leboncoin.fr

lobstr.io

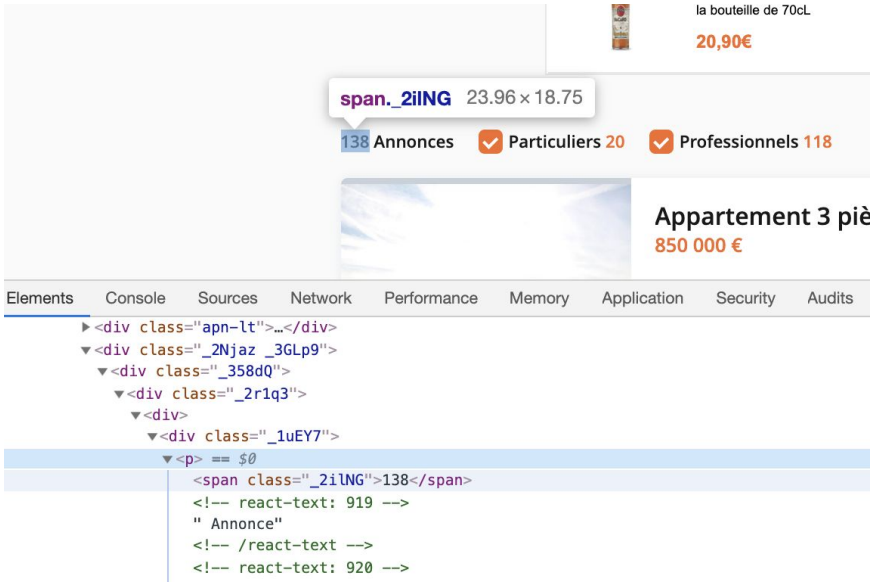
The purpose of this case is to test your abilities through 3 main area of competences :
- XPath
- Regular Expression (Regex)
- Requests/LXML with Python

You will find a doc named sample.html which is the source code of the following URL : https://www.leboncoin.fr/recherche/?category=9&locations=Cassis_13260.

1. XPATH

Find the right XPath to find the very element in **sample.html**
https://www.w3schools.com/xml/xpath_intro.asp
Google Sheet (IMPORTXML)

1.a. (Example) Find the total number of “Annonces”



//span[@class=”_2ilNG”]/text()

1.a Find XPath for title of annonce
1.b Find XPath for price of annonce
1.c Find XPath for kind of annonce (“Pro”, bold)
1.d Find XPath for type of annonce (“Ventes Immobilières”)
1.e Find XPath for city of annonce
1.f Find XPath for zip code of annonce
1.g Find XPath for date

NB : one XPath must work for at least 2 URLs, and return a list of values for every annonce

2. REGEX

Find the right Regex to find the element we are looking for in **sample.json**
https://regex101.com/

2.a. (Example) Find the total number of “Annonces”

```
33
34      ],
35      "data":{
36          "total":138,
37          "total_all":138,
38          "total_pro":118,
39          "total_private":20,
40          "total_active":0,
41          "total_inactive":0,
42          "pivot":"35:1554795997000",
43          "ads":[
44              {
```

(?<=\"total\"\:)\d+

2.b. Find the total number of private “Annonces” : “total_private”
/* ECRIRE LE REGEX ICI */

2.c. Find the total number of pro “Annonces” : “total_pro”
2.d. Find all the unique IDs of “Annonce” : “list_id”

```
306                 "list_id":1598602783,
```

2.e. Find the prices : “price”

```
317                 "price":[
318                     420000
319                 ],
```

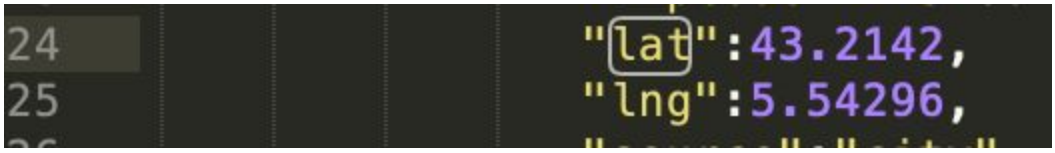
2.f. Find the subject : “subject”

```
456                 "subject":"Appartement 3 pièces 68 m²",
```

2.g. Find the number of rooms : “rooms”

2.h.a Find lat

2.h.b Find lng



3. REQUESTS/LXML

In this final case, you have to build a script, in Python, using **REQUESTS/LXML**, that will scrape annonces data from https://www.leboncoin.fr/recherche/?category=9&locations=Cassis_13260, from page 1 to page 4.

