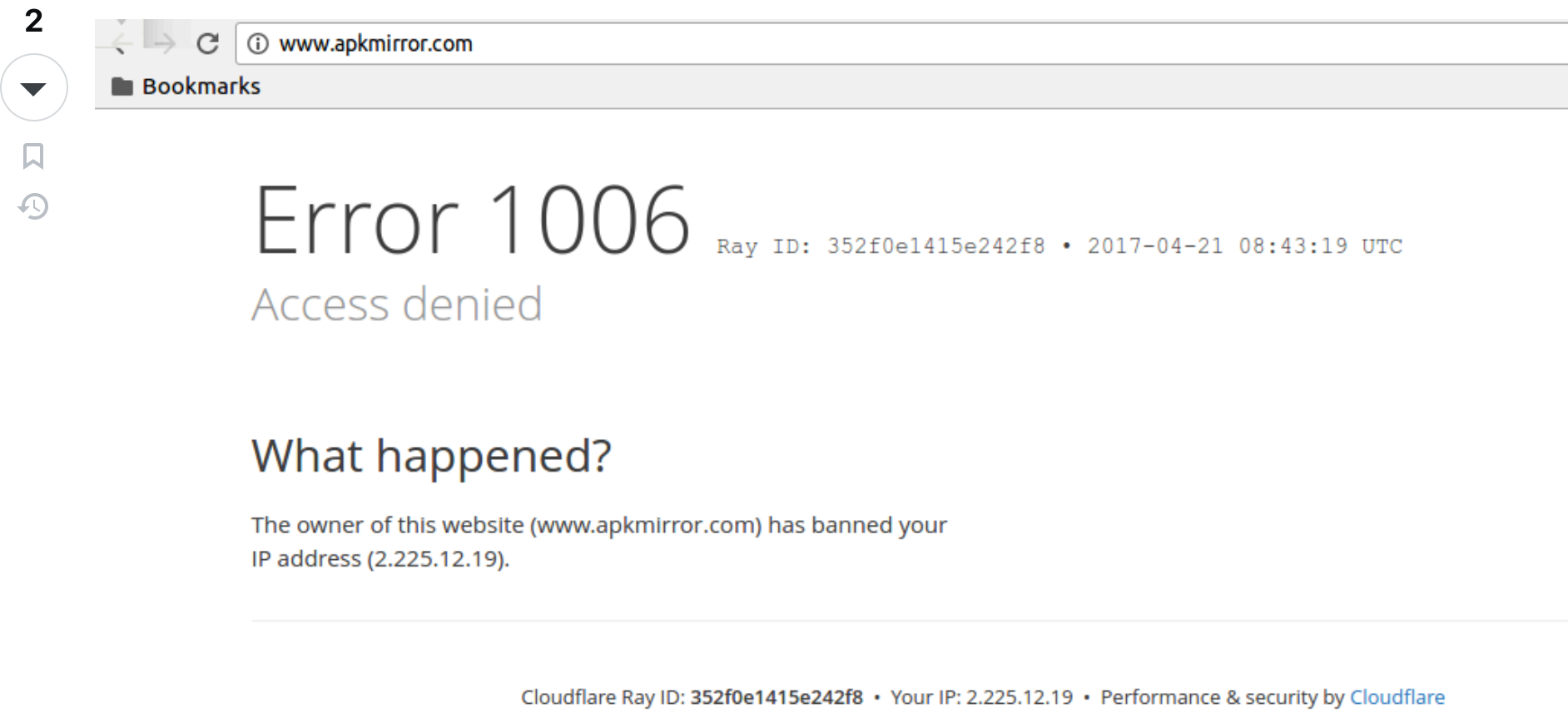


# How to use Privoxy and Tor for a Scrapy project

Asked 7 years, 11 months ago    Modified 7 years, 11 months ago    Viewed 3k times

I'm trying to scrape from <http://www.apkmirror.com>, but currently I'm not able to access the site anymore in my browser because its says the owner banned my IP address (see below).



I'm trying to get around this by using Privoxy and Tor, similar to what is described in <http://blog.michaelyin.info/2014/02/19/scrapy-socket-proxy/>.

Firstly, I installed an started [Privoxy](#), which by default listens at port 8118. I've added the following line to `/etc/privoxy/config` :

```
forward-socks5      /                  127.0.0.1:9050 .
```

I also have Tor running, which is listening at port 9050 as verified using

```
kurt@kurt-ThinkPad:~$ netstat -tulnp | grep 9050
(Not all processes could be identified, non-owned process info
will not be shown, you would have to be root to see it all.)
tcp        0      0 127.0.0.1:9050        0.0.0.0:*            LISTEN
-
```

As far as I can tell using `wget` , it is working. For example, if I `wget apkmirror.com` using a proxy I get a response:

```
kurt@kurt-ThinkPad:~$ wget www.apkmirror.com -e use_proxy=yes -e
http_proxy=127.0.0.1:8118
--2017-04-24 11:02:32--  http://www.apkmirror.com/
Connecting to 127.0.0.1:8118... connected.
Proxy request sent, awaiting response... 200 OK
Length: 185097 (181K) [text/html]
Saving to: 'index.html.2'

index.html.2      100%[=====] 180,76K  --.-KB/s   in 0,004s

2017-04-24 11:02:44 (42,7 MB/s) - 'index.html.2' saved [185097/185097]
```

whereas without the proxy I get `ERROR 403: Forbidden` :

```
kurt@kurt-ThinkPad:~$ wget www.apkmirror.com
--2017-04-24 11:01:24--  http://www.apkmirror.com/
Resolving www.apkmirror.com (www.apkmirror.com)... 104.19.134.58,
104.19.136.58, 104.19.133.58, ...
Connecting to www.apkmirror.com (www.apkmirror.com)|104.19.134.58|:80...
connected.
HTTP request sent, awaiting response... 403 Forbidden
2017-04-24 11:01:24 ERROR 403: Forbidden.
```

