Capstone Project - The Battle of the Neighborhoods (Week 2) : Report

Applied Data Science Capstone by IBM/Coursera

Table of contents

Introduction: Business Problem

In this project we will try to find an optimal location for a new business. Specifically, this report will be targeted to stakeholders interested in opening a **point of service (shop, bar, café)** in **Toronto**, Canada.

Here we will try finding if someone wants to open a new point of service in the city which location is best suited for it keeping in mind the competitors and which income group of people will be attracted most to it based on the **population of the neighbourhood**.

Since there are lot of places for buys in Toronto, we will try to detect **locations that are not already crowded with shops**. We would also prefer locations **as close to city center as possible**, assuming that first two conditions are met, because downtown its most populated and has good transport infrastructure,

We will use our data science powers to generate a few 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.

Data

Based on definition of our problem, factors that will influence our decission are:

- All existing venues in the neighborhood (any type of shops)
- Age group of people with their income
- Distance of neighborhood from city center

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

- centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- number of venues and their type and location in every neighborhood will be obtained using **Foursquare API**

Methodology

The main moto of this project is to find best location to open a new business in Toronto, Canada based on competition in different locality and their population.

So, to do this I have used 2 different data sets available as mentioned above. Those 2 data set contains Locality information of Toronto, different age group of people in the people, population.

To solve the problem I am going to use “K-Means Clustering Algorithm ". K-means clustering 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 pointsare 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 new data Labels for the training data (each data point is assigned to a single cluster)

Also, I will be utilizing different maps in-order to give a clear vision to the target audience.

Steps we took for the analysis:

- Collected required **data: location and type (category) of every venue within our lat and lng. We have also the type of venue in particular locality.
- Explored the **venue density**' across different areas of Toronto - we will use **K- mean** to identify a few promising areas close to center with low number of venues and their type.
- Explored the most promising areas and within those create **clusters of locations that meet some basic requirements** established in discussion with stakeholders: we will take into consideration locations with **less venues in radius of 500 meters**, We will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to explore neighborhood.

Analysis

Data identification, capturing and cleaning.

Search & Identify the relevant data source and capture it, here we are using Wikipedia to get data about Toronto, Canada. Then we remove all the redundant value (data cleaning). Then we combine neighborhood. Now the data is clean and ready to use.

Combining different data source and sorting neighborhood based on Longitude and latitude

Now, we will combine neighborhood dataset with postal address and dataset with Latitude & Longitude and save them it separate data frame. The resultant data frame with contain details about Postal code, Borough, Neighborhood, Latitude & Longitude. Then visualize it using folium map.

Explore the Toronto's neighborhoods

Firstly, we explored all the neighborhoods in the city of Toronto, using the Latitude & Longitude data, usin g Foresquare API to get the Venue venues available in Toronto. Explore the unique categories in the neighborhood. Filter the Venues details for all possible 'Venues'. Find each neighborhood along with the top most common venues. Identify the top 10 venues for each neighborhood.

Clustering

With an assumption of 10 clusters, use K-Cluster algorithm to come up with 10 different clusters in Toronto with similar set of Venues. Explore each cluster and determine the discriminating venue categories that distinguish each cluster. Identify the clusters & Boroughs/Neighborhoods with Maximum nu mber venues and their types.

Results and Discussion

Our analysis shows that although there is a great number of venues in Toronto, there are pockets of low venue density fairly close to city center. We have 4 boroughs and 74 neighborhoods inside geography coordinate of **43.653963, -79.387207**.

Based on our initial assumption of the cluster with maximum number of venues will have the best possibility to have a new venue due to the need in the area. Based on the resultant clusters it lo oks like Cluster 1 and Cluster 10 have higher number of venues then rest of the clusters.

It is entirely possible that there is a very good reason for small number of venues in any of those areas, reasons which would make them unsuitable for a new venue regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

Conclusion

Purpose of this project was to identify areas Toronto with low number of points of service in order to aid stakeholders in narrowing down the search for optimal location for a new business. By calculating venues 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 venues. Clustering of those locations was then performed in order to create major zones of interest (containing gre atest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders. Based on the work carried out, we see that the city center is full of institutions providing services.

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