Julien Pascal
https://julienpascal.github.io/

Capstone Project - The Battle of Neighborhoods

Table of content

I. Introduction

Paris is a vibrant and complex city. For someone who has not been living in Paris for many years, the city may look impenetrable. **Which neighborhoods are great for a coffee? Which neighborhoods are famous for its markets? Where are the best bars?** When confronted with these questions, tourists and new residents generally use a guide or use websites such as Yelp or Tripadvisor. My experience with these platforms has often been disappointing because the website (or the guide) is not tailored to my tastes and preferences. For this capstone project, I would like to offer an alternative based on data mining and clustering.

Data on Paris neighborhoods amenities (bars, cafés, museums, bakeries, etc.) can easily be collected and treated to generate a map of Paris. Based on the user's preferences, we can direct the user towards a specific neighborhood in Paris. This approach can be seen as **a refinement of the traditional tourism websites, with the addition of a data-driven customization layer**. Having a data-driven map of Paris is also helpful for people moving to Paris when deciding where to live. The French capital is extremely expensive. Yes, the Marais is great, but maybe you prefer living in the much cheaper 19th or 20th? To answer this question, **we need data and a robust clustering methodology.**

II. Data

To create a data-driven map of Paris' neighborhoods, I use data from **Foursquare API.** Foursquare defines itself as "a location technology platform dedicated to improving how people move through the real world". In practice, people use the Foursquare website or app to find places. They can then rate the venue, give a rating, add photos and/or a description. The Foursquare API allows researchers and data scientists to retrieve observation points that were created by users.

I also use data from **Wikipedia** to get information on Paris. The following page contains the name of Paris areas (called "arrondissements"), as well as some basic information such as area and population.

III. Methodology

A. Software choice

To generate new insights on Paris' neighborhoods, I use **Python**. Python has become the dominant language within the realm of data science because it is both powerful and simple to use (as well as being free). To manipulate data, I use **pandas**. To geocode locations, I use **geopy**. For the visual exploration of neighborhoods, I use **folium**. For the clustering algorithm, I use **scikit-learn**.

B. Algorithm choice

What makes neighborhoods similar? "SImilarity" is a broad concept and may encapsulate many dimensions. In this analysis, I measure similarity using shops and amenities (restaurants, bars, cafés, museum, theatre, etc.). If two neighborhoods possess the same types of shops and other amenities, their similarity score will be high. One may think of more sophisticated notions of similarity, also including the type of residents living in each neighborhood (age, nationality, etc.), or including the cost of housing. These more sophisticated approaches could be explored in a subsequent report.

In terms of clustering algorithm, I make use of the **k-means algorithm**. The k-means algorithm is a relatively simple clustering algorithm that works well on large scale data sets and that is guaranteed to converge. Some of the disadvantages of the k-means algorithm is that the outcome depends on starting values and that it does not detect outliers. In subsequent analysis, one may use the DBSCAN algorithm instead, which is great in detecting outliers in a sample.

