Battle of two cities

Capstone Project Report

Introduction/Business Problem

New York City and London are one of the most popular and the most visited cities in the world.

This Capstone project will focus on some of the Venue Categories in New York City and London, using the **Foursquare Location data**. This involves fetching a sample of Venue data from the *Foursquare API* and arranging them on the basis of the most frequent venue categories, along with the venues falling under that category.

*For e.g. From a sample of **100** Venues, New York City has four popular bakeries whereas in the 100 sample data set of venues in London, we find only one popular bakery.*

The Business Problem is more of an informative reflection on the comparison of these two great cities. Thus, the target audience could comprise of the popular demographic of tourists traveling to/from these locations to explore the venues hosted here, which would be really helpful, especially in a time-crunch. This project tries to showcase the venues from both these cities and forms a comparison of sorts between the two cities.

Data

To form a comparison analysis between the two cities, the data that will be used in this project will be *Foursquare Location data* (by accessing the **Foursquare API**) and then mapped onto Libraries like Seaborn and Folium maps for visualization.

For e.g. the New York Venues grouped according to the venue categories (from a sample of 100 venues) would be fetched by these steps:

- Hit the **explore** Foursquare API endpoint of both these cities
- Fetch 100 of the venues and their categories and save these in separate Pandas DataFrames

- Group these venues according to their categories to form a grouped data of the highest occuring Venue Categories in that dataset
- Visualise that data using Seaborn and Folium Maps

Methodology

The methodology used in this project can be summarised under the following points:

1. Fetch Geolocation parameters (latitude, longitude) of the location and access the *Foursquare* location API to **get info of top 100 venues**
2. For both the cities, **group the venues** according to their categories and **gather statistical information**
    a. Frequency of occurrence and mean of these frequencies for a comparative analysis of the venues in the two cities
    b. K-means clustering as an unsupervised learning technique to categorize venue categories of the two cities
3. **Visualize** the analysis using libraries like: **Folium** (for geographically visualizing the venues), **Seaborn** (for visualizing statistical calculations)
4. For a **comparative analysis**, combine the data of the two cities on the basis of their *common venue categories*, and re-collect statistical information for this combined collection
5. Draw **conclusion** on the information collected from this analysis

Results

The exploratory data analysis of the above project resulted in the following insights about the venues of the two prime locations:

New York City

- Highest number of Ice-cream shops - ***6*** around East Village
- ***1*** Pub and ***2*** Bars - spread across Astor Place and East Village
- ***4*** Pizza Places and ***4*** Coffee Shops, spread across East Village *(again!)*
- *Ramen and Thai restaurants* (***4***) are more likely to be found sandwiched between 1st and 3rd Avenue, East Village
- Unlikeliness, or the least number of occurrences (***1***) of common venue categories are
    - Arts and Crafts Store
    - Bakery

  - Book Store
  - Burger Joint
  - Greek Restaurant
  - Pub
  - Sandwich Place
  - Spa
  - Tea Room
  - Wine Shop

London

- ***4*** Ice-cream shops - around Leicester Square
- ***4*** Pubs and ***1*** Bar - spread across Leicester Square, London Charing Cross, and Piccadilly Circus
- Probability of finding *Pizza Places* (***1***) is lesser as compared to a *Bar* or *Pub*
- *Ramen and Thai restaurants* (***4***) are more likely to be found sandwiched between Leicester Square and Piccadilly Circus
- Unlikeliness, or the least number of occurrences (***1***) of common venue categories are
  - Bar
  - Bakery
  - Book Store
  - Burger Joint
  - Greek Restaurant
  - Wine Shop
  - Mexican Restaurant
  - Pizza Place
  - Spa
  - Tea Room

Discussion

- While the *Foursquare* Location API is highly useful in exploring the huge collection of venues and locations around any place on Earth, we are restricted as users of **Sandbox Accounts** in terms of accessing more number of venues in one day

- While executing *K-means clustering* on the sample data sets, the algorithm gives an optimum result when clustering on numerical data, but we need a more intelligent Machine Learning Algorithm to cluster **textual data**, as clustering on the basis of these values may not produce the desired results

- The **geocoder** library does not always fare well in producing the Geolocation attributes of a given location. However, I found the **Geopy** library as more reliable

Conclusion

In conclusion, the following points of executing a Data Science project can be inferred:

- **Foursquare Location** API exposes a wide array of methods to extract and use location-specific data in analysis and application development

- **Folium** is a great Leaflet library, capable of getting accurate locations and representations on a geographic map
- **Seaborn**, compatible with libraries like Pandas, Numpy, MatplotLib, provides a large number of plots and analysis tools, to create informative and stunning visualizations
- **Machine Learning** algorithms ease the process of exploratory analysis and modeling of the data of a project
- The exploratory analysis using the aforementioned technologies and methods helps achieve the goal to serve people, especially tourists, interested in visiting the locations.
- The results showed the venues and places where any person can visit in a given time frame, in an optimum way.