Scrapy and proxies

Asked 14 years, 2 months ago Modified 3 years, 11 months ago Viewed 82k times

54

How do you utilize proxy support with the python web-scraping framework Scrapy?

python scrapy

Share Improve this question Follow

edited Mar 18, 2011 at 16:06
bdd
3,434 5 33 43

asked Jan 17, 2011 at 6:17
no1
945 2 9 9

1 you need to be more specific with your question... – radman Jan 17, 2011 at 6:26

9 Answers

Sorted by: Highest score (default)

56

Single Proxy

1. Enable `HttpProxyMiddleware` in your `settings.py` , like this:

```
DOWNLOADER_MIDDLEWARES = {
    'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware': 1
}
```

2. pass proxy to request via `request.meta` :

```
request = Request(url="http://example.com")
request.meta['proxy'] = "host:port"
yield request
```

You also can choose a proxy address randomly if you have an address pool. Like this:

Multiple Proxies

```
class MySpider(BaseSpider):
    name = "my_spider"
    def __init__(self, *args, **kwargs):
        super(MySpider, self).__init__(*args, **kwargs)
        self.proxy_pool = ['proxy_address1', 'proxy_address2', ...,
'proxy_addressN']

    def parse(self, response):
        ...parse code...
        if something:
            yield self.get_request(url)

    def get_request(self, url):
        req = Request(url=url)
        if self.proxy_pool:
            req.meta['proxy'] = random.choice(self.proxy_pool)
        return req
```

Share Improve this answer Follow

edited Feb 25, 2018 at 7:33
Kurt Peek
57.8k 102 345 564

answered Dec 16, 2013 at 10:25
Amom
709 5 6

13 The documentation says that the `HttpProxyMiddleware` is setting the proxy inside every Requests meta attr, so enabling ProxyMiddleware AND setting it manually would make no sense – Rafael T Dec 22, 2014 at 20:16

1 I should have copied this code. I glanced it and then coded myself, but proxy functionality was not working. Now I see the proxy value was set to `request.headers` instead of `request.meta` . Stupid me (face palm)! I went to see the `HttpProxyMiddleware` code, it skips if someone has already set `request.meta['proxy']` , so there is no need to list it in the settings github.com/scrapy/scrapy/blob/master/scrapy/... – TG Gowda Jul 21, 2017 at 3:48

1 I am not sure I understand the difference between the two, is `BaseSpider` your original spider and `MySpider` or is `MySpider` is the actual modified spider and `BaseSpider` refers to `scrapy.Spider` ? – ishandutta2007 Dec 19, 2019 at 10:46

1 Modern version does not have this field anymore `AttributeError: 'Request' object has no attribute 'meta'` . Likely now it should look as this `yield scrapy.Request(url=url, callback=self.parse, meta={'proxy': proxy})` – Gleichmut Sep 22, 2023 at 13:17

From the Scrapy FAQ,

54

Does Scrapy work with HTTP proxies?

Yes. Support for HTTP proxies is provided (since Scrapy 0.8) through the HTTP Proxy downloader middleware. See `HttpProxyMiddleware` .

The easiest way to use a proxy is to set the environment variable `http_proxy` . How this is done depends on your shell.

```
C:\>set http_proxy=http://proxy:port
csh% setenv http_proxy http://proxy:port
sh$ export http_proxy=http://proxy:port
```

if you want to use https proxy and visited https web,to set the environment variable `http_proxy` you should follow below,

```
C:\>set https_proxy=https://proxy:port
csh% setenv https_proxy https://proxy:port
sh$ export https_proxy=https://proxy:port
```

Share Improve this answer Follow

edited Jun 20, 2020 at 9:12
Community Bot
1 1

answered Jan 17, 2011 at 6:29
ephemient
205k 39 288 400

Thanks ... So I need to set this var before running scrapy crawler it's not possible to set it or change it from the crawler code – no1 Jan 17, 2011 at 11:59

