B0 - HUB

B-INN-000

Introducing to Web scraping and monitoring

Create your own monitor

INTRODUCTION

The goal of this workshop is to introduce the web scraping and the monitoring of web pages.

Web scraping means that you will scan a web page in order to recover information present on it

To do this, we will use Python with requests and beautifulsoup4 libs.

SET-UP

You must git clone this repository

```
Terminal
~/B-INN-000> pip install -r requirements.txt
```

You have now downloaded the required libraries for the project.

STARTING

In the first step, we will try to just discover what is web scraping.

Start your project by importing the libraries.

```
from bs4 import BeautifulSoup
import requests
import sys
```

Now let's see the requests's documentation

Be simple and try to just gest a request with a simple url. Do not hesitate to ask questions !

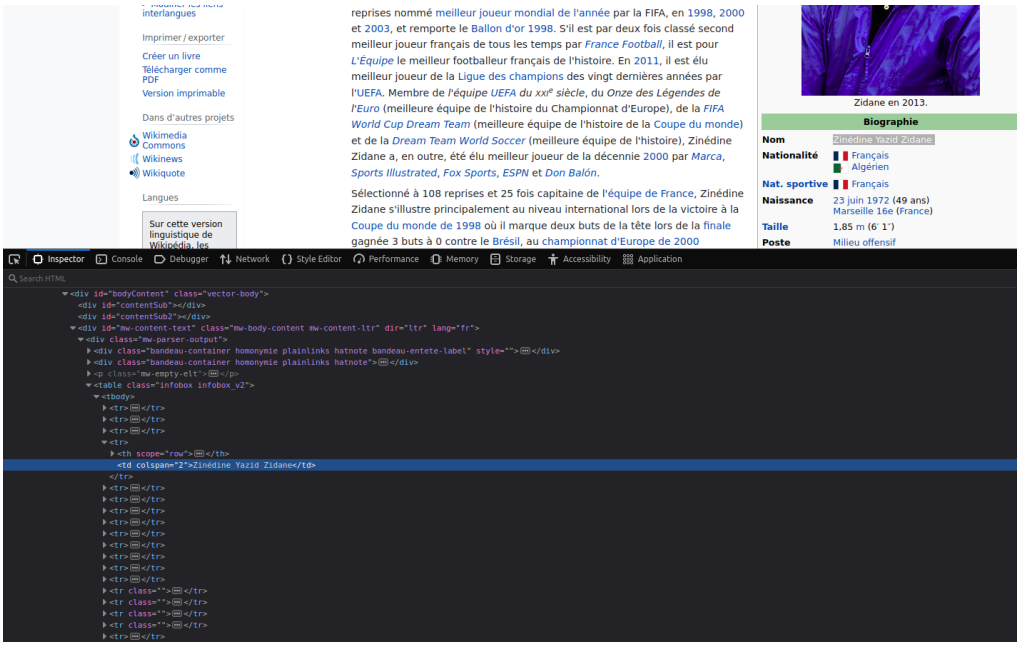
It should look like this..

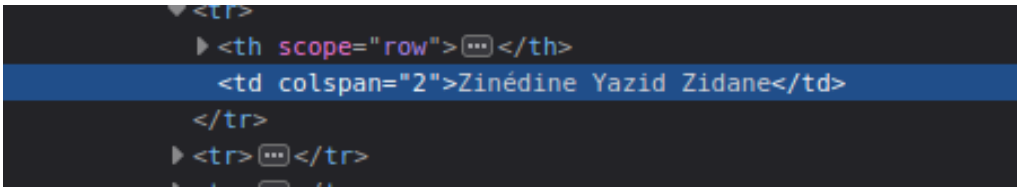
```
from bs4 import BeautifulSoup
import requests
import sys

response = requests.get('your_url')

if response.status_code != 200:
    print(f'response code = {response.status_code}')
    sys.exit(84)
next...
```

Check status code value, sometimes website blocks scraping(ex: Linkedin), so you have to go another way

Wikipedia's web pages are easy to start with !

OK now you know how to make requests so what?
Take a look at the following picture.





So if I want to recover the “Nom”, It should look like this.

```
name = soup.find('td', {'colspan':'2'}).text
print(name)
```

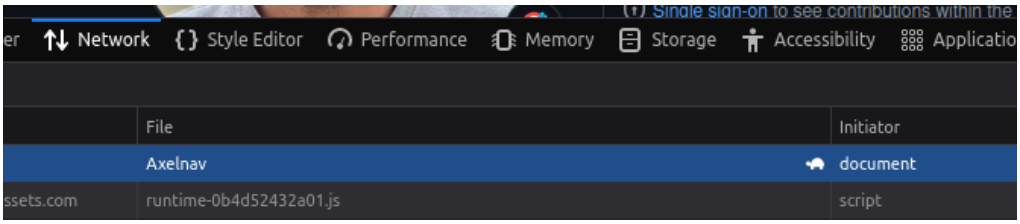
```
Terminal
~/B-INN-000> python3 yourfilename.py
Zinédine Yazid Zidane
```

In order to go further away, you can try to recover the Zidane’s picture, the “Poste”, etc.. .

