

Forum:[Hack Forums](#)

Thread: [Downloading source code automatically with python](#)

0u1law - June 25, 2020, 5:42 am

hey guys.

i need some help in the following case:

i want to write a program that uses my browser. i have a list of links i want to visit and download the source code automatically. thats it.

which packages do i need for this ?

if someone have a piece of code i would be glad!

PSEUDO-CODE:

```
list_of_links = [... ... ... ...]
for each in list_of_links:
----open browser
----show sourcecode
----download sourcecode
----close browser
```

and if there are another ways to do this in a more efficient way please let me know

and just for the curiosity is there a method to do this with bash scripting?

best regards

0u1law

Sickness - June 25, 2020, 5:52 am

To automate it better you can use Selenium, just initialize a driver, save the sourcecode on a string and save it on a file and done.

Harley Quinn - June 25, 2020, 10:43 am

(Jun 25, 2020 05:42 AM)0u1law Wrote: [\[CITE\]](#) hey guys.

i need some help in the following case:

i want to write a program that uses my browser. i have a list of links i want to visit and download the source code automatically. thats it.

which packages do i need for this ?

if someone have a piece of code i would be glad!

PSEUDO-CODE:

```
list_of_links = [... ..]
for each in list_of_links:
----open browser
----show sourcecode
----download sourcecode
----close browser
```

and if there are another ways to do this in a more efficient way please let me know

and just for the curiosity is there a method to do this with bash scripting?

best regards

Ou1law
Hi ?

Yes You can use Selenium, or Requests w/ BeautifulSoup.

Hit me up with a DM and i'll see what i can do for you. I'll show you how to do it also so you know and can show someone else one day!

Harley

4Txe4y33C9zvNhMj - June 25, 2020, 11:31 am

(Jun 25, 2020 05:52 AM)Sickness. Wrote: [\[CITE\]](#) To automate it better you can use Selenium, just initialize a driver, save the sourcecode on a string and save it on a file and done.
Selenium is extremely inefficient.

Selenium was made for automated quality assurance, not for datascraping. Since OP just wants the actual source code, he can just use the Python.requests module to download that code.

This should work:
Code import requests

```
list_of_links = ["https://hackforums.net/", "https://google.com/", "https://wikipedia.org/"]
```

```
for link in links:
    result = requests.get(link)
    print("downloading link...")
    #do whatever you want with the content
    with f as open(link + "_download.txt", "w+")
        f.write(result)
```

h0lyl337 - June 25, 2020, 11:42 am

(Jun 25, 2020 11:31 AM)4Txe4y33C9zvNhMj Wrote: [\[CITE\]](#) Selenium is extremely inefficient.

Selenium was made for automated quality assurance, not for datascraping. Since OP just wants the actual source code, he can just use the Python.requests module to download that code.

This should work:
Code import requests

```
list_of_links = ["https://hackforums.net/", "https://google.com/", "https://wikipedia.org/"]
```

```
for link in links:
```

```
    result = requests.get(link)
```

```
    print("downloading link...")
```

```
    #do whatever you want with the content
```

```
    with f as open(link + "_download.txt", "w+")
```

```
        f.write(result)
```

i agree, requests would be perfect, bypassing the web browser and use less resource.

Sickness - June 25, 2020, 11:44 am

(Jun 25, 2020 11:31 AM)4Txe4y33C9zvNhMj Wrote: [\[CITE\]](#) Selenium is extremely inefficient.

Selenium was made for automated quality assurance, not for datascraping. Since OP just wants the actual source code, he can just use the Python.requests module to download that code.

This should work:

Code import requests

```
list_of_links = ["https://hackforums.net/", "https://google.com/", "https://wikipedia.org/"]
```

```
for link in links:
```

```
    result = requests.get(link)
```

```
    print("downloading link...")
```

```
    #do whatever you want with the content
```

```
    with f as open(link + "_download.txt", "w+")
```

```
        f.write(result)
```

You are right, but for pages that got protection for bots requests wont work always.

4Txe4y33C9zvNhMj - June 25, 2020, 12:49 pm

(Jun 25, 2020 11:44 AM)Sickness. Wrote: [\[CITE\]](#) You are right, but for pages that got protection for bots requests wont work always.

In the end you can always fake the actual HTTP request your browser sends. Even if they dynamically get content from an API endpoint you can scrape that too.

Pawnhub - June 25, 2020, 11:43 pm

(Jun 25, 2020 11:44 AM)Sickness. Wrote: [\[CITE\]](#) You are right, but for pages that got protection for bots requests wont work always.

Selenium wouldn't do any better in this case. I get more sites blocking selenium when cracking then request's.

But if you could 100% be able to mimic a proper request (which can be outstandingly hard with hidden variables and cookies etc) request's would still be the most viable.

Although, selenium can read JS unlike request's. So if the form he wants to scrape from loads from a lframe then yes selenium is most viable since request's can't.

SIM-4771460 - July 4, 2020, 4:48 pm

Requests or Scrapy can do your Job.

I suggest selenium because It handles Javascript easily, and you can use headless chrome/firefox driver for speed,

0u1law - July 9, 2020, 12:46 pm

(Jul 4, 2020 04:48 PM)Gosyne ? Wrote: [\[CITE\]](#) Requests or Scrapy can do your Job.

I suggest selenium because It handles Javascript easily, and you can use headless chrome/firefox driver for speed,
how can i use it headless?

SIM-4771460 - July 9, 2020, 12:47 pm

(Jul 9, 2020 12:46 PM)Ou1law Wrote: [\[CITE\]](#) how can i use it headless?

you can use headless chrome/firefox, just put --headless argument in code, or some headless browser.

Ou1law - July 9, 2020, 12:48 pm

(Jul 9, 2020 12:47 PM)Gosyne ? Wrote: [\[CITE\]](#) you can use headless chrome/firefox, just put --headless argument in code, or some headless browser.
thanks a lot!