Capstone Project
The Battle of Neighborhoods in Paris (Week 1)

Christophe Coudé
December 09, 2020

1. Introduction

1.1 Background

There are currently around 12,000 restaurants in Paris for 2,150 million inhabitants. That is to say, on average, about 1 restaurant per hectare. That's a lot, the competition is tough. Many restaurants are closing, especially in these times of crisis. A newcomer might then wonder where to set up his (or her) new restaurant. He may then have the choice between :

- finding a place where there are very few restaurants of his specialty in order to attract customers of the local neighborhood who want to change cuisine or,

- setting up his restaurant where there are already many restaurants of his own specialty because even if there is competition, this means that there is an already existing clientele.

The restaurateur must therefore know what is the distribution of "cuisines" in the different neighborhoods in order to be able to make his choice.

1.2 Business problem and interest

The aim of this project is therefore to offer new restaurant owners a global vision of the distribution of restaurants in Paris thanks to a clustering which makes it possible to know for each district which are the most represented cuisines or regions and those less represented.
According to the strategy chosen by the owner, he (or she) will be able to choose in which district to try settling.

Therefore, the problem is :

- to find where the restaurants are located and to know their specialty, for example using the *Foursquare API,* and then
- to use machine learning to bring out the general culinary trends and tastes of each neighborhood, for example the unsupervised learning method *Clustering*.

2. Data acquisition and cleaning

2.1 Data sources

Worldwide venues description can be accessed through the **Foursquare API** once an account has been created on their site there.

The different categories of cuisines are referenced on their page of categories.

We can see an example of hierarchy describing the kind of cuisine and their ID:

Food
4d4b7105d754a06374d81259
     Afghan Restaurant
     503288ae91d4c4b30a586d67
     African Restaurant
     4bf58dd8d48988d1c8941735
          Ethiopian Restaurant
          4bf58dd8d48988d10a941735

To request Foursquare, we need the category of restaurant and the GPS coordinates of the neighborhoods of Paris.

It's quite difficult to find accurate GPS coordinates... Each site gives different coordinates... So, for the center of Paris, needed to center the Folium map, I chose the site **latitude.to** that I scraped with the library BeautifulSoup.

However, for the coordonates of the 20 districts of Paris, I couldn't find correct coordinates on the Web, so I "manually" used ***Google Map*** to approximately determine them.

2.1 Data cleaning

There was another difficulty. Foursquare uses circles to demarcate the areas where it will look for the venues. But the Parisian districts are not round.

So I had to manually draw several circles (from 1 to 3) inside the neighborhoods in order to get as many restaurants as possible, even though Foursquare requests only return 30 venues per category each time. So it may be not exactly accurate, but it's the principle.

I couldn't draw a unique big circle encompassing each district because Foursquare only brings back 30 venues per search research. I would have lost more venues.

I entered the coordinates of the centers of the districts are in my local file *district_centers_coordinates.csv*.

	Arrd	Arrd_lat	Arrd_lng
0	1	48.862994	2.335592
1	2	48.868651	2.342876
2	3	48.863132	2.359781
3	4	48.854801	2.358448
4	5	48.844402	2.350722

And I have entered the centers of the circles and their radius in my local file *district_coordinates.csv.*

	Arrd	Arrd_lat	Arrd_lng	rad
0	1	48.865789	2.327739	500
1	1	48.863079	2.336408	500
2	1	48.861244	2.345377	500
3	2	48.869364	2.333671	150
4	2	48.868489	2.341267	300

Here's an example of how the approximation renders for the 17th district :
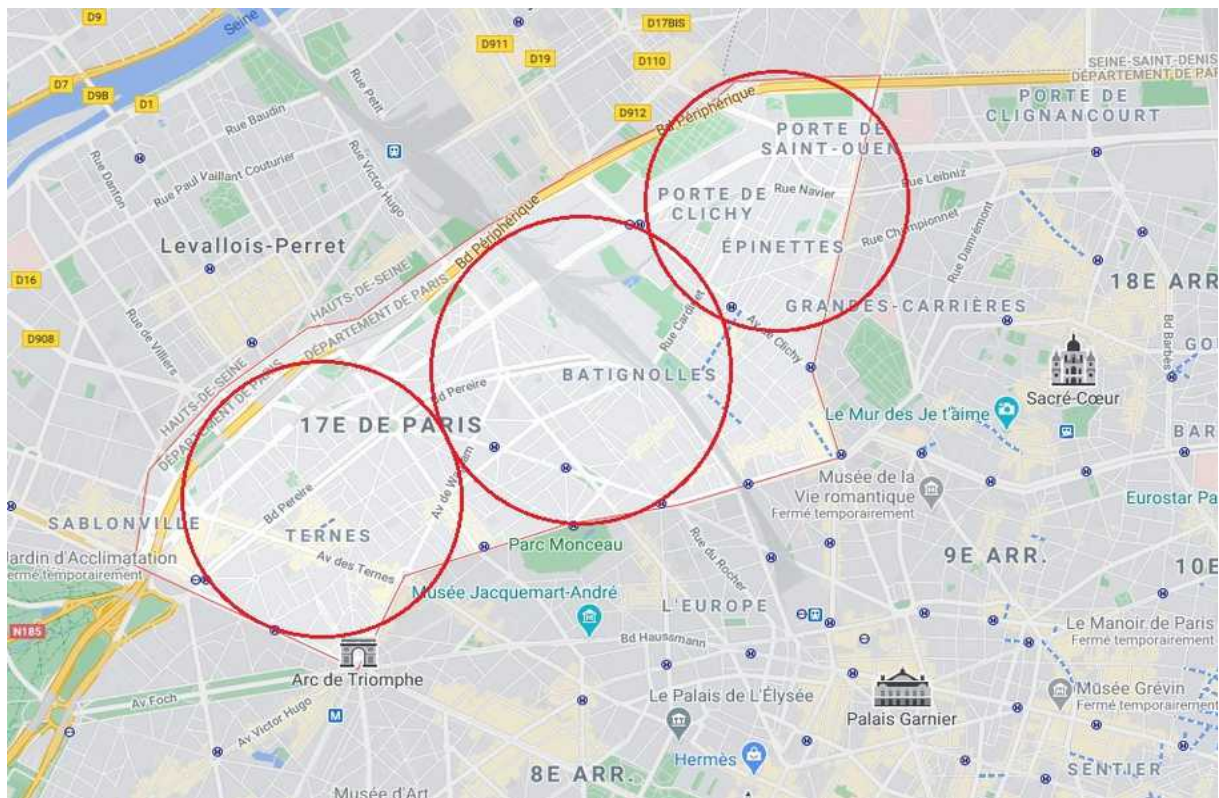
2.2 Feature selection

There are about 250 food categories in Foursquare but it was too much because Fousquare limits to 950 requests per day for a free account. So I restricted to 69 categories. Multiplicated by the 46 "circles" in the 20 districts, that makes 3174 requests to Foursquare, i.e. 4 days.

I also added a « region » feature to geographically group most of the cuisine categories. Foursquare had already did it, for example Asia, Africa and so on. But as some venues were only to be find in sub-categories or only in regions, I decided to create a feature to group every country of a continent.

I just made three exceptions to try to highlight some specificities in Paris. French, Indian and Japanese venues. But they can be respectively regrouped in Europa and Asia. For example, we'll see that Japanese restaurants are often more common (5th region place) than America (6th place).