GITHUB SCRAPER

Now, we will see how to scrap when the website is hardest than Wikipedia. Like when you need to log in. For the example, we will use GitHub, so go to your Github profile.

So get back to the Inspector and go in the network tab and refresh your page if it seems empty. Look in the initiator column for the type “document”. It should look like this.



Now make a right click on this document and select Copy->Copy as cURL.
From now on, go to Curl converter, paste it and select “Python”.

Now copy and paste the result in your_file_py.

It should look like this.

```
from bs4 import BeautifulSoup as bs
import requests

cookies = {
    ... ,
}

headers = {
    ... ,
}

response = requests.get(url=url, cookies=cookies, headers=headers)
```

MONITORING

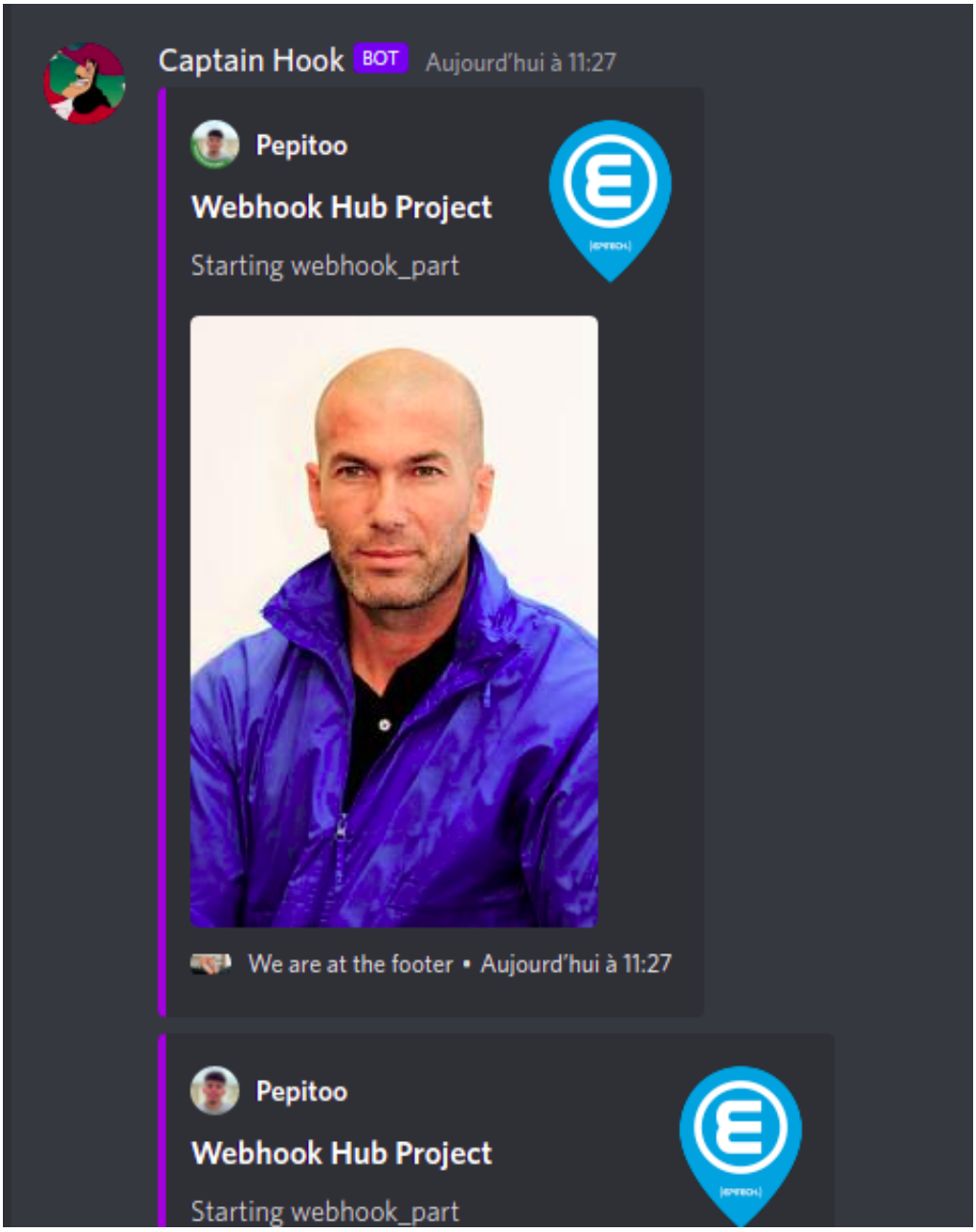
Monitoring means you keep an eye on.