20 You can even set the proxy on a per-request base with: request.meta['proxy'] = 'your.proxy.address' – Pablo Hoffman Jan 25, 2011 at 19:35

3 How do you authenticate the proxy? – Lionel Nov 20, 2011 at 16:59

2 @ephemient How can we tell if `scrapy` is using the proxy? – ocean800 Jun 19, 2017 at 22:58

@ocean800 I use scrapy to scrape a website that shows your current IP to see if it's using the proxy. That way I can load the page via a chrome and see my actual IP and compare it to what scrapy sees on the same page. – Shannon Cole Jun 24, 2018 at 12:53

33

1-Create a new file called “middlewares.py” and save it in your scrapy project and add the following code to it.

```
import base64
class ProxyMiddleware(object):
    # overwrite process request
    def process_request(self, request, spider):
        # Set the location of the proxy
        request.meta['proxy'] = "http://YOUR_PROXY_IP:PORT"

        # Use the following lines if your proxy requires authentication
        proxy_user_pass = "USERNAME:PASSWORD"
        # setup basic authentication for the proxy
        encoded_user_pass = base64.encodestring(proxy_user_pass)
        request.headers['Proxy-Authorization'] = 'Basic ' + encoded_user_pass
```

2 – Open your project’s configuration file (./project_name/settings.py) and add the following code

```
DOWNLOADER_MIDDLEWARES = {
    'scrapy.contrib.downloadermiddleware.httpproxy.HttpProxyMiddleware': 110,
    'project_name.middlewares.ProxyMiddleware': 100,
}
```

Now, your requests should be passed by this proxy. Simple, isn’t it ?

Share Improve this answer Follow

edited Jul 17, 2017 at 18:25
André C. Andersen
9,405 3 58 82

answered Apr 18, 2015 at 10:46
Shahryar Saljoughi
3,069 25 45

I implement your solution which looks correct, but I keep getting a Twisted error: twisted.web._newclient.ResponseNeverReceived: [<twisted.python.failure.Failure <class 'twisted.internet.error.ConnectionDone'>>] ANY ADVICE??? – ccdpowell May 7, 2015 at 1:09

3 Take care to use `base64.b64encode` instead of `base64.encodestring` as the latter adds a newline character to the encoded base64 result...! See stackoverflow.com/a/32243566/426790 – Greg Sadetsky Feb 28, 2016 at 3:03

How can we change proxy after 20 request to not to be banned? – Ekrem Gurdal Jul 6, 2018 at 7:37

2 `scrapy.contrib` is deprecated , it should be just `scrapy` – ishandutta2007 Dec 19, 2019 at 11:12

How do you get `project_name` ? – rom Mar 21, 2021 at 0:05

that would be:

9

```
export http_proxy=http://user:password@proxy:port
```

Share Improve this answer Follow

answered Jan 18, 2013 at 14:58
laurent alsina
171 2 2

I use this yet I just received `[<twisted.python.failure.Failure <class 'twisted.web._newclient.ParseError'>>]` – Allan Ruin Mar 30, 2014 at 15:41

5

Here is what I do

Method 1:

Create a Download Middleware like this

```
class ProxiesDownloaderMiddleware(object):

    def process_request(self, request, spider):

        request.meta['proxy'] = 'user:pass@host:port'
```

and enable that in `settings.py`

```
DOWNLOADER_MIDDLEWARES: {
    'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware': 110,
    'my_scrapy_project_directory.middlewares.ProxiesDownloaderMiddleware': 600,
},
```

That is it, now proxy will be applied to every request

Method 2:

Just enable `HttpProxyMiddleware` in `settings.py` and then do this for each request

```
yield Request(url=..., meta={'proxy': 'user:pass@host:port'})
```

