

A Tale of Two Cities – London vs. NYC

Introduction/Business Problem

The international business traveler often faces a dilemma: Where do I find places or things to do similar to what I have at home?

For example, a New Yorker may be accustomed to having a bevy of pizza parlors at their disposal. Sample clustering reveals no fewer than five popular pizza locations nearby via Foursquare. Therefore, he or she might be surprised to learn they are not as prevalent in the city of London, where Foursquare sampling returns only a single popular location.

Fortunately, data repositories such as Foursquare allow us to compare comparable venue types across diverse locations to give travelers options approximating the comforts of home. This project will do just that, utilize Foursquare data to fetch samplings of venues in both locations based on the most frequent categories and allow comparison of the data.

The goal of this project is to provide data for travelers between London and NYC to compare venues similar to those at their home location and have a more enjoyable travel experience with that data at hand.

Data

As the goal of this project is to compare venues between the two cities, we find that the Foursquare data (accessed via API) is sufficient for our purposes. Visualization will be accomplished via visualization tools such as Folium or Seaborn.

An example of how the data collection/extrapolation/visualization will be accomplished is as follows: Examination of clustered segments via Foursquare API, allocation of the 100 top venue types into dataframes, grouping of venues by category and mapping into the visualization tool.

Methodology

Methodology utilized can be summarized as follows:

1. Acquire and apply latitude/longitude geocode parameters for both locations and query Foursquare API for top 100 venues for both
2. Group venues for both locations by categories.
3. Use Folium and/or Seaborn for data visualization.
4. Compare data for both cities via common venues under comparative analysis.
5. Analyze results and make/report conclusions.

Results

Final data analysis revealed the following about both locations:

NYC

- Highest concentration of ice cream shops - 6 clusters
- Clustering of 4 pizza parlors and 4 coffee shops in one area
- A cluster of three bars/pubs around the East Village area
- Least number of occurrences (unlikeliness) was interesting for a cosmopolitan major

city. The following venue types had one occurrence each:

- Bakeries
- Bookstores
- Wine Shops
- Burger Restaurants

London

- Highest concentration of ice cream shops - 4 clusters
- Only a single pizza parlor in cluster area
- A cluster of five bars/pubs in three areas, significantly more than NYC
- Least number of occurrences (unlikeliness) was also interesting for a cosmopolitan

major city. The following venue types had one occurrence each:

- Bakeries
- Bookstores
- Wine Shops
- Burger Restaurants
- Tea Room (Very surprising)
- Mexican Restaurants

Discussion

- While Foursquare API is useful for mass venue analysis, its sandbox accounts limit calls per day, rendering mass analysis unwieldy.

- Numerical data analysis yielded good data, but textual analysis seemed bound by machine learning limitations.
- Geocoder data proved to be hit and miss in numerous runs

Conclusion

- Foursquare API has great potential for utilizing geodata in analysis development of applications
- Machine learning algorithms eased the analysis process, but its current limitations were also on display.
- Visualization tools such as Folium and Seaborn help tremendously in presenting geo-specific data to front end users in a manner which is visual and easy to comprehend.
- Results were shown to the user as planned, making the program an overall success