That why we can use Discord in order to make some Webhook which will alert us.

We will use Discord Hook

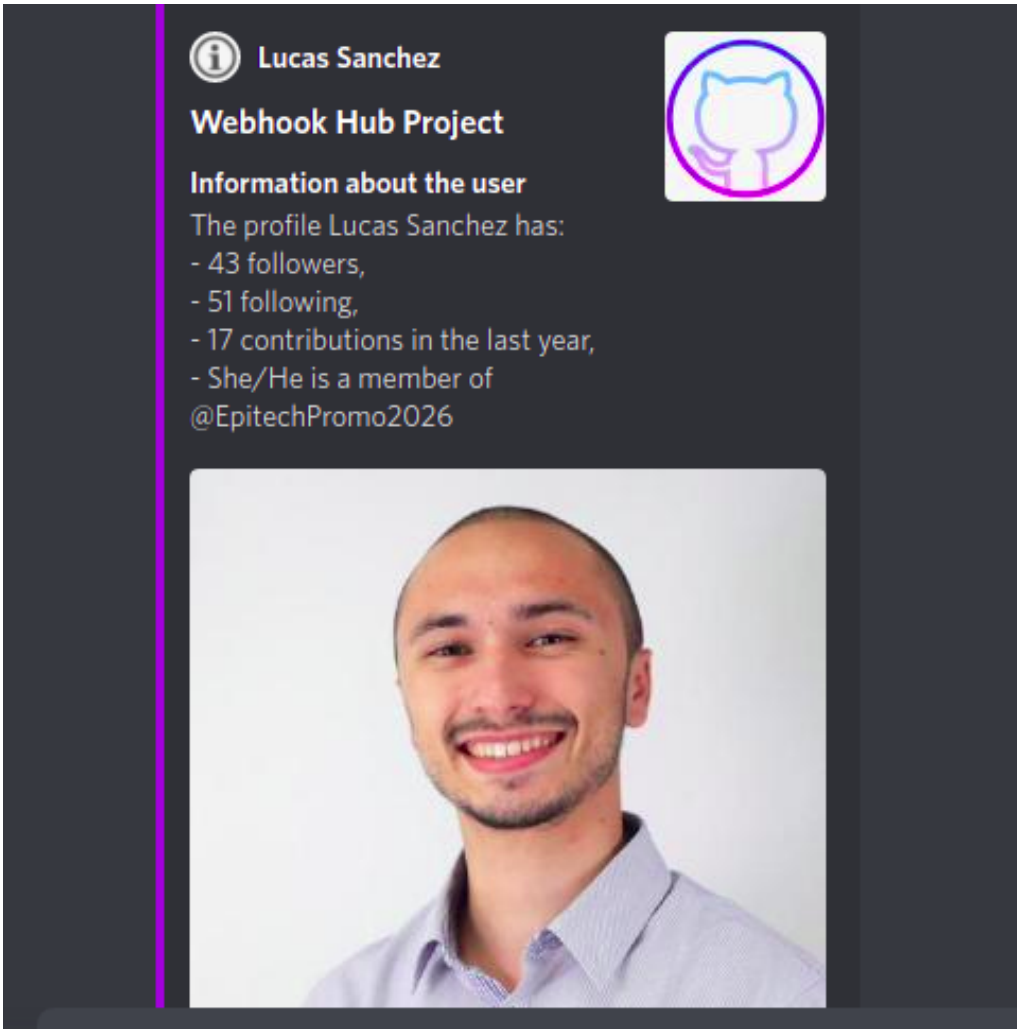
```
from discord_webhook import DiscordWebhook, DiscordEmbed

webhook = DiscordWebhook(url='your_url_webhook')

embed = DiscordEmbed(title='your_title', description='Starting webhook_part', color='
    your_color_in_hexa')
embed.set_author(name='your_name', url='your_url', icon_url='your_ico_url')
embed.set_image(url="https:"+link_of_zizou_head)
embed.set_thumbnail(url='your_thumbnail_url')
embed.set_footer(text='footer_text', icon_url='url_photo_footer')
embed.set_timestamp()

webhook.add_embed(embed)

response = webhook.execute()
```

Example of Webhook sends with the Zinédine Zidane's Wikipedia.



Try to go further !

An example, below shows you scraping on Github.



A monitor it simply a loop which compares previous and last data.

Try to discover the libraries and continue on your way.

Do not hesitate to ask me some questions !

Thanks for listening and I hope you discover something which was interesting for you !