Now for the Python code. I've written the following (simplified) spider:

```
import scrapy

DEBUG = True

class TorSpider(scrapy.spiders.SitemapSpider):
```

```
name = "tor-spider"

sitemap_urls = ['https://www.apkmirror.com/sitemap_index.xml']
sitemap_rules = [(r'.*-android-apk-download/$', 'parse')]

if DEBUG:
    custom_settings = {'CLOSESPIDER_PAGECOUNT': 20}

def parse(self, response):
    item = {'url': response.url}
    yield item
```

I've also added the following lines to settings.py :

```
import os
os.environ['http_proxy'] = "http://localhost:8118"

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

from what I understand from <https://doc.scrapy.org/en/latest/topics/downloader-middleware.html#module-scrapy-downloadermiddlewares-httpproxy>, if I set the http\_proxy environment variable the HttpProxyMiddleware should work. However, if I try to scrape using the command

```
scrapy crawl tor-spider -o test.json
```

I get the following response:

```
2017-04-24 10:59:17 [scrapy.utils.log] INFO: Scrapy 1.3.3 started (bot: proxy_spider)
2017-04-24 10:59:17 [scrapy.utils.log] INFO: Overridden settings:
{'NEWSPIDER_MODULE': 'proxy_spider.spiders', 'FEED_URI': 'test.json',
'SPIDER_MODULES': ['proxy_spider.spiders'], 'BOT_NAME': 'proxy_spider',
'ROBOTSTXT_OBEY': True, 'FEED_FORMAT': 'json'}

2017-04-24 10:59:18 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.closespider.CloseSpider',
'scrapy.extensions.feedexport.FeedExporter',
'scrapy.extensions.logstats.LogStats',
'scrapy.extensions.telnet.TelnetConsole',
'scrapy.extensions.corestats.CoreStats']
2017-04-24 10:59:18 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
'scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
'scrapy.downloadermiddlewares.retry.RetryMiddleware',
'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
'scrapy.downloadermiddlewares.stats.DownloaderStats']
2017-04-24 10:59:18 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
'scrapy.spidermiddlewares.referer.RefererMiddleware',
'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
'scrapy.spidermiddlewares.depth.DepthMiddleware']
2017-04-24 10:59:18 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2017-04-24 10:59:18 [scrapy.core.engine] INFO: Spider opened
2017-04-24 10:59:18 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0
pages/min), scraped 0 items (at 0 items/min)
2017-04-24 10:59:18 [scrapy.extensions.telnet] DEBUG: Telnet console listening
on 127.0.0.1:6024
2017-04-24 10:59:18 [scrapy.core.engine] DEBUG: Crawled (403) <GET
https://www.apkmirror.com/robots.txt> (referer: None)
2017-04-24 10:59:18 [scrapy.core.engine] DEBUG: Crawled (403) <GET
https://www.apkmirror.com/sitemap_index.xml> (referer: None)
2017-04-24 10:59:18 [scrapy.spidermiddlewares.httperror] INFO: Ignoring
response <403 https://www.apkmirror.com/sitemap_index.xml>: HTTP status code is
not handled or not allowed
2017-04-24 10:59:18 [scrapy.core.engine] INFO: Closing spider (finished)
2017-04-24 10:59:18 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 519,
'downloader/request_count': 2,
'downloader/request_method_count/GET': 2,
'downloader/response_bytes': 3110,
'downloader/response_count': 2,
'downloader/response_status_count/403': 2,
'finish_reason': 'finished',
'finish_time': datetime.datetime(2017, 4, 24, 8, 59, 18, 927878),
'log_count/DEBUG': 3,
'log_count/INFO': 8,
'response_received_count': 2,
'scheduler/dequeued': 1,
'scheduler/dequeued/memory': 1,
'scheduler/enqueued': 1,
```

```
'scheduler/enqueued/memory': 1,
'start_time': datetime.datetime(2017, 4, 24, 8, 59, 18, 489419)}}
2017-04-24 10:59:18 [scrapy.core.engine] INFO: Spider closed (finished)
```

In short, I'm still getting the 403 error with the scraper despite trying to scrape anonymously using Privoxy/Tor. Am I doing something wrong?


python proxy scrapy privoxy



Share Improve this question Follow

asked Apr 24, 2017 at 9:44

 **Kurt Peek**  
57.8k 102 345 564

1 Answer



Sorted by: Highest score (default) 

  
**3**  



akpmirror is using cloudflare to protect themselves (among other things) against scraping and bots.

Most probably they have scrapy's standard user agent blacklisted. So in addition to using a tor IP (which btw can also be easily blacklisted) you should also set a user agent header that looks like a real browser:

in settings.py

```
USER_AGENT = "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:53.0) Gecko/20100101 Firefox/53.0"
```

 (see <https://doc.scrapy.org/en/latest/topics/settings.html#user-agent> for details)

Share Improve this answer Follow

answered Apr 24, 2017 at 10:33 

 **Done Data Solutions**  
2,286 21 33

Start asking to get answers

Find the answer to your question by asking.

Ask question

Explore related questions

python proxy scrapy privoxy

See similar questions with these tags.