Scrapping

Internet rvest selenium
web technologies

Чувакин Сергей

R meetup

7 декабря 2019 г.

Internet

WWW vs. Internet ?

Internet

- Серверы
- Клиенты
- Третья сторона...

HTML
CSS

HTML - язык разметки, который отвечает за содержание страницы.

CSS - язык стилей, который отвечает, за то, как будет выглядеть страница.

HTML

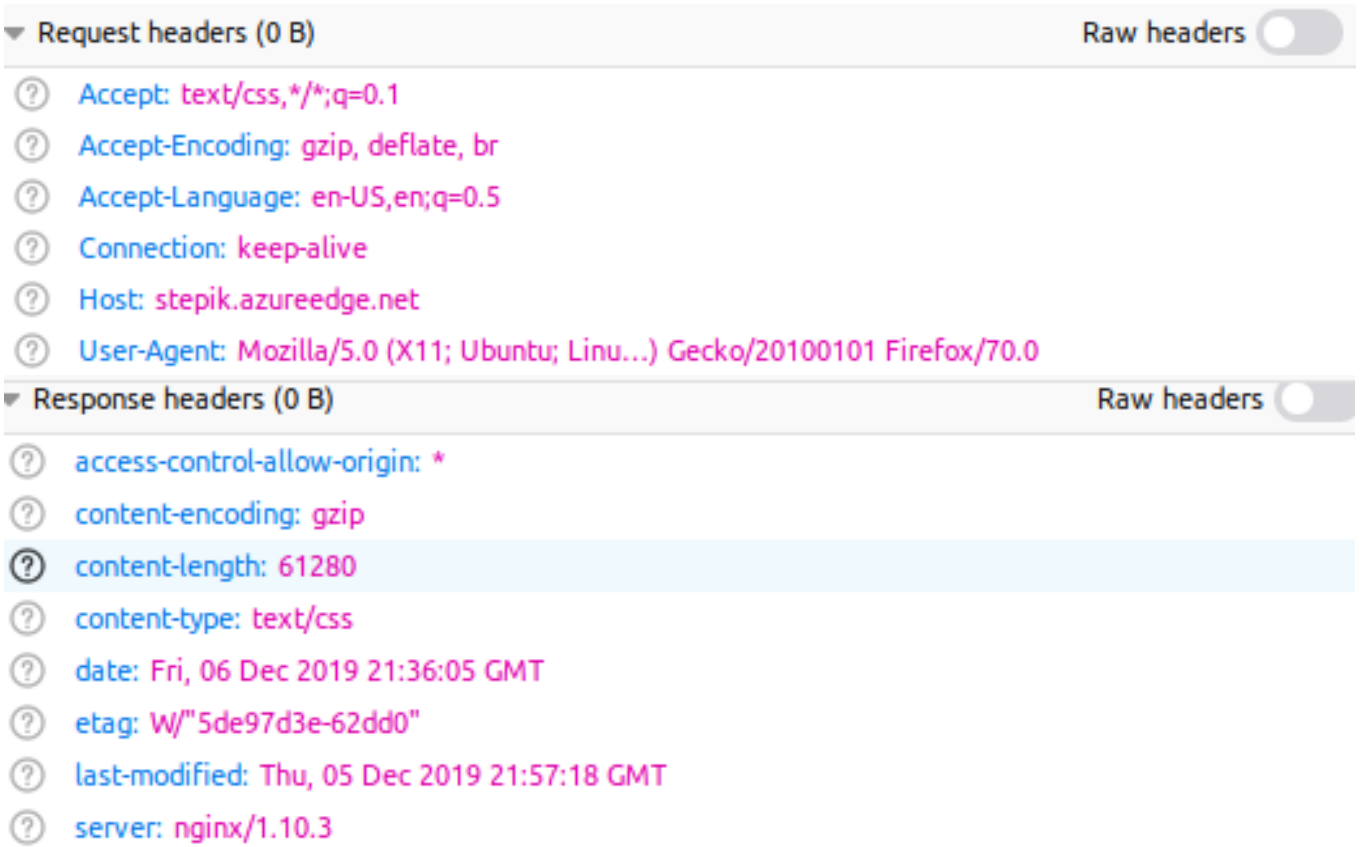
CSS

Пример: ...

