Coursera IBM Data Science Professional Certificate

Applied Data Science Capstone

By

Dharani Ravichandran

Contents

Introduction

Using the data gathered in the previous week work, there are a list of questions we can answer with the clustering algorithm.

Business Problem

Identifying boroughs in Toronto which are very desirable and similar to the downtown Toronto. This is to explore the restaurant kinds that has an optimal chance of success. The location for the restaurant and venue to choose for successful operations

Target audience

The principal stakeholders targeted are business explorers and entrepreneurs. The location choosing suggestions for explorers and tourists.

Data Section

Foursquare location data is leveraged to explore and compare the different boroughs. Foursquare is used to gather and group live data on the most popular food venues per borough. Foursquare developer license, the Client ID and Client Secret, it is possible to make API requests in order in order to retrieve venue information. By feeding a function with borough name and its geometric coordinates, using Foursquare API different venues in the category Food are extracted. After performing one-hot encoding and grouping together the rows by borough, the different dataframe are combined into a single dataframe (with non-numeric values removed) in order to perform the clustering operation.

Data cleaning

Datasets gathered for the analysis are in csv format and loaded into Jupyter Notebook, which is used for the analysis. NaN values substituted.

Feature selection

There are 17 unique venue categories in the Neighborhood captured. Each of which can be grouped and given weights to decide if it is the interesting place to reside or start a business.

1. Bus Line
2. Check Cashing Service

3. Convenience Store
4. Discount Store
5. Fast Food Restaurant
6. Field
7. Grocery Store
8. Hockey Arena
9. Market
10. Park
11. Pharmacy
12. Pizza Place
13. Restaurant
14. Sandwich Place
15. Tennis Court
16. Trail
17. Women's Store

Methodology

K-means Clustering is used for this analysis. It is a simple unsupervised machine learning algorithm that groups a dataset into a user-specified number (k) of clusters. Steps followed

Importing necessary libraries:

Used the Folium library to generated maps to visualize neighborhoods on and how they cluster together. Foursquare API is used to explore the boroughs and segment them. List data scraped from the website for each borough from their given latitude and longitude information. Used GitHub repository in the study to complete the documentation and sharing. Python's pandas library is used for data analysis. Pandas makes importing, analyzing, and visualizing data much easier. It builds on packages like NumPy and matplotlib to give you a single, convenient, place to do most of your data analysis and visualization work.

Importing, combining and cleaning datasets

Toronto neighborhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighborhoods in Toronto. You will be required to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a *pandas* dataframe so that it is in a structured format like the New York dataset. Once the data is in a structured format, you can replicate the analysis that we did to the New York City dataset to explore and cluster the neighborhoods in the city of Toronto.

Retrieving and processing Foursquare data

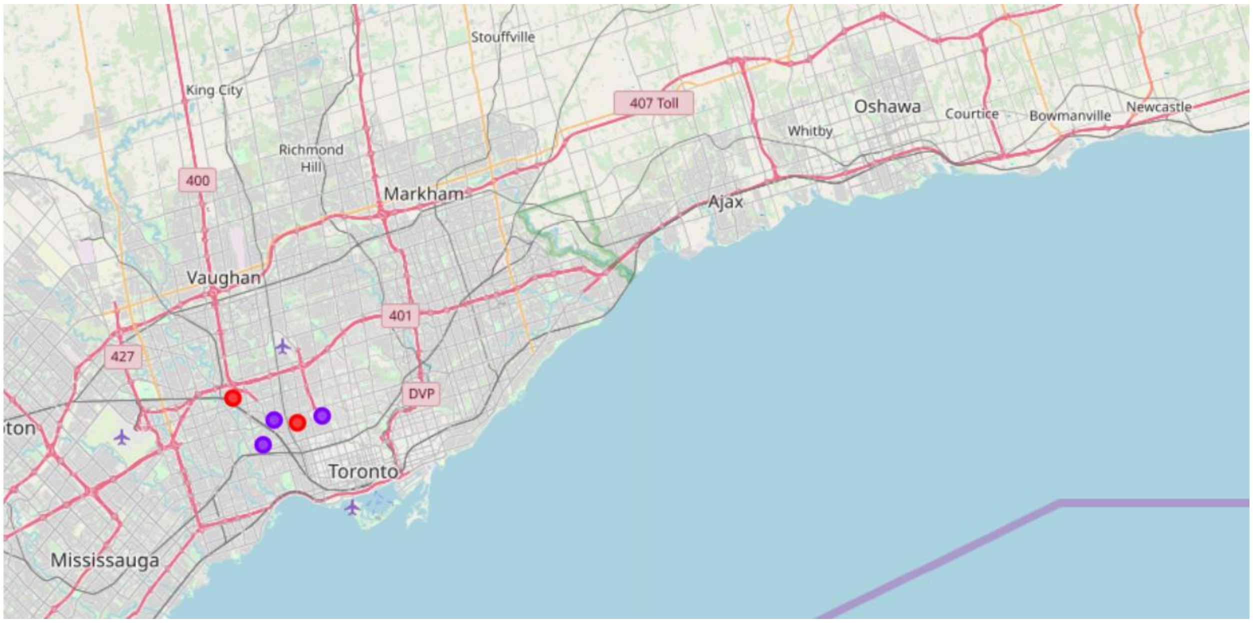
The venue data grouped to show the boroughs and the frequency. This information can answer a number for seeking data Science professionals.

```
----Caledonia-Fairbanks----
                  venue  freq
0                  Park  0.33
```

```
1                Market  0.17
2  Fast Food Restaurant  0.17


----Del Ray,Keelesdale,Mount Dennis,Silverthorn----
            venue  freq
0  Discount Store  0.25
1  Sandwich Place  0.25
2      Restaurant  0.25


----Humewood-Cedarvale----
          venue  freq
0         Trail  0.25
1  Tennis Court  0.25
2         Field  0.25


----The Junction North,Runnymede----
               venue  freq
0           Bus Line  0.25
1  Convenience Store  0.25
2      Grocery Store  0.25


----Weston----
      venue  freq
0      Park   1.0
1  Bus Line   0.0
2     Trail   0.0
```

After analyzing each borough using all datasets including Foursquare data and applying K-means Clustering, clustering boroughs and determined the clusters under study

Visualizing cluster and presenting cluster information



Results

The neighborhood Weston stands out from the four others in the York borough of Toronto, for fitness enthusiasts. The other neighborhoods, also have parks but are further away from the highway, as can be seen on the map. The neighborhood could be a good for the choice of restaurants with vegan and healthy food choices. The most common venues are check cashing services, tennis courts and a bus line etc. The neighborhood Fairbanks is desirable for restaurant choices and residents preferring the easy to access to public transport.