Share Improve this answer Follow

edited Mar 21, 2021 at 8:08

answered Mar 9, 2021 at 7:00
Umair Ayub
21.4k 14 80 154

Where do I find the value for `myproject` ? – rom Mar 21, 2021 at 0:11

@rom its root directory name of your Scrapy project – Umair Ayub Mar 21, 2021 at 8:08

sorry, can you give example? – rom Dec 14, 2021 at 0:59

1 @rom its name of folder where your scrapy project is in, its that simple – Umair Ayub Dec 14, 2021 at 5:09

@UmairAyub, some scrapy projects are done without this template form. In case anyone else is interested you have to create `$ scrapy startproject <your-project-name>` – Gleichmut Sep 22, 2023 at 13:06

4

As I've had trouble by setting the environment in /etc/environment, here is what I've put in my spider (Python):

```
os.environ["http_proxy"] = "http://localhost:12345"
```

Share Improve this answer Follow

answered Nov 18, 2015 at 7:58
user494599

Might as well add `os.environ["https_proxy"]` to it. Worked for me having both. – James Koss May 10, 2019 at 5:08

4

There is nice middleware written by someone [1]: https://github.com/aivarsk/scrapy-proxies "Scrapy proxy middleware"

Share Improve this answer Follow

answered Dec 1, 2015 at 1:58
Niranjan Sagar
829 1 15 17

3

In Windows I put together a couple of previous answers and it worked. I simply did:

```
C:>  set http_proxy = http://username:password@proxy:port
```

and then I launched my program:

```
C:/.../RightFolder> scrapy crawl dmoz
```

where "dmzo" is the program name (I'm writing it because it's the one you find in a tutorial on internet, and if you're here you have probably started from the tutorial).

Share Improve this answer Follow

answered Oct 27, 2015 at 13:20
Andrea Ianni
839 12 24

3

I would recommend you to use a middleware such as scrapy-proxies. You can **rotate proxies, filter bad proxies or use a single proxy** for all your request. Also,using a middleware will save you the trouble of setting up proxy on every run.

This is directly from the GitHub README.

- Install the scrapy-rotating-proxy library

  ```
  pip install scrapy_proxies
  ```

- In your settings.py add the following settings

```
# Retry many times since proxies often fail
RETRY_TIMES = 10
# Retry on most error codes since proxies fail for different reasons
RETRY_HTTP_CODES = [500, 503, 504, 400, 403, 404, 408]

DOWNLOADER_MIDDLEWARES = {
    'scrapy.downloadermiddlewares.retry.RetryMiddleware': 90,
    'scrapy_proxies.RandomProxy': 100,
    'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware': 110,
}

# Proxy list containing entries like
# http://host1:port
# http://username:password@host2:port
# http://host3:port
# ...
PROXY_LIST = '/path/to/proxy/list.txt'

# Proxy mode
# 0 = Every requests have different proxy
# 1 = Take only one proxy from the list and assign it to every requests
# 2 = Put a custom proxy to use in the settings
PROXY_MODE = 0

# If proxy mode is 2 uncomment this sentence :
#CUSTOM_PROXY = "http://host1:port"
```

Here you can change **retry times**, set a **single or rotating proxy**

- Then add your proxy to a list.txt file like this

```
http://host1:port
http://username:password@host2:port
http://host3:port
```

After this all your requests for that project will be sent through proxy. Proxy is rotated for every request randomly. It will not affect concurrency.

Note: if you donot want to use proxy. You can simply comment the scrapy_proxy middleware line.

```
DOWNLOADER_MIDDLEWARES = {
    'scrapy.downloadermiddlewares.retry.RetryMiddleware': 90,
#    'scrapy_proxies.RandomProxy': 100,
    'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware': 110,
}
```

Happy crawling!!!

Share Improve this answer Follow

answered Aug 13, 2019 at 10:26

Amit
1,011 9 24

Start asking to get answers

Find the answer to your question by asking.

Ask question

Explore related questions

python scrapy

See similar questions with these tags.