Интернет запрос - структурированная конструкция позволяющая получать содержимое страниц с удаленных серверов.
Сценарий: request - response.

- headers
- body
- cookies
- data
- parameters (params)
- cache

headers

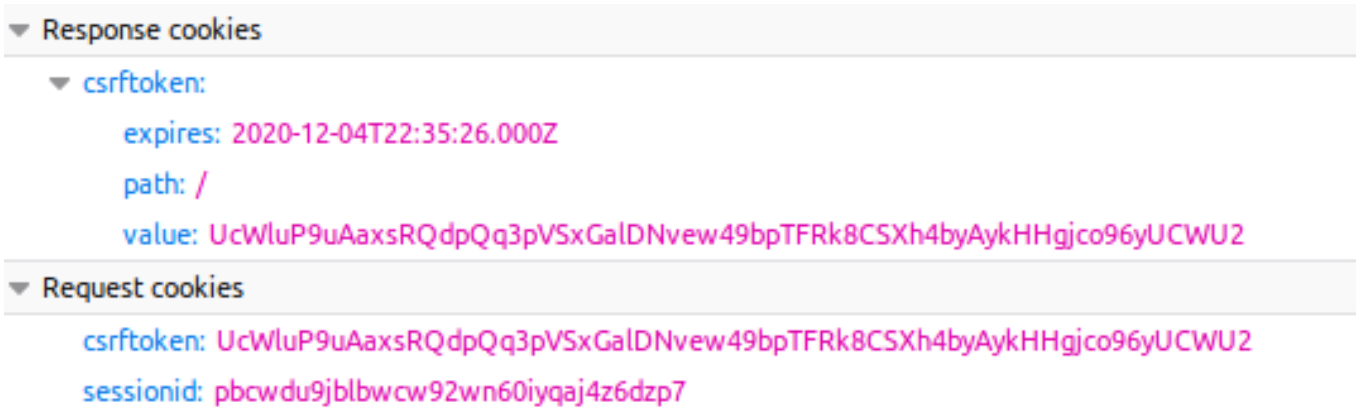

Request headers (0 B)
Raw headers
Accept: text/css,*/*;q=0.1
Accept-Encoding: gzip, deflate, br
Accept-Language: en-US,en;q=0.5
Connection: keep-alive
Host: stepik.azureedge.net
User-Agent: Mozilla/5.0 (X11; Ubuntu; Linu...) Gecko/20100101 Firefox/70.0
Response headers (0 B)
Raw headers
access-control-allow-origin: *
content-encoding: gzip
content-length: 61280
content-type: text/css
date: Fri, 06 Dec 2019 21:36:05 GMT
etag: W/"5de97d3e-62dd0"
last-modified: Thu, 05 Dec 2019 21:57:18 GMT
server: nginx/1.10.3

