Coursera IBM Data Science Professional Certificate

Applied Data Science Capstone

By

Dharani Ravichandran

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 similar to the downtown Toronto. This is to explore the restaurant kinds that has an optimal chance of success.

Target audience

The principal stakeholders targeted are business explorers and entrepreneurs.

Data Section

Besides the datasets from week 2 and 3, Foursquare location data is leveraged to explore and compare the different boroughs. Foursquare is used to gather data on the most popular food venues per borough. Having signed up for a Foursquare developer, using 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 dataframes are combined into a single dataframe (with non-numeric values removed) in order to perform the clustering operation.

Data cleaning

Both datasets are fortunately very clean to begin with. Both files are in csv format and contain no empty values. They are easily loaded into Jupyter Notebook, which is used for the analysis.

Feature selection

Toronto boroughs, their respective geometric coordinates and relevant (mostly) demographic data, Foursquare is used to gather data on the most popular food venues per borough. Foursquare API provides access to a massive database consisting of venues from all around the world including a rich variety of information such as addresses, tips, photos and comments. Having signed up for a Foursquare developer, using 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 dataframes are combined into a single dataframe

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

1. Importing necessary libraries

2. Importing, combining and cleaning datasets

3. Retrieving and processing Foursquare data

4. Analyzing each borough using all datasets including Foursquare data

5. Using K-means Clustering, clustering boroughs

6. Determining cluster

7. Visualizing cluster and presenting cluster information

Results

In this section all the findings of the above clustering and data visualization will be documented.