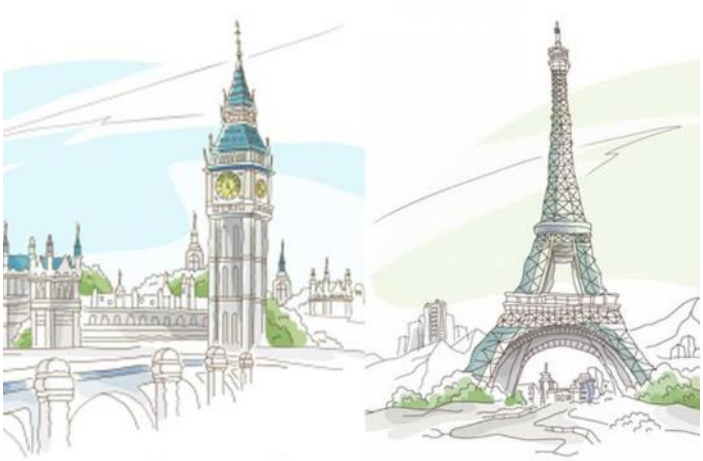

QUANTITATIVE SIMILARITY ANALYSIS OF LONDON AND PARIS NEIGHBORHOODS

Are you moving from London (Paris) to Paris (London)? Let's see which neighborhood of your future city will suit the most!

Marc Gou

This report has been drafted in the context of the course: IBM – Applied Data Science Capstone

Contents

Introduction

London and Paris are without doubt the two most economically powerful cities in Europe. Both of them are home to many of British and French national flagship companies, they have also attracted countless foreign companies to establish their national/European base.

They are also two of the most visited cities in the world. The large number of opportunities combined to their offers in the domain of art culture, sport, education, social system… have attracted many of the talents to settle up. Especially, a lot of Parisian are living in London and many Londoners are considering to move to Paris.

In fact, London is considered as being the fifth French speaking city in Europe, has been for long a primary destination for Parisian people. On the other hand, Paris is one of the city with the highest number of British citizens, and with the actual political context, we may reasonably expect even more London-based people moving to Paris.

The goal of this project is to provide insights for individuals or families which are moving from one of those cities to the other, which district(s) would fit them the most by using a quantitative approach and real-world location and venue data.

As a first step, we will cluster the different district of those two cities based on the venues available at each neighborhood. And in the second step, we will classify all the neighborhood of those cities into defined groups, and also given any district of one city, find the most similar district to that one in the other city.

Data

For this project we will mainly use the Foursquare location data. Foursquare is a technology company providing mobile search-and-discovery services. Foursquare features also a developer API that lets third-party applications make use of Foursquare's location data. In March 2013, the Foursquare API had 40,000 registered developers. The API powers searches third-party apps, including Evernote, Uber, Flickr and Jawbone (*source: Wikipedia*).

In order to get the neighborhood data of both cities, we are using the information from Wikipedia:

1. London: https://en.wikipedia.org/wiki/List_of_areas_of_London (only neighborhoods with Post town = “London” are included in the scope of this project)
2. Paris: https://fr.wikipedia.org/wiki/Liste_des_quartiers_administratifs_de_Paris

These geographical coordinates of the neighborhoods (called « areas » for London and « quartiers » for Paris) listed above are located via Nominatim (from geopy), which is a tool to locate addresses, neighborhoods and interest points by using the data of Openstreetmap.

Methodology

Preparation of data

First, we downloaded the data from Wikipedia and load them as dataframes. Then we used Nominatim Geolocator to add the geographical coordinate’s data to all the neighborhoods. With the

geographical coordinate's we used the function "Explore" from NetSuite API to get all the venues within a radius of 1000 meters. We only keep the categories of venues that exist in both cities.

All neighborhoods with no geographical coordinates or with no venues arounds are dropped from the dataframe.

At that stage, we have one dataframe for each city containing the number of venues from each categories by neighborhoods.

Refer to the sections 1, 2, and 3 of the Jupyter Notebook

Clustering

For the clustering, we used the k-Means algorithm:

- For London, as we have +- 180 neighborhoods, we selected k = 18
- For Paris, as we have +- 80 neighborhoods, we selected k = 8

Refer to the section 4 of the Jupyter Notebook

Classification Model

As we would like to know for a given neighborhood from one city, what would be the similar neighborhoods from the other cities, we model this as a multi-class classification problem. Two steps are implemented to determine the solution:

First step: Build two models, lond_predSVM for London and paris_predSVM for Paris with the following parameters:

- Algorithm used: Support vector machine with the parameters *gamma* and *decision_function_shape* adapted to a multiclass classification problem
- Independent variable: Y = Cluster Label
- Dependant variables X = Normalized values of each category of venues of the neighborhoods

We split the dataset into a training set (75%) and a testing set (25%) in order to assess the accuracy of the model:

- lond_predSVM 's Accuracy according to Jaccard Index: 0.7778
- paris_predSVM's Accuracy according to Jaccard Index: 0.7500

Second step: We used the models obtained to predict what would be the best cluster in the other city for a given neighborhood of one city:

- Independent variable: Y = Cluster Labels of the other city
- Dependant variables X = Normalized values of each category of venues of the analyzed city neighborhoods
- Model used: The model computed for the other city

Refer to the sections 5, and 6 of the Jupyter Notebook

Results

The final results are registered in the two following dataframes:

- **london_final_wParis_df**: London neighborhoods (column “Location”) with their corresponding computed cluster in Paris (column “Paris Cluster”)
- **paris_final_wLond_df**: Paris neighborhoods (column “Quartiers”) with their corresponding computed cluster in London (column “London Cluster”)

In order to visualize the results obtained, we used the library Folium to show the location of the best neighborhoods in the other city found for the following selected neighborhoods:

- Westminster, London: The most similar neighborhoods in Paris are the neighborhoods of the cluster 0
- Blackwall, London: The most similar neighborhoods in Paris are the neighborhoods of the cluster 3
- Place-Vendôme, Paris: The most similar neighborhoods in London are the neighborhoods of the cluster 16
- Val-de-Grâce, Paris: The most similar neighborhoods in London are the neighborhoods of the cluster 1

Refer to the section 7 of the Jupyter Notebook

Discussion

We can observe that despite we have selected k = 18 for London and k = 8 for Paris, there are three to four dominant clusters for each city. In consequence, the recommendation at the end always pointed to those dominant clusters. Therefore one area of improvement maybe redefining the value k in K-Means Clustering to obtain more insightful results.

Another area that would be interesting to be explored is the classification method used, here we used SVP without considering its accuracy to the other algorithms available. It can be interesting to compare its results to some other well-known classification algorithms such as decision tree, k-NN, random forest…

From a less technical perspective, we can also improve the models by using more diverse data, such as criminality, average age of the population, education level.

Conclusion

The objective of this project is accomplished, this work would allow to any people who lives in a specific neighborhood of London or Paris and who need to relocate to the other city in a near future to identify, with a quantitative approach, the neighborhoods in the other city that would be the best-fit from a venue category availability perspective.

Even though numerous areas of improvement exists as described in the section “Discussion”, this project is a good first step to dive into the topic.