body

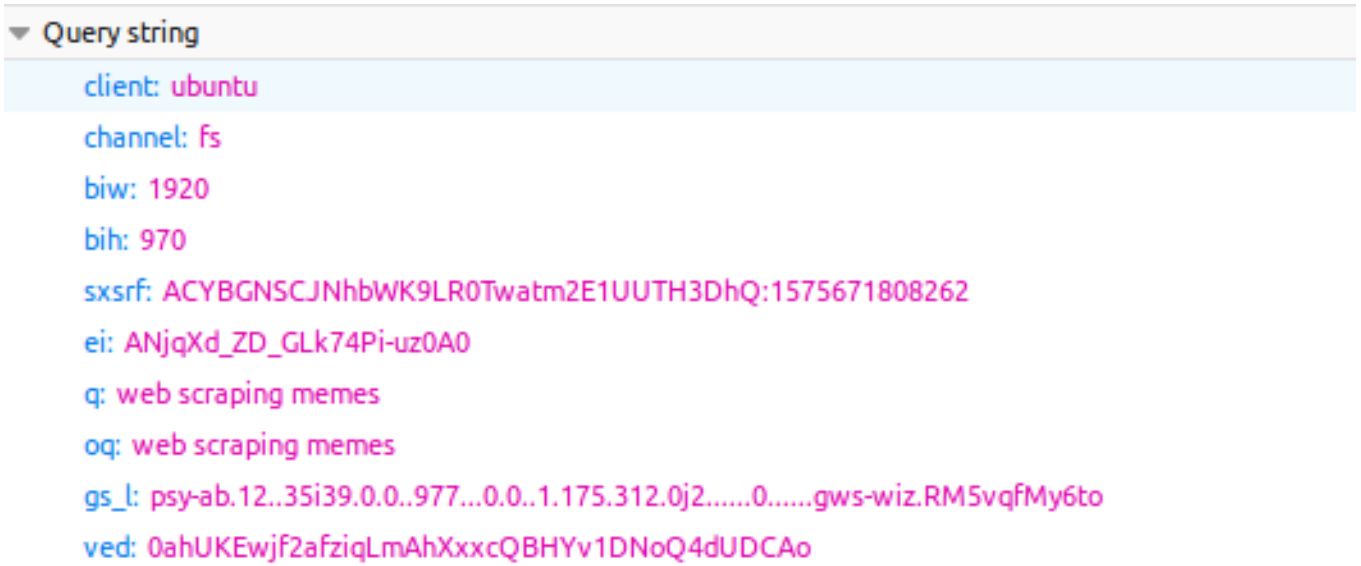
Методы:

