IBM Data Science Professional Certification Capstone

CLUSTERING AND SEGMENTING US COLLEGES

Introduction

While searching for colleges and universities that best fit one's needs, the question of the location of that college often comes up. Prospective students are interested in knowing what kind of a place they will be spending the next 3-4 years residing in and what kind of amenities will be available to them.

Furthermore, the location of a college also serves as an indicator of the kind of people that frequent it as the venues surrounding a focal point estate are usually in accordance with the tastes of the masses that frequent it.

The idea behind this project was to take up the data of colleges across the united states and leverage the foursquare API to find the most popular venues found near and around each of these institutions.

Clustering algorithms could then be used to cluster like neighborhood colleges into the same groups so that prospective students could explore the location of the colleges they are interested in a little better and also find colleges similar to ones they already prefer based on encompassing venues.

The data used

The primary database that is used in the project was obtained from an online data source called opendatasoft. The dataset is quite large and has detailed information of about 7500 colleges across the United states. Data fields like the ranking, the tier, the postal code, etc of each of the colleges have been included into this dataset.

The most important feature of this dataset which made it optimal for usage in the current project is that it already contains an attribute called coordinates which houses the latitude and longitude values of each college as a string type variable.

This makes the task of coding the project easy and the outputs more reliable as the requirement of the geocoder library is eradicated. We could now directly wrangle and clean this dataset and send it off systematically to the foursquare API to get desired results.

The final dataset that is to be queried by the user is created using the results returned by the foursquare API based on the kind of venues that exist around each college, te one hot encoding of each location is done and based on the results, the top 20 most popular venues present around each college selected.

This dataset is then sent into the K Means clustering algorithm which returns suitable cluster labels. These cluster labels along with the name, address, zip code, and the top 20 most popular venues at each college are stored in a data frame which becomes the final dataset of the project.

Methodology Section

The project was divided into four sections. In the first section, the dataset was imported and cleaned using data cleaning and wrangling techniques. In the second section, all the required libraries and packages were installed. In the third section, the foursquare API was contacted to get relevant results, and in the final section, the results were incorporated into the dataset and the data was clustered to obtain the results.

1. CREATING THE DATASET AND SOURCING THE DATA

In this section, the dataset as obtained from the opendatsoft website was uploaded and integrated into the jupyter labs environment. It was discovered that the dataset was quite vast and constituted a lot of entries and attributes that were not of use in the present context.

Thus the relevant attributes were located and identified and a new dataframe df was created with only the relevant information. Once this was done, it came to attention that the coordinates column did have the right required information but in an unusable format.

To solve this issue, the column was converted to a list from string variable type and then separated to distinct latitude and longitude attributes. In addition to this several other data cleaning and wrangling strategies were applied. The null value problem was addressed and a total of about 100 samples were taken from this vast dataset for ease of computation.

2. IMPORTING PACKAGES AND LIBRARIES

In this sectional, all of the packages that were integral for the working of the project were imported into the notebook. This included the pandas and numpy libraries for data frame manipulation, the cluster package from scikit learn to implement the main clustering algorithm, the folium library to visualize the results, the requests library to handle API requests and the json library to handle json files as returned by the foursquare API.

3. LEVERAGING THE FOURSQUARE API

In this section, the client credentials for the foursquare API were generated. A function called getNearbyVenues() was defined and was used to iteratively call the foursquare API

for each data entry and get the relevant venue information. The table also created a new dataframe called nearby_venues to store the results of the process. A call to the function with the dataframe as generated in section 1 was made and results obtained.

4. CLUSTERING THE COLLEGES

In this section, the venue data for each college was grouped by college name and then one hot encoded to get the top 20 most popular venues in and around each university. After doing this, the data was fed into the clustering algorithm.

The algorithm used for this purpose was the K means clustering algorithm. This was done because K means is one of the easiest and uncomplicated algorithms for clustering. Moreover, if run a sufficient number of times, it provides pretty accurate predictions.

After clustering, the K Means object was used to obtain the clustering labels. These along with information of each college and the top 20 most popular venues in and around each of these places were compiled to create a new database which served as the final output of the project.

The folium library was used to visualize the results of the clustering on the map, and a rudimentary UI was created to allow the user to find colleges similar to the ones they are interested in as well as the popular venues at each of these locations.

Results

As a result of the above exercise, a data frame with relevant cluster labels and venue information of each of the colleges ws generated. This could be used by the user to find colleges similar to the ones they are currently interested in on the basis of surrounding venues. Moreover information about the kind of venues they will find around the college they are interested in is also provided.

When the results are visualized on a map, it is observed that most of the colleges have similar cluster labels indicating that the venues around each college try to be what the mainstream millennial culture portrays. In other words since the requirements and tastes of a majority of college students are thequite similar, so are the venues that cater to these needs.

Discussions

During the course of project development, the biggest problem I came across was the data processing capabilities of utilized platforms. Since the original database was quite large, when used in its entirety, the datasets would take quite long to load and the kernel would expire before the computation was complete wasting time and internet resources.

Moreover, since the number of API calls that can be made to the foursquare API with a basic account are limited, the entire dataset could note be used.

If there was a respite in these problems even slightly, more optimum clustering results could be obtained.

Conclusions

Under this project, the skills learned during the course of the entire specialization were put to application. We got to address all of the practical problems that come along with the data science and analytics processes.

As a result of the project, a dataset was created which could be queried by the user to obtain relevant information. The data was set in the realtime and the code could be run periodically to get the most up to date results.

The utility developed under this exercise can be embedded into any existing online portal for university aspirants. It may also be converted into a web application with an interactive GUI to enhance user experience and share important insights.