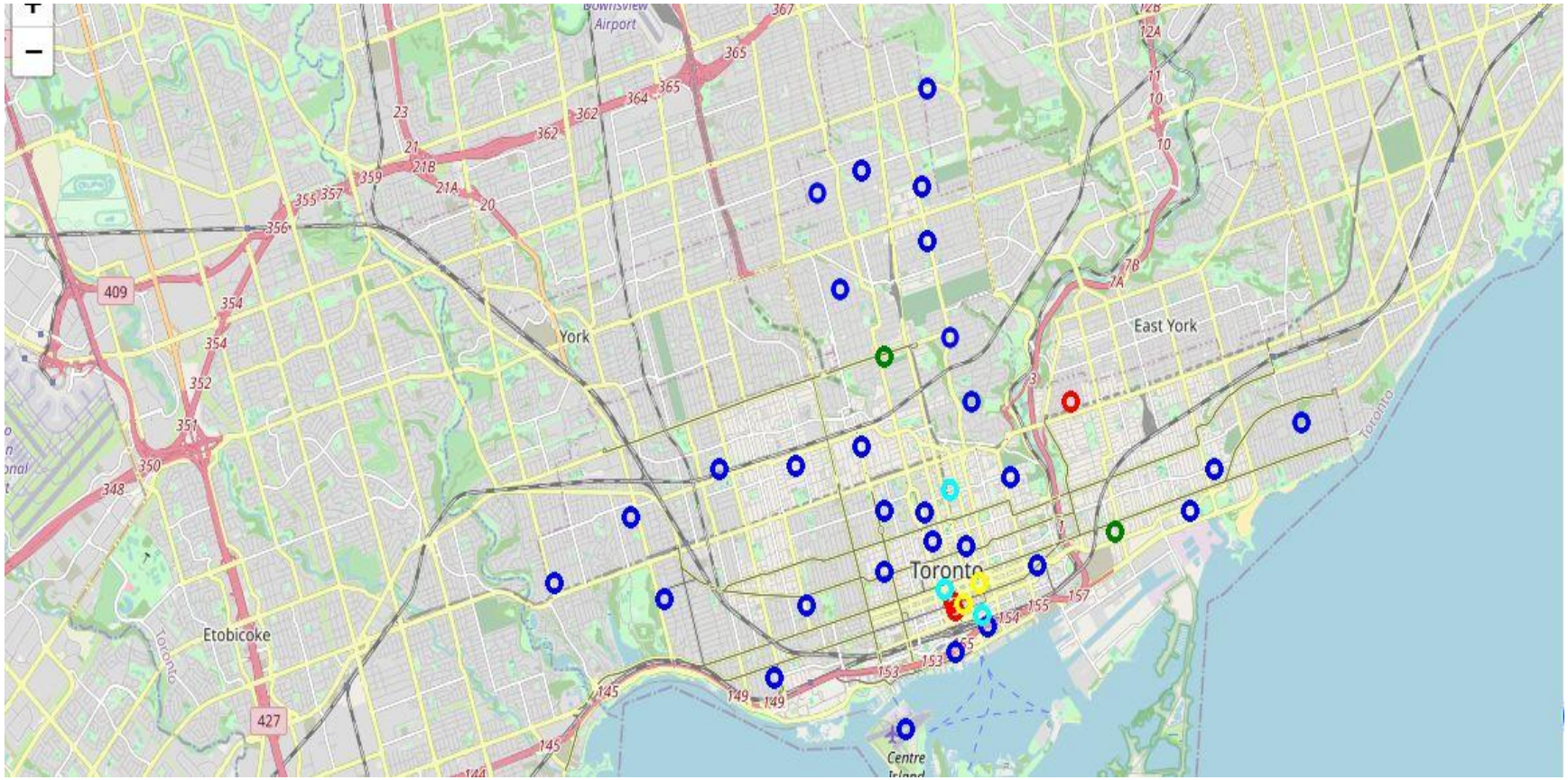
Here, I made a justification to specifically look for “American restaurants”. Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 5 clusters based on their frequency of occurrence for “American food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

RESULT

CLUSTERS

The results from k-means clustering show that we can categorize Toronto neighborhoods into 5 clusters based on how many American restaurants are in each neighborhood:

DISCUSSION



- Cluster 0: Neighborhoods with good number of American restaurants.
- Cluster 1: Neighborhoods with no American restaurants.
- Cluster 2: Neighborhoods with a more number of American restaurants
- Cluster 3: Neighborhoods with a more number of American restaurants
- Cluster 4: Neighborhoods with a few number of American restaurants

CONCLUSION

The results are visualized in the above map with Cluster 0,Cluster 1, Cluster 2.

Most of the American restaurants are in cluster 0,cluster 3 and cluster 4 which is around St.James town, Stn A PO Boxes, Church and Wellesley, Richmond, Adelaide, King,Commerce Court, Victoria Hotel,First Canadian Place, Underground city,Toronto Dominion Centre, Design Exchange. areas.

Looking at nearby venues it seems cluster 1 might be a good location as there are no American restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic American restaurant in these locations with minimum restaurants.