VenueFinder API for Toronto City

Karan Pillay

04/26/2021

1. **Introduction**

   VenueFinder allows you to find all the venues located in the city of Toronto. Explored neighborhoods of Toronto city with the help of Foursquare API and later used this feature to group the neighborhoods into clusters. Used k means to complete this task. Finally used Folium library to visualize the neighbors in the Toronto city and their emerging clusters.

2. **Business Problem**

   One of the most difficult parts of planning your wedding, birthday parties or any celebration is choosing the right venue. There are many factors which can impact the search like budget, number of guests, aesthetics, and accommodations. There are so many things to consider, and people to please that what used to be a fun experience has turned into more of a daunting task. Therefore, the primary aim of VenueFinder is to effortlessly find and recommend venues in Toronto city.

3. **Data Acquisition and Data Cleaning**

   The dataset used in this project is the Toronto neighborhood data, a Wikipedia page exists that has all the information needed to explore and cluster the neighborhoods in Toronto based on other demographics. Firstly, the dataset was scarped from Wikipedia page using beautiful soup object later it was wrangled, cleaned, and structured into a pandas dataframe. Also used a geospatial_coordinates.csv which includes latitudes and longitudes data as geocoder package was not responding. Finally merged the two datasets to obtain a final dataset which contains the following columns.

a) PostalCode
b) Borough
c) Neighborhood
d) Latitude
e) Longitude

4. Exploratory Data Analysis

In data cleaning process and manipulation process observed that the final_df contains 103 rows and 5 columns. Dropped ‘Postal Code’ column as it was similar to ‘PostalCode’ later checked for any NULL values in the dataset using isnull().sum() function in pandas fortunately there were no NULL values. On further analysis found that there are 15 Boroughs and 103 Neighborhoods in Toronto city.

5. Methodology

In this project we will direct our efforts on detecting the venues in the Toronto City. We will be using folium that is built on the data wrangling strengths of the Python ecosystem and the mapping strengths of the leaflet.js library. It helps in manipulation of data in Python, then visualize it in on a Leaflet map via folium, we will use it to display venues in neighborhood of Toronto city. In the second step of our analysis, we will use Foursquare API to explore neighbors and segment them based on the venues across different locations in the Toronto city. In the final step we will use k means clustering algorithm to cluster venues based on locations and display top 5 venues to be selected from the list of venues in Toronto city.

Out[16]:

Table 1. Map of Toronto City using Folium

Out[62]:

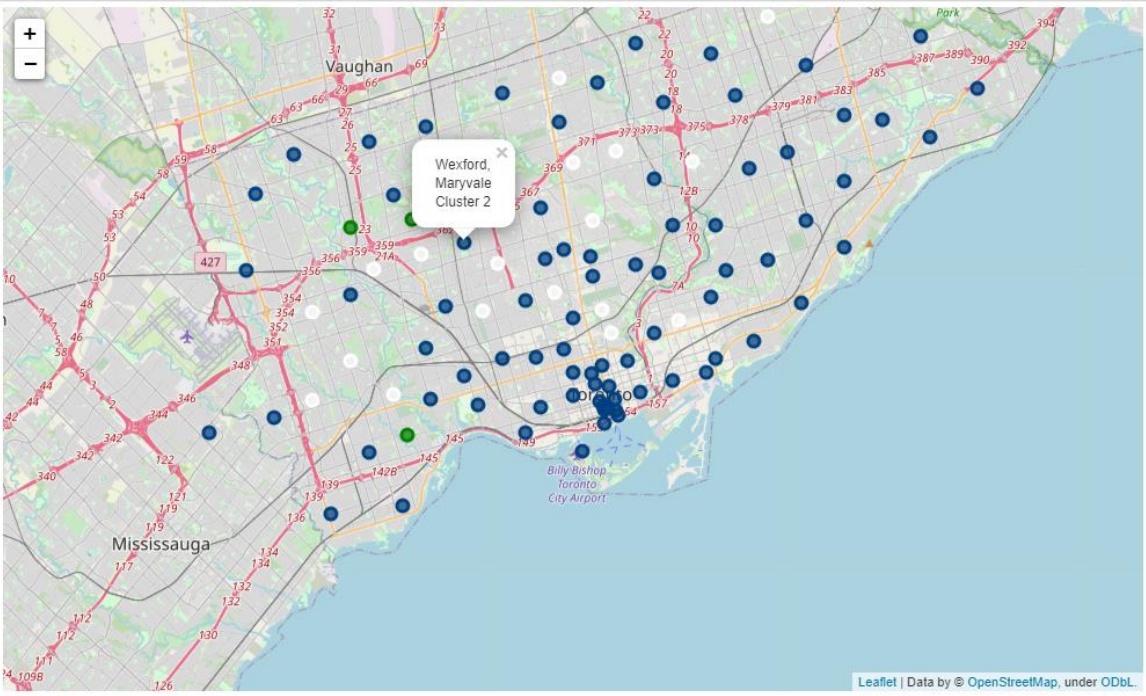

Table 2. Clustering using kmeans

6. Results

Our analysis shows that there are majority of venues clustered in the North York are which belongs to cluster 2, the most common venues in all 3 clusters are restaurants. There are only 3 clusters in cluster label 1 which shows the first most common venue of basketball field. In cluster label 0 the most common venue is park. We can observe from the above map how the venues are clustered from each other based on distance metrics. You can also observe there are null values which are replaced by 0 for cluster label 0 that hinders efficiency of our kmeans model. It is evident from the clustering model that the clusters generated could have been better. It can be improvised using a greater k value. We can also observe that there are majority venues in North York neighborhood in city of Toronto.

Using Foursquare 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, also searched for the top 5 neighborhood venues.

Purpose of this analysis was to only provide info on venues in Toronto city - it is entirely possible that there is a very good reason for absurd recommendation of venues like Dog Run in any of those areas, reasons which would make them unsuitable for venue selection. 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 considered, and all other relevant conditions met.

7. Conclusion

The purpose of this project was to identify venues in the neighborhood of the city of Toronto, we used web scraping to scrap the data from Wikipedia then clean, manipulate and structure it into usable format. Folium made it easy to plot and visualize the map of Toronto city and later the clusters too. In modeling step, we used kmeans which is a unsupervised classification

algorithm forming 3 clusters of venue data points. Observed that majority of the venues are in the North York which is in downtown region. Using Foursquare 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.

Final decision on optimal restaurant 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 (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.