



# Scrapy gets NoneType Error when using Privoxy Proxy for Tor

Asked 8 years, 8 months ago    Modified 8 years, 8 months ago    Viewed 852 times



I am using Ubuntu 14.04 LTS.

7



I tried Polipo, but it kept refusing Firefox's connections even if I added myself as allowedClient and hours of researching with no solution. So instead, I installed Privoxy and I verified it work with Firefox by going to the Tor website and it said Congrats this browser is configured to use Tor. This confirms that I should be able to scrape Tor websites.



However when I used Scrapy, I get an error that no one seems to have...?



```
2016-07-14 02:43:34 [scrapy] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.httppauth.HttpAuthMiddleware',
'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
'myProject.middlewares.RandomUserAgentMiddleware',
'myProject.middlewares.ProxyMiddleware',
'scrapy.downloadermiddlewares.retry.RetryMiddleware',
'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
'scrapy.downloadermiddlewares.chunked.ChunkedTransferMiddleware',
'scrapy.downloadermiddlewares.stats.DownloaderStats']
2016-07-14 02:43:34 [scrapy] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
'scrapy.spidermiddlewares.referer.RefererMiddleware',
'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
'scrapy.spidermiddlewares.depth.DepthMiddleware']
2016-07-14 02:43:34 [scrapy] INFO: Enabled item pipelines:
['myProject.pipelines.MysqlPipeline']
2016-07-14 02:43:34 [scrapy] INFO: Spider opened
2016-07-14 02:43:34 [scrapy] INFO: Crawled 0 pages (at 0 pages/min), scraped 0
items (at 0 items/min)
2016-07-14 02:43:34 [scrapy] DEBUG: Telnet console listening on 127.0.0.1:6023
2016-07-14 02:43:34 [Tor] DEBUG: User-Agent: Mozilla/5.0 (Macintosh; Intel Mac
OS X 10_7_3) AppleWebKit/534.55.3 (KHTML, like Gecko) Version/5.1.3
Safari/534.53.10 <GET http://thehiddenwiki.org>
2016-07-14 02:43:34 [scrapy] ERROR: Error downloading <GET
http://thehiddenwiki.org>
Traceback (most recent call last):
  File "/usr/local/lib/python2.7/dist-packages/twisted/internet/defer.py", line
1126, in _inlineCallbacks
    result = result.throwExceptionIntoGenerator(g)
  File "/usr/local/lib/python2.7/dist-packages/twisted/python/failure.py", line
389, in throwExceptionIntoGenerator
    return g.throw(self.type, self.value, self.tb)
  File "/usr/local/lib/python2.7/dist-
packages/scrapy/core/downloader/middleware.py", line 43, in process_request
    defer.returnValue((yield download_func(request=request,spider=spider)))
  File "/usr/local/lib/python2.7/dist-packages/scrapy/utils/defer.py", line 45,
in mustbe_deferred
    result = f(*args, **kw)
  File "/usr/local/lib/python2.7/dist-
packages/scrapy/core/downloader/handlers/__init__.py", line 65, in
download_request
    return handler.download_request(request, spider)
  File "/usr/local/lib/python2.7/dist-
packages/scrapy/core/downloader/handlers/http11.py", line 60, in
download_request
    return agent.download_request(request)
  File "/usr/local/lib/python2.7/dist-
packages/scrapy/core/downloader/handlers/http11.py", line 259, in
download_request
    agent = self._get_agent(request, timeout)
  File "/usr/local/lib/python2.7/dist-
packages/scrapy/core/downloader/handlers/http11.py", line 239, in _get_agent
    _, _, proxyHost, proxyPort, proxyParams = _parse(proxy)
  File "/usr/local/lib/python2.7/dist-
packages/scrapy/core/downloader/webclient.py", line 37, in _parse
    return _parsed_url_args(parsed)
  File "/usr/local/lib/python2.7/dist-
packages/scrapy/core/downloader/webclient.py", line 20, in _parsed_url_args
    host = b(parsed.hostname)
  File "/usr/local/lib/python2.7/dist-
packages/scrapy/core/downloader/webclient.py", line 17, in <lambda>
    b = lambda s: to_bytes(s, encoding='ascii')
  File "/usr/local/lib/python2.7/dist-packages/scrapy/utils/python.py", line
117, in to_bytes
    'object, got %s' % type(text).__name__)
TypeError: to_bytes must receive a unicode, str or bytes object, got NoneType
```



I looked up this "to\_byte" error but I go to the source code for Scrapy.

