How to use Scrapy with both Splash and Tor over Privoxy in Docker Compose

Asked 7 years, 8 months ago Modified 7 years, 8 months ago Viewed 3k times

6

I'm trying to run a Scrapy spider with two 'extensions':

1. Splash for rendering JavaScript,
2. Tor-Privoxy to provide anonymity.

As an example, I'm using the scraper of `quotes.toscrape.com` in https://github.com/scrapy-plugins/scrapy-splash/tree/master/example. Here is my directory structure:

```
.
├── docker-compose.yml
└── example
    ├── Dockerfile
    ├── scrapy.cfg
    └── scrashtest
        ├── __init__.py
        ├── settings.py
        └── spiders
            ├── __init__.py
            └── quotes.py
```

where the `example` directory is cloned from the `scrapy-splash` repository. I've added the following `docker-compose.yml` file:

```
version: '3'

services:
  scraper:
    build: ./example
    environment:
      - http_proxy=http://tor-privoxy:8118
    links:
      - tor-privoxy
      - splash

  tor-privoxy:
    image: rdsubhas/tor-privoxy-alpine

  splash:
    image: scrapinghub/splash
```

where in the `settings.py` file I've changed the `SPLASH_URL`:

```
# SPLASH_URL = 'http://127.0.0.1:8050/'
SPLASH_URL = 'http://splash:8050'
```

Because Splash is running not on the localhost, but in a separate linked container named `splash`. The `Dockerfile` for the `scraper` is

```
FROM python:alpine
RUN apk --update add libxml2-dev libxslt-dev libffi-dev gcc musl-dev libgcc
openssl-dev curl bash
RUN pip install scrapy scrapy-splash
COPY . /scraper
WORKDIR /scraper
CMD ["scrapy", "crawl", "quotes"]
```

The problem is that when I run this using `docker-compose build` and `docker-compose up`, I get the following logs:

```
Starting examplecompose_tor-privoxy_1
Starting examplecompose_splash_1
Recreating examplecompose_scraper_1
Attaching to examplecompose_splash_1, examplecompose_tor-privoxy_1,
examplecompose_scraper_1
splash_1       | 2017-07-11 16:10:13+0000 [-] Log opened.
splash_1       | 2017-07-11 16:10:13.794595 [-] Splash version: 3.0
tor-privoxy_1  | 2017-07-11 16:10:13.568 7f08e999eee8 Info: Privoxy version
3.0.23
tor-privoxy_1  | 2017-07-11 16:10:13.568 7f08e999eee8 Info: Program name:
privoxy
tor-privoxy_1  | Jul 11 16:10:13.578 [notice] Tor v0.2.6.10 (git-
58c51dc6087b0936) running on Linux with Libevent 2.0.22-stable, OpenSSL 1.0.2d
and Zlib 1.2.8.
tor-privoxy_1  | Jul 11 16:10:13.578 [notice] Tor can't help you if you use it
wrong! Learn how to be safe at
https://www.torproject.org/download/download#warning
splash_1       | 2017-07-11 16:10:13.795925 [-] Qt 5.9.1, PyQt 5.9, WebKit
602.1, sip 4.19.3, Twisted 16.1.1, Lua 5.2
splash_1       | 2017-07-11 16:10:13.796204 [-] Python 3.5.2 (default, Nov 17
2016, 17:05:23) [GCC 5.4.0 20160609]
tor-privoxy_1  | Jul 11 16:10:13.578 [notice] Configuration file
```

```
"/etc/tor/torrc" not present, using reasonable defaults.
tor-privoxy_1  | Jul 11 16:10:13.581 [notice] Opening Socks listener on
127