

Forum:[Hack Forums](#)

Thread: [Best way to scrape protected website](#)

arxadaer - February 1, 2020, 7:44 am

Hi,
as the title says, I'm trying to scrape a website that block me.
To do this, I would like to use a simple http request library (like requests and session from python) and not a browser simulator (like Selenium, Puppeteer, PhantomJS, etc...). I already tried successfully with those, but I really want to do it with an http request.
I have already set the basic stuff, like changing user-agent, refer, and other headers. I have "top of the shelf" proxies, too. The website in question is very likely hosted by cloudflare, so they're protecting it in some other ways for sure.
My main question is: what should I focus on? Should I find someone path of the server that are not protected, or maybe find some work around with cookies?
Let me know if you have other advices, too :)

I would be very grateful if someone would help me for free, but since I need to find a serious and reliable solution, the help could also be paid.

Thank you

SIM-3991524 - February 20, 2020, 8:55 pm

If you're just going after some text you could take a take a "full page" screenshot and then pass that through ocr (see pytesseract). From there you could use a regex

hannalexise - February 24, 2020, 3:43 am

You should also mimic the human behavior as well. For example, instead of scraping the same sort of sites, it is better to mix it with other unrelated stuff to prevent it from being detected. Some site has the mechanism to detect such abnormal behavior e.g., someone only request a certain sort of url, etc. Also give them an appropriate delay between requests.

Ri?dle - February 24, 2020, 1:01 pm

I don't know how to get around the cloudflare issue while using the python requests module, but if the reason you're not wanting to use selenium is because of selenium is popping up a visible browser on your screen you can do

```
Code chrome_options = Options()
chrome_options.add_argument("--headless")
driver = webdriver.Chrome(chrome_options=chrome_options)and I believe it's different for other
firefox, and other browsers, just type in python selenium [your browser name] headless into google
and it'll show you how,
```

or if your problem with using selenium is that you don't like the way you're scraping text using selenium's methods (I.E Code find_element_by_id, find_elements_by_class, find_element_by_xpath ect..) you could always pass the page source Code driver.page_source to BeautifulSoup as you normally would do with requests

gua - February 24, 2020, 2:36 pm

Idk what programming language u know but u can use delphi for this, and indy components.
Changing proxy in every 5-10 requests, also changing requests that u are sending (randomize time

from-to) ..

Also randomizing user agents etc.. By doing this u can basically scrape 24/7

if u are interested i can show u on pm some scraper that i made ?

SIM-3991524 - February 24, 2020, 5:38 pm

(Feb 24, 2020 02:01 PM)Ri?dle Wrote: [\[CITE\]](#) I don't know how to get around the cloudflare issue while using the python requests module, but if the reason you're not wanting to use selenium is because of selenium is popping up a visible browser on your screen you can do

```
Code chrome_options = Options()
chrome_options.add_argument("--headless")
driver = webdriver.Chrome(chrome_options=chrome_options)and I believe it's different for other
firefox, and other browsers, just type in python selenium [your browser name] headless into google
and it'll show you how,
```

or if your problem with using selenium is that you don't like the way you're scraping text using selenium's methods (I.E

Code find_element_by_id, find_elements_by_class, find_element_by_xpath ect..) you could always pass the page source

Code driver.page_source to BeautifulSoup as you normally would do with requests
yep, selenium is extra easy in python, would recommend this method assuming the site is not too cryptic