

Hi,

As part of our recruitment process, we would like to ask you to work on this challenge case. We expect that you spend no more than 2 to 4 hours on this assignment. We realise that this is an investment for you and you will probably take some precious time to do it, but we believe that such a process will help us to select the best candidates ... in the most objective way. Octoparse is one of the best scraping platforms, completing nicely the possibility to extract open data with python or other scripts, and knowing how to use it will be certainly a nice tool for your future projects.

**Part 1: general questions**

- a) can you explain what is the general concept of web data crawling?
- b) what is the base concept of "xpath"?
- c) can you present shortly the approach that you will be using in order to collect any data from a web site using its HTML / XML Dom tree?

**Part 2 : practical case**

In this simple study case, we would like you to build a scraper using the free version of the scraping platform 'Octoparse' (<https://www.octoparse.com/pricing>). Your target will be to extract all user reviews from a sub-category of the Amazon.fr web site.

From the menu => High-Tech : Photo et Caméscopes : **Caméras de surveillance** :

[https://www.amazon.fr/s/ref=sr\\_nr\\_n\\_0?srs=14868146031&fst=as%3Aoff&rh=n%3A13921051%2Cn%3A%2113910671%2Cn%3A13910691%2Cn%3A598657031&bbn=13910691&ie=UTF8&qid=1551293624&rnd=13910691](https://www.amazon.fr/s/ref=sr_nr_n_0?srs=14868146031&fst=as%3Aoff&rh=n%3A13921051%2Cn%3A%2113910671%2Cn%3A13910691%2Cn%3A598657031&bbn=13910691&ie=UTF8&qid=1551293624&rnd=13910691)

For every product & user review you will have to extract the following fields:

Product Name
Amazon Product Code (ASIN)
Price
Average User Rate for the Product
Total Number of User Reviews for the Product
Name of the User posting the Review
Rate (equivalent to 1 to 5 stars)
Date of the Review
Short title of the Review
Full title of the Review

Please send us the extracted data in an Excel file, as well as a file with the Octoparse 'task' exported from the system.

Have a good fun and we are looking to meet you and discuss with you about your approach, choices and ideas!

*The Innoscape team*