- GET
- HEAD
- POST
- PUT
- PATCH
- DELETE
- etc

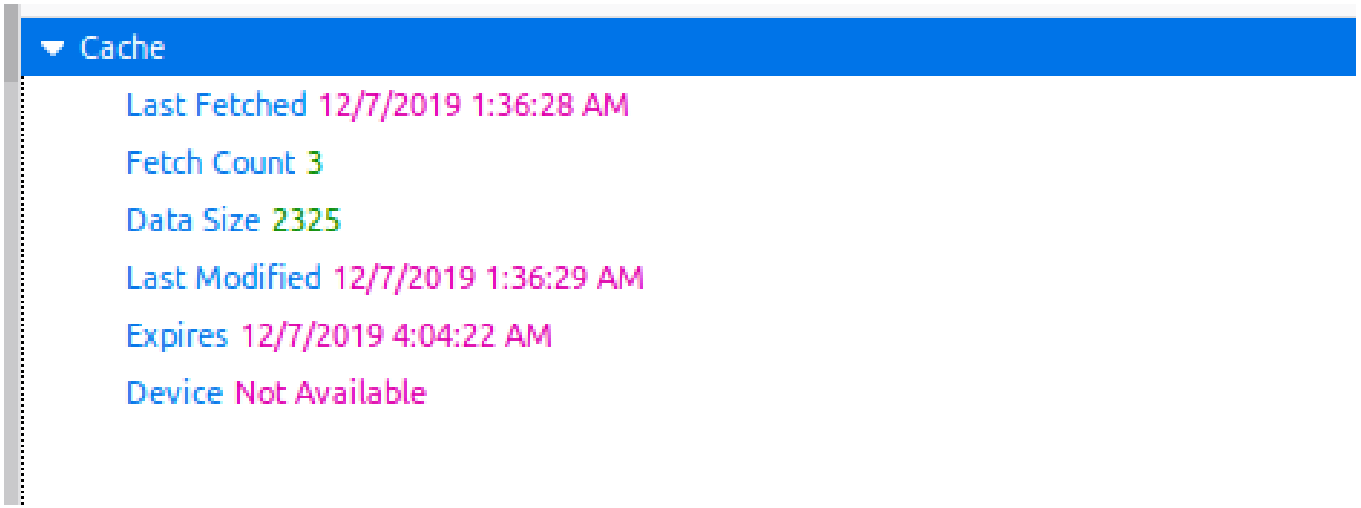
cookies


Response cookies
csrftoken:
expires: 2020-12-04T22:35:26.000Z
path: /
value: UcWluP9uAaxsRQdpQq3pVSxGalDNvew49bpTFRk8CSXh4byAykHHgjco96yUCWU2
Request cookies
csrftoken: UcWluP9uAaxsRQdpQq3pVSxGalDNvew49bpTFRk8CSXh4byAykHHgjco96yUCWU2
sessionid: pbcwdu9jblbwcw92wn60iyqaj4z6dzp7

params


Query string
client: ubuntu
channel: fs
biw: 1920
bih: 970
sxsrf: ACYBGNSCJNhbWK9LR0Twatm2E1UUTH3DhQ:1575671808262
ei: ANjqXd_ZD_GLk74Pi-uz0A0
q: web scraping memes
oq: web scraping memes
gs_l: psy-ab.12..35i39.0.0..977...0.0..1.175.312.0j2.....0.....gws-wiz.RM5vqfMy6to
ved: 0ahUKEwjf2afziqLmAhXxxcQBHYv1DNoQ4dUDCAo

Cache


Cache
Last Fetched 12/7/2019 1:36:28 AM
Fetch Count 3
Data Size 2325
Last Modified 12/7/2019 1:36:29 AM
Expires 12/7/2019 4:04:22 AM
Device Not Available

Что еще?

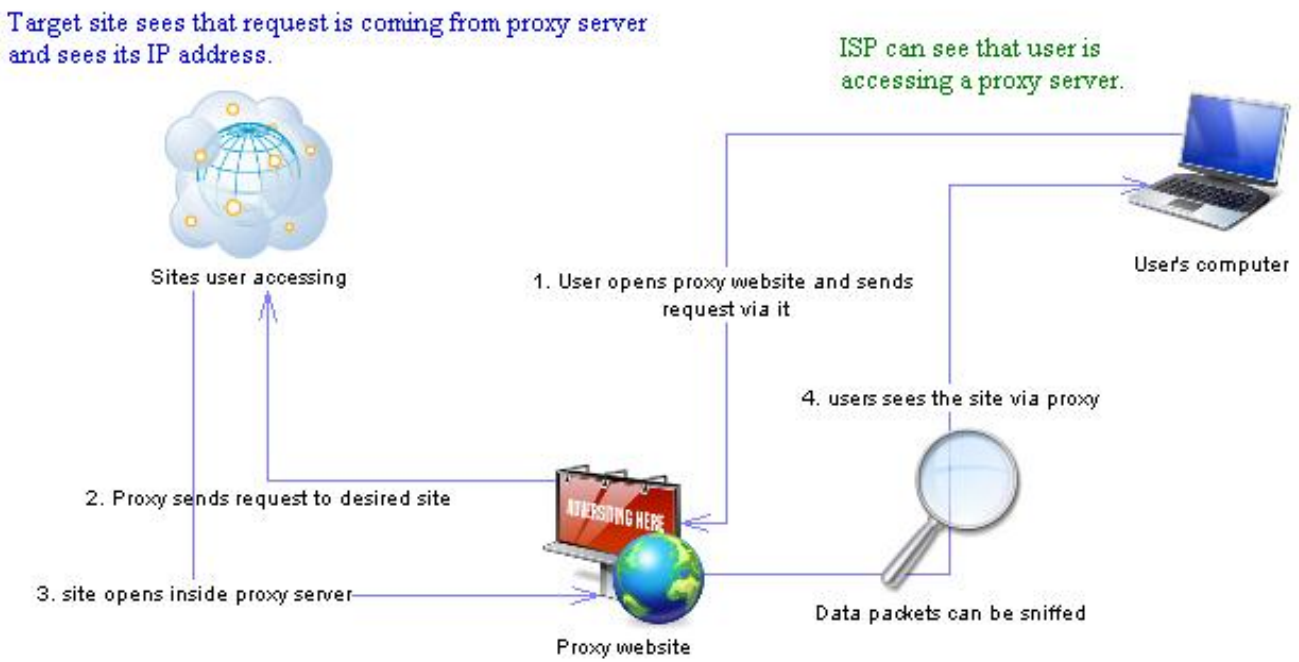
- Протоколы
- Архитектура DOM
- MVC фреймворки
- frontend-backend
- Sessions
- Работа с СУБД
- etc