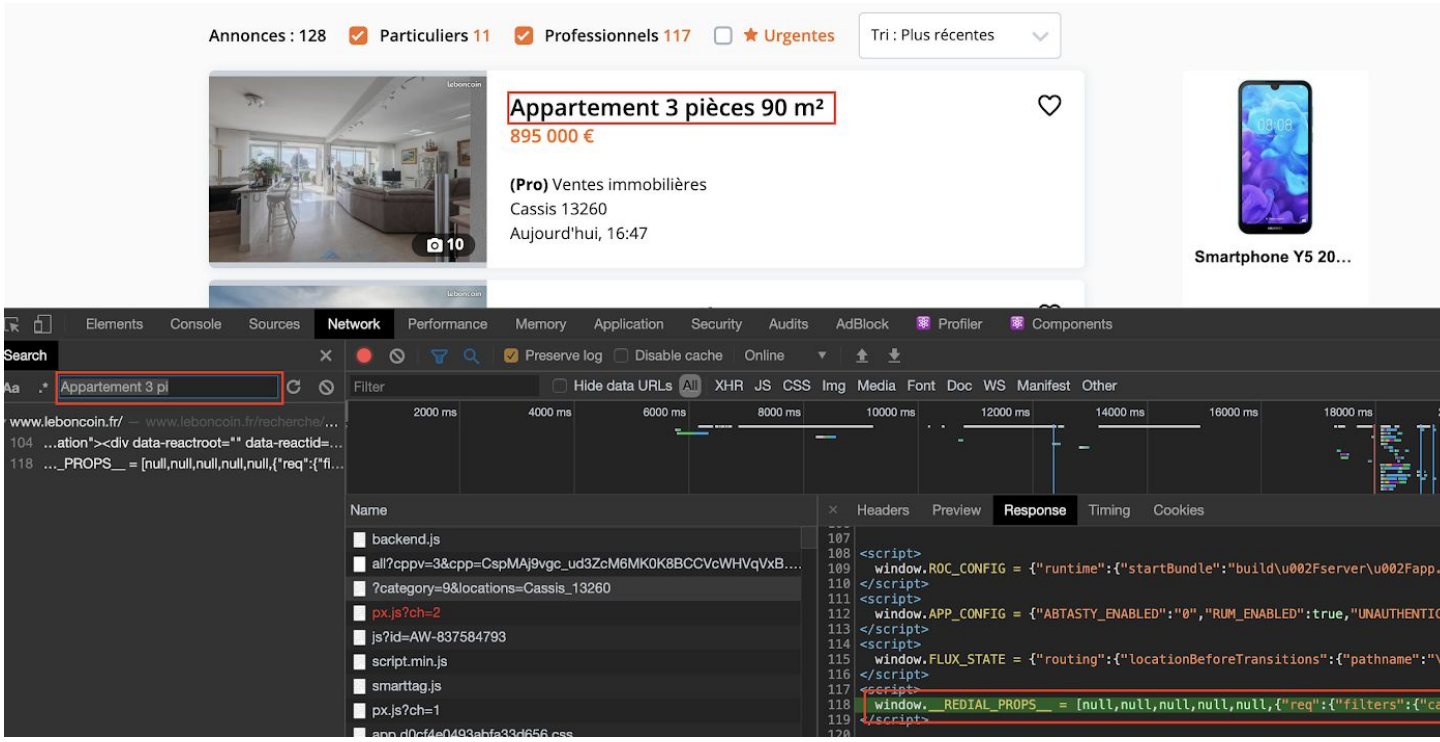
- write script.py able to scrape data
- save data in .csv file
- host your script on Github
- send to us :)

An already made extract is available : **output.xlsx**.

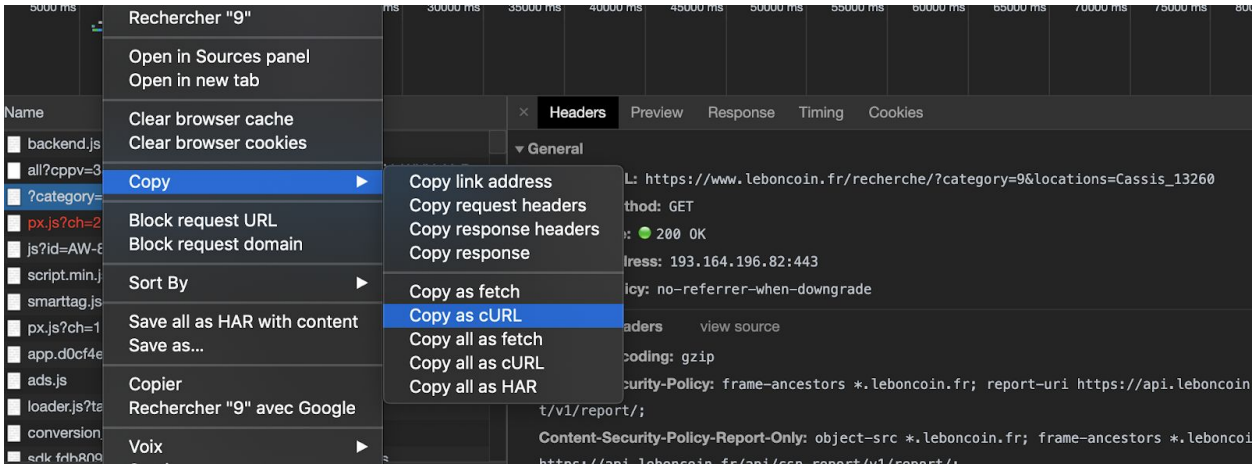
TECHNICAL ADVICE

NB: https://www.leboncoin.fr/ is protected with *datadome*, French leading bot-mitigation services. Please, control your request speed and request attributes — headers, cookies...

Use the **Google Chrome Search Tool**, to search the data you want to parse, and identify the very request that originated it.



Thereafter, use **Copy as CURL**, to get the nature of the request sent: URL, headers, cookies etc.



Endly, **convert it to `requests`** with https://curl.trillworks.com/: