Facebook

One of the sources we need to scrap was the website Facebook. In this report, we're going to explain in a first time which data we scrap and the format of these data. In a second time, we're going to develop our scraping method and we're going to finish with some statistics.

This website has a page dedicated to public pages where we can find publications in specific pages (exemple : Airbus) about airlines. We've decided to deal with 40 pages, and we retrieve only the first comment on each publication.

I Format of scraped data :

The table below summary all data we scraped and give their type and a description of what is the column.

Variable name	Type	Description
Data_Source	string	Source of scraping
Date_publication	date	Date of publication
Likes	int	Number of likes on publication
Description	string	Description of publication
Nb_sharing	int	Number of share on publication
Nb_comments	int	Number of comments on publication
Review	string	Comment in english

II Scraping method:

We scrap our data into several steps :

- For each pages, select informations publication thanks to soup and PhantomJS method.
- You select the description of the publication as well as its date and the number of shares, likes and comments. A comment is also taken.
- DataFrame will be exported in a json file.

In order to retrieve the comments, you have to go through Beautiful Soup and PhantomJS (scroll pages), so that you can to retrieve the page soup. In the soup, we have all tags and information pages. In order to retrieve the data we go through the htlm code of the page .

One of the big problem we have when we scrap Facebook is when you want to see more comments, the button "see more comments" is approximately as the same place than the like and share buttons. So phantomjs think you want to click on like/share buttons and no on "see more contents" button. That's why we retrieve only the first comment on each publications.

We don’t have a robot on this website because the date format we retrieve is not the same than the expected format, even if we change the format of facebook date. If you need all new comments and publications, you need to launch all the code.

III Statistics:

The final dataset counts 395 rows and 72 columns (7 not NA) and weigh 751 Ko.

We scrap the data on 40 pages.

The following statistics shows the mean, the median, the min, the max and the ecart-type of the number of comments of each page (for example, in mean, we have 30 commets by page)

	Nb_comments
count	395.000000
mean	30.493671
std	50.895053
min	1.000000
25%	2.000000
50%	5.000000
75%	34.000000
max	184.000000