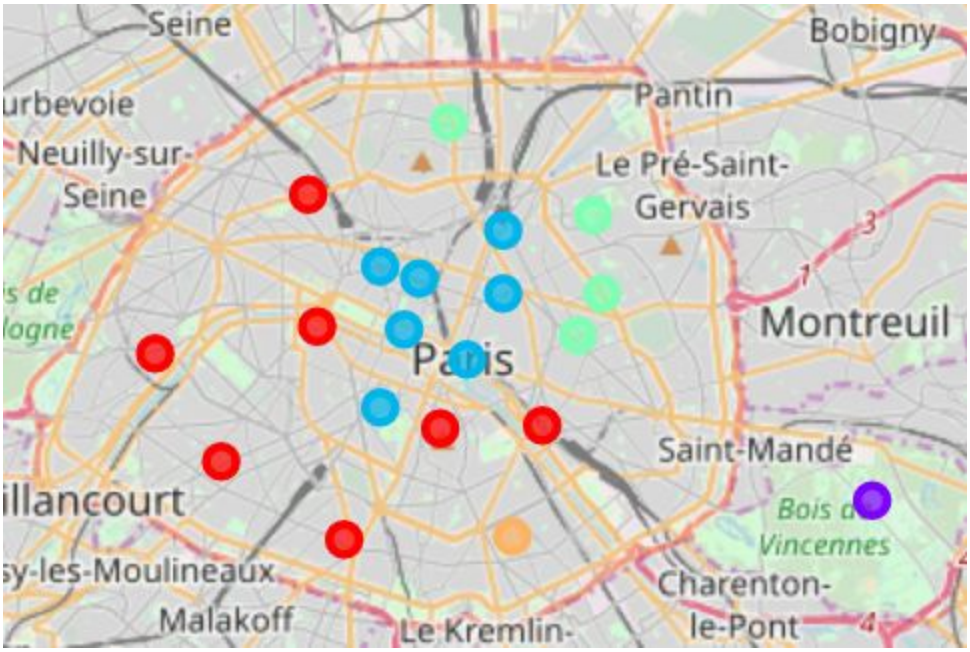
IV. Results

A. Paris clusters

I choose an arbitrary number of clusters (k=5). Results are presented in **Figure 1**. The clustering algorithm detects distinctions that are familiar to people living in Paris. The first cluster (in red) is composed of neighborhoods from the West of Paris (5th, 7th, 8th, 14th, 15th, 16th and 17th arrondissements). The second cluster (in purple) is composed of the

12th arrondissement. The third cluster (in blue) is composed of the central arrondissements (1st, 2nd, 3rd, 4th, 6th, 9th and 10th). The fourth cluster (in light green) is composed of the North and East arrondissements (11th, 18th, 19th, and 20th). The fifth cluster (in orange) is composed of the 13 arrondissement.

Figure 1. Paris arrondissements clusters



B. Properties of clusters

Each cluster has its particularities. I use word clouds to represent the type of amenities associated with each cluster. The first cluster (Figure 2) contains many french restaurants, hotels and historic sites (the Eiffel tower belongs to this cluster). The second cluster (Figure 3) contains many theaters, stadium and sport facilities (The Bois de Vincennes belongs to this cluster). The third cluster is very diverse, with a healthy mix of diverse restaurants, hotels, museums and art shops. The fourth cluster is dominated by bars, bistrots and pizza places. The fifth cluster contains many asian cuisine restaurants.

