Tripadvisor

One of the sources we need to scrap was the website Tripadvisor. 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. To respond to our problematic, we chose to scrape the informations that airline's users post on Tripadvisor after their flights.

This website has a page dedicated to each company where we can find comments connected to the companies flights. We've decided to deal with the 240 biggest companies repertoried in Tripadvisor.

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	Website used to scrape
Title	string	Title of the comment
Airline_Name	string	Name of the airline company
Review	string	User's review
Date_Review	date	Date of the posted review
Date_Flown	date	Date of the flow,
Hashtags	list of string	List of hashtags (possibilities from 0 to 3)
Contributions_Pers	int	Number of reviews posted on Tripadvisor
Nb_Pertinent_Comments	int	Number of reviews « liked » by other users
Overall_Customer_Rating	int	Global note of the user /10
Cleanliness	int	Note of the cleanliness /5
Food_And_Beverages	int	Note of the food and beverages /5
Inflight_Entertainment	int	Note of the inflight entertainment /5
Registration	int	Note of the registration /5
Seat_Comfort	int	Note of the quality of the seat comfort /5
Seat_Legroom	int	Note of the place for the user's space for legs /5
Value_For_Money	int	Note of the value for money

		/5
Overall_Service_Rating	int	Note of the overall service rating /5

II Scraping method:

We scrape our data into several steps :

- For each airline company : select « all languages » reviews to get the maximum of informations, then

- For each page of reviews : Getting all the informations described above for each block of review

- Then, the algorithm detect the language of the review and translate it automatically in english

- To finish, all the informations collected are stored into a table

- This table is finally exported in Json format to be analysed by the other groups

All this automatisation have been done with the driver PhantomJS but there is another version (V1) unsing Chrome Driver working on the same way.

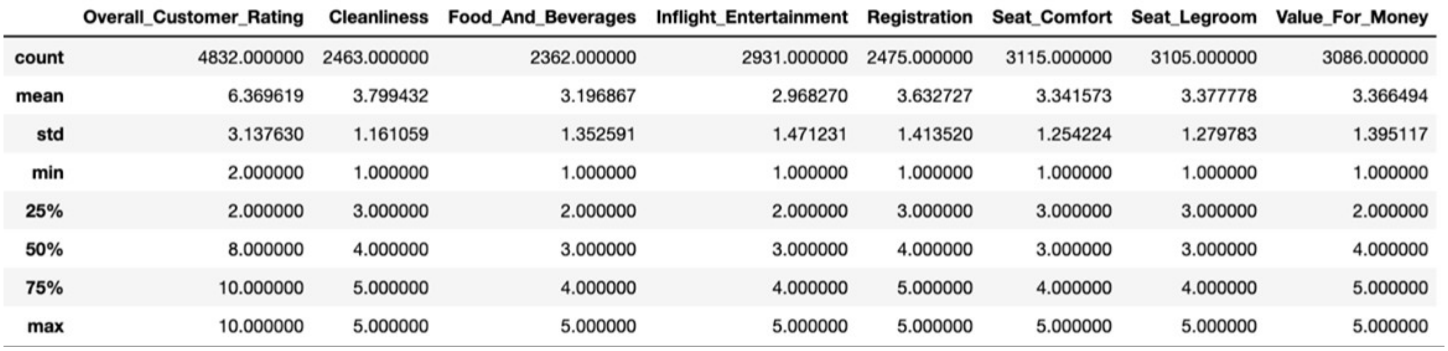
The final function allow you to retrieve the review posted after a given number of days. For exemple, if you wan't to take all the data posted on the current day, type « Lets_Scrape(1) » ; for today and yesterday type : « Lets_Scrape(2) » and for the seven last days type : « Lets_Scrape(7) ».

Tripadvisor is a site where you can get « spotted » as a robot very easily, so we had to put some items simulating « human behaviour » like scrolling a the end of every page, and waiting a little before clicking on buttons for exemple.
This has hard consequences into the time that the function take to get data. There are approximatively 1.700.000 comments available, and getting 5000 of them cost us 12 hours.

III Statistics:

The final dataset counts 4832 rows and 65 columns

The following table shows for the reviews some statistics about all the quantitative informations



	Overall_Customer_Rating	Cleanliness	Food_And_Beverages	Inflight_Entertainment	Registration	Seat_Comfort	Seat_Legroom	Value_For_Money
count	4832.000000	2463.000000	2362.000000	2931.000000	2475.000000	3115.000000	3105.000000	3086.000000
mean	6.369619	3.799432	3.196867	2.968270	3.632727	3.341573	3.377778	3.366494
std	3.137630	1.161059	1.352591	1.471231	1.413520	1.254224	1.279783	1.395117
min	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	2.000000	3.000000	2.000000	2.000000	3.000000	3.000000	3.000000	2.000000
50%	8.000000	4.000000	3.000000	3.000000	4.000000	3.000000	3.000000	4.000000
75%	10.000000	5.000000	4.000000	4.000000	5.000000	4.000000	4.000000	5.000000
max	10.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000