Web scrapping behind authentication

Web scrapping of a private GitHub repository with `requests` and `Beautiful Soup`.

1. First steps

In [115]:
```
import requests
from bs4 import BeautifulSoup
```

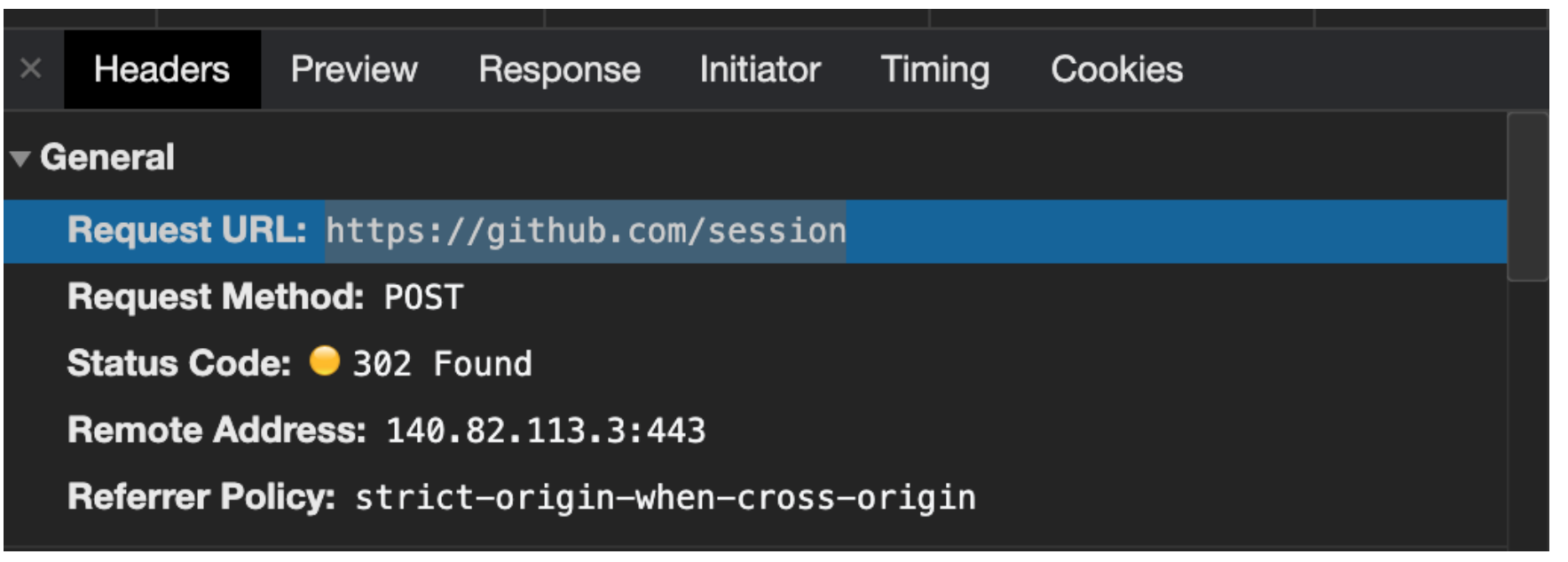
2. Getting data

We need to know two things:

1. The `URL` to which the POST request (sign in) will be sent.
2. The payload, or data that will be sent.

How?

1. Clear all the data from the *network* tab in the developer tools and sign in.
2. Go to *session* and save the URL. In this case: https://github.com/session



In [116]:
```
url = "https://github.com/session"

# Create the session object
s = requests.Session()

response = s.post(url)
```

In [117]:
```
response
```

Out[117]:
```
<Response [403]>
```

It doesn't work. We need to pass it more information. Go to the network tab again and right click to *session*, then *copy*, and then *copy as cURL*.

You can paste it into this website: https://curl.trillworks.com/, to get the request written in Python.

From that request we need two things: `the cookies and the data`.

In [ ]:
```
# Complete this with your personal session data.

cookies = {
    #...
}

headers = {
    #...
}

data = {
    #...
}

s = requests.Session()

response = requests.post('https://github.com/session', headers=headers, cookies=cookies, data=data)
```

Now we're signed in. Let's request something from a repository.

I choosed this one because is a `private repository`.

In [119]:
```
s = requests.Session()

s.post('https://github.com/session', data=data, cookies=cookies, headers=headers)
result = s.get("https://github.com/paulawoloszyn/php_login")
result
```

Out[119]:
```
<Response [200]>
```

It worked! Now let's scrap something.

3. Scrapping

We'll start using the Beautiful Soup library. We already have in the 'result' variable the content of our page.

In [120]:
```
# Save the content of the website on a new variable.

content = result.content

# And create the soup.

soup = BeautifulSoup(content)

# print(soup.prettify())
```

Some example of scrapping, to check if we are on the website.

In [136]:
```
samples = soup.find_all("h2")
# print(samples)
```

In [141]:
```
for a in samples:
    a = a.get_text().strip()
    print(a)
```

```
Learn Git and GitHub without any code!
Latest commit
Git stats
Files
Drop to upload your files
About
Releases
Packages 0
Languages
```