Figure 2. Word cloud 1st cluster

Figure 2. Word cloud 1st cluster

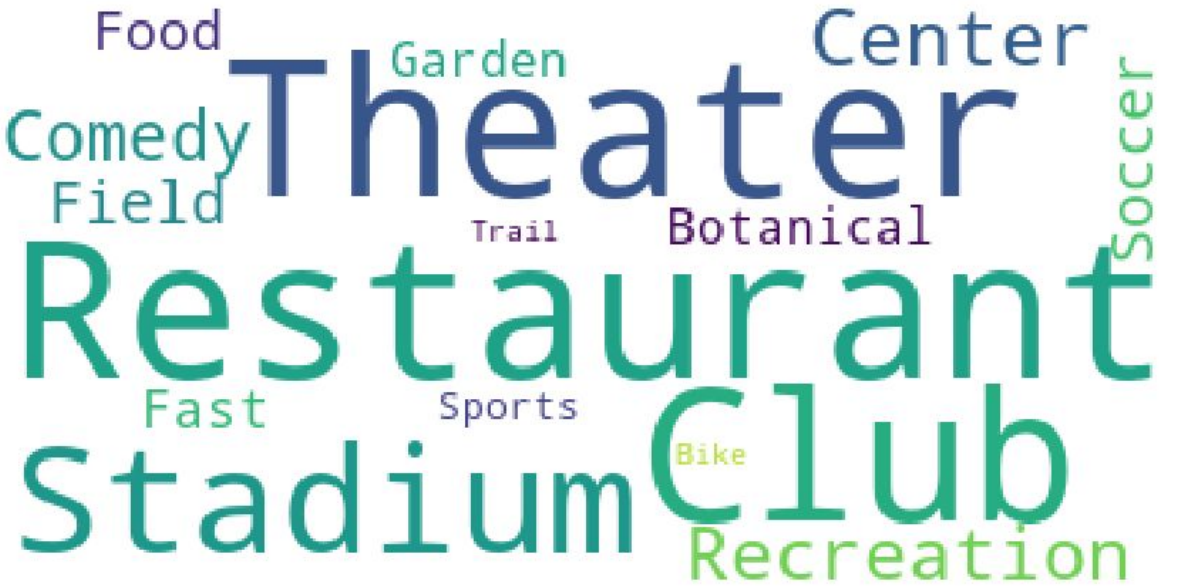
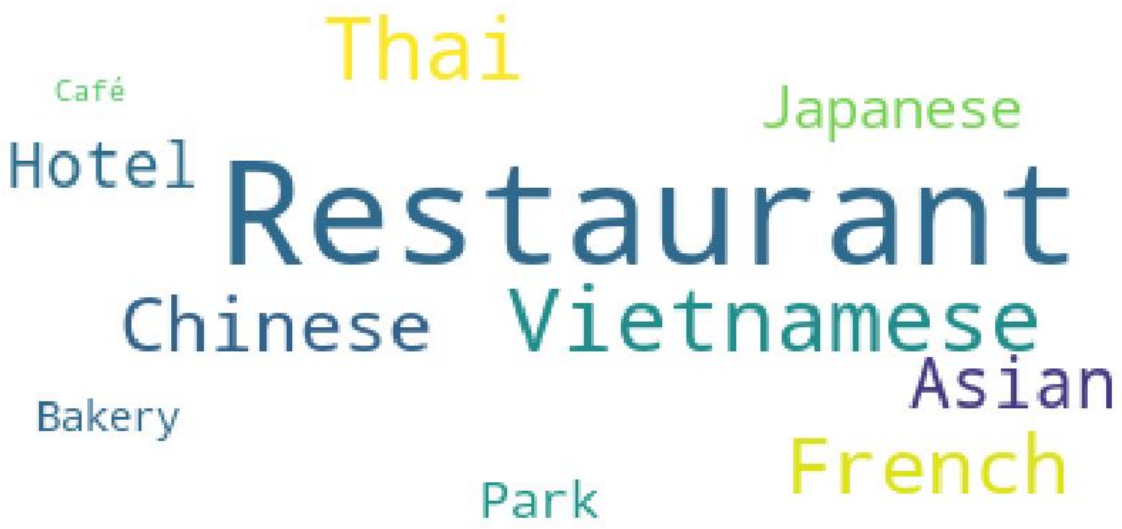


Figure 3. Word cloud 3rd cluster

Figure 4. Word cloud 4th cluster

Figure 4. Word cloud 5th cluster



C. Where to live in Paris?

- If what you are looking for is a very “French” experience, in a mostly quiet and residential area, the first cluster is for you.
- If you want to enjoy the vibrance of Paris, but still value fresh air and sport, the 2nd cluster, near the Bois de VIncennes, is for you.
- If you are looking for a mix of art, museums and diverse cuisine, go for a place in the 3rd cluster.
- For the best bars/bistros and pizza experience in town, the 4th cluster is a must.

V. Discussion

The present analysis could be improved along several dimensions. First of all, different algorithms could be used to check for the robustness of the present classification. Using DBSCAN would be particularly attractive, as the number of clusters would be automatically selected. Second of all, one could replicate the present analysis using data at a finer geographical scale, using an Iris as the geographical unit. Third of all, one could try to incorporate user ratings in the analysis. Finally, using additional datasets could be valuable to generate more precise clusters. Having data on house prices and rents could be particularly relevant.

VI. Conclusion

A data-driven approach using the Foursquare API and clustering techniques generates new insights on the neighborhoods in Paris. Five clusters have been identified. Depending on your preferences, it is now easier for you to decide where to live in Paris or what arrondissement to visit during your stay.