I know this code works without the proxy because it scraped my localhost website and other websites, but not Tor obviously since it's needs the proxy to access onion websites.

What is going on?



Middlewares.py

```
class RandomUserAgentMiddleware(object):
    def process_request(self, request, spider):
        ua = random.choice(settings.get('USER_AGENT_LIST'))
        if ua:
            request.headers.setdefault('User-Agent', ua)
            #this is just to check which user agent is being used for request
            spider.log(
                u'User-Agent: {} {}'.format(request.headers.get('User-Agent'),
            request),
                level=log.DEBUG
            )

class ProxyMiddleware(object):
    def process_request(self, request, spider):
        request.meta['proxy'] = settings.get('HTTP_PROXY')
```

Settings.py

```
USER_AGENT_LIST = [
    'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/535.7 (KHTML, like Gecko)
    Chrome/16.0.912.36 Safari/535.7',
    'Mozilla/5.0 (Windows NT 6.2; Win64; x64; rv:16.0) Gecko/16.0
    Firefox/16.0',
    'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_7_3) AppleWebKit/534.55.3
    (KHTML, like Gecko) Version/5.1.3 Safari/534.53.10'
]

DOWNLOADER_MIDDLEWARES = {
    'myProject.middlewares.RandomUserAgentMiddleware': 400,
    'myProject.middlewares.ProxyMiddleware': 410,
    #'scrapy.contrib.downloadermiddleware.useragent.UserAgentMiddleware': None
    'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware': None
    # Disable compression middleware, so the actual HTML pages are cached
}

HTTP_PROXY = 'localhost:8118'
```

python


proxy

scrapy


polipo

Share Improve this question Follow

asked Jul 14, 2016 at 15:47

 **Arrow**

721 2 7 16


- 1 Use HTTP\_PROXY = 'http://localhost:8118' – paul trmbrth Jul 15, 2016 at 7:19
- 1 And that sir was the answer. Feel fee to post it and I'll upvote you + mark as solution! – Arrow Jul 15, 2016 at 14:31 
- For the record, I created github.com/scrapy/scrapy/issues/2127 – paul trmbrth Jul 15, 2016 at 15:19
- Nice! May this help others along the way. – Arrow Jul 15, 2016 at 15:38
- Apparently this question was answered otherwise, but I got the same error when I was hit by github.com/scrapy/scrapy/pull/1857 -- a nonstandard error code from the server. – Isolated Ostrich Jul 23, 2016 at 21:02

1 Answer


Sorted by:


Highest score (default)


⬆




8









Internally, [Scrapy uses urllib\(2\) 's \\_parse\\_proxy](#) to detect proxy settings. From [urllib docs](#):

The urlopen() function works transparently with proxies which do not require authentication. In a Unix or Windows environment, set the http\_proxy, or ftp\_proxy environment variables to a URL that identifies the proxy server before starting the Python interpreter.

% http\_proxy="http://www.someproxy.com:3128"  
% export http\_proxy  
% python  
...

And when using proxy key in meta , Scrapy expects the same syntax, that is it must contain the scheme, for example 'http://localhost:8118' .

[This is in the docs](#), albeit a bit burried:

You can also set the meta key proxy per-request, to a value like http://some\_proxy\_server:port .


Share

Improve this answer

Follow

edited Jul 15, 2016 at 15:00

answered Jul 15, 2016 at 14:48

 [paul trmbrth](#)

20.7k45567

Start asking to get answers

Find the answer to your question by asking.

Ask question

Explore related questions

- python
- proxy
- scrapy
- polipo

See similar questions with these tags.

