Project Proposal

Exploring the AirBnB dataset of New York to guide travellers

E-558 A Network Tour of Data Science

Samuel FURTER, Sylvain LUGEON, Paul MOSSER, Florian HARTMANN

3 December 2019

1. Aim of the project

Since a few years, AirBnB has become an easy way to find accommodations for more or less long stays in big cities. With the amount of accommodations and the number of tourists constantly increasing, it has been crucial to find the best spots in the city for your stay. The analysis of this huge amount of data can provide useful information to the travellers such as the most popular districts, the most expensive ones and even the most adapted to a particular search.

2. Data acquisition

We found the Airbnb data from New York City on *Kaggle* but we will download it from Airbnb directly. Since the data are given by the "producer", the dataset is well furnished with almost 50'000 accommodation and zero missing data. They are already in a *csv* file thus we just have to select the features we want to keep for our analysis.

3. Data exploration

The first step of this project will be to construct graphs that represents the AirBnB dataset. Our idea is to take the AirBnB listings as nodes and to compute the weights between the nodes according these features : location of the accommodation, price per night, type of the room, owner of the listings.
The weights of the edges could be based on one, or a combination, of the above features. Each graph could represent a different aspect of the AirBnB network.
We will then explore the graphs, and highlight the most interesting characteristics (edge distribution, number of connected components, etc.). The creation and the exploration are closely related, as it is very important to have good graphs. We will modify the weight calculation according to the graphs exploration.

Data exploitation

The exploitation of the graphs will follow these different axis:

- Identify clusters in the graphs representing the neighborhoods. Compare them with the ground truth.
- Consider the price of the listing as a function on the graph. Training a Machine Learning model that could output the price of an AirBnB according to the inputs a traveller would enter in a search engine (location, type of room, duration of the stay).
- We could imagine clustering owners that have similar listings together and try to see if there is a correlation with the area for example.