AJAX

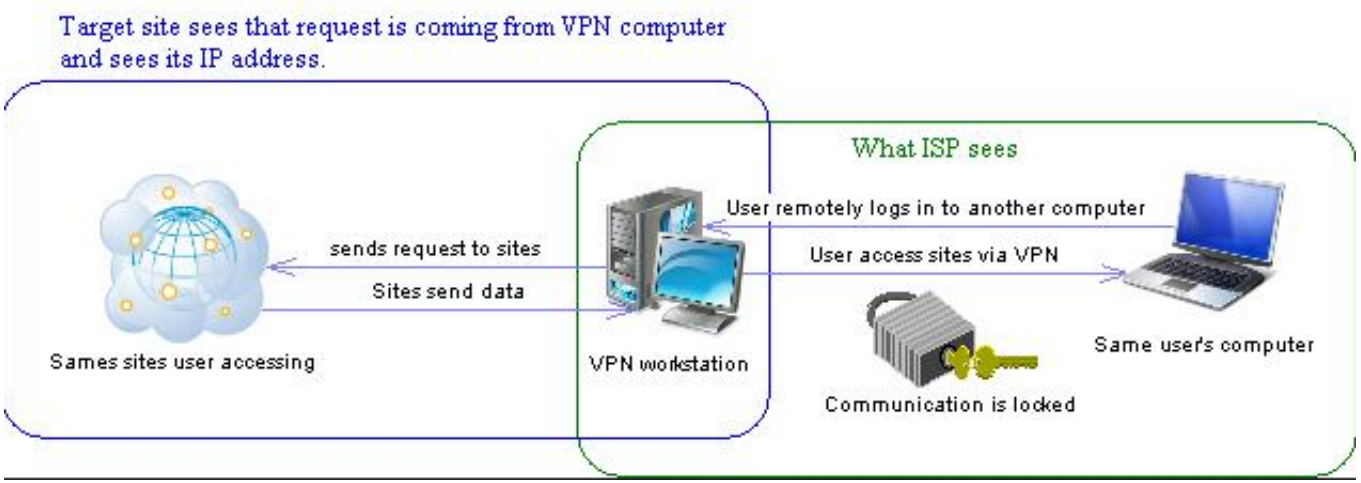
AJAX - технология верстки страниц с помощью ЯП (java script).

пример - http://zpp.rospotrebnadzor.ru/Forum/Appeals

proxy


Target site sees that request is coming from proxy server and sees its IP address.
ISP can see that user is accessing a proxy server.
Sites user accessing
User's computer
1. User opens proxy website and sends request via it
4. users sees the site via proxy
2. Proxy sends request to desired site
3. site opens inside proxy server
Data packets can be sniffed
Proxy website

vpn


Target site sees that request is coming from VPN computer
and sees its IP address.
What ISP sees
User remotely logs in to another computer
User access sites via VPN
sends request to sites
Sites send data
Sames sites user accessing
VPN workstation
Same user's computer
Communication is locked