Exercise Homework 1

This homework will be intense. You can work on the scraping part of this homework during the TA session with Luis' help. The TA session will be there to help you - make good use of this and prepare a little before by doing step 2) before the TA session and then working on steps 3) following. Please read all the instructions carefully before starting.

1. Get together in groups as randomized here. (Link to Google Sheet)
2. Download the material for the homework and get the booking scrape code to run. Prove that you did by attaching your code (this is only relevant for those who fail later steps).
3. Due date is 5th of February.
4. **Submission must be done as follows:**
   - One single .zip file per group.
   - File name should be on the format *group#_surname1_surname2.zip*; with all the surnames of the members included.
   - Include codes and pdfs and upload to google class (only a single submission per group).

Mini Research Project

Part 1: Scraping

Design and implement a mini research project in which you research the effect of a big annual event in Barcelona on rental prices on booking by scraping data for at least two separate weeks (important note that search results go across different pages) for Barcelona and at least one more city.

1. Identify a (future) event that makes a lot of people come to Barcelona. Think about music festivals, local festivities etc.
2. Think of the time periods to scrape and what second city to scrape. The second city will be your control group. Explain your choices in written.
3. Design a careful scraping pipeline that follows the advises seen in class and TAs. The

basic points to bear in mind are:

- Organize the data you need, format and structure to store it **beforehand**. Try to foresee how you will need to read in the data to answer your questions. If you want, you can include some few lines explaining your pipeline strategy at the beginning.
- Codes should be as automated as possible. That is, you don't want to rely on human intervention to get your data.
- Use **only** the packages we have seen in the course. Although firefox is recommended, you can also use chrome as your scraping browser.
- Document your codes and make them robust and efficient.

Part 2: Text analysis

4. For each of the hotels extract the text on the description (and possibly other text metadata) and do the following:
   (a) Pre-process the text by removing stop words and stemming. Customize your stopword list if needed.
   (b) Create two wordclouds before and after pre-processing for each city (a total of four). Comment on the changes in the wordclouds.

Part 3: DiD

5. Suppose you want to understand the effect on prices of the events you have choosen. We will apply Differences-in-differences methodology in DiD regressions with the dataset you have collected. Watch out that you get the indexes right when answering the following questions.
   (a) Write down a fixed effects regression equation that allows you to derive a difference-in-difference estimate of the effect of the event on prices. Think of controls to add, why is this relevant? Explain why you need a second city for this.
   (b) How would you use text features from the description as controls? Think about the text in the descriptions you scraped. How would this help? Why would terms like "Barcelona" not help?
   (c) Now suppose we want to decompose the treatment effect by hotel quality. Can you use the text description here? How would you use them to study heterogenous treatment effects? Write down a regression equation.