Case study

Main task

You will have one week to do a scrapping script for this problem. The evaluation will be done by judging the code quality and the reliability/reusability of the script - jupyter notebook will not be accepted, and clear instructions on how to run the script should be provided (don't forget the python version and the requirements.txt).

Please keep in mind that:
1. Code needs to be easy to read and well commented on.
2. Scrapping can be done in different ways. It's necessary to analyze the website and choose the best way to do it.
3. The script should be reliable. If we have some connection problems or the website is down, the script should be able to save intermediate results and continue from there in another moment.

Recomended libraries:

Pandas, Scrapy, Selenium, BeautifulSoup, Requests, Scrapy-Splash, ...

Analysis

Acquired the data, given you still have time, you can provide a jupyter notebook with some analysis of the data. A deep analysis is not mandatory, but it will be a plus. The analysis should be done so we can easily understand the data and the reviews.

Some ideas are:
1. Word cloud of the reviews.
2. Aspect-term extraction.
3. Product's most important characteristics, with its sentiments
4. ...

Website

The website to be scrapped is https://www.huffandpuffers.com/ and the items are under the category "disposables"

Deliverables

The deliverables are:
1. The script to scrap the data with clear instrunctions of how to run it.
2. The data scrapped.
3. The analysis of the data as a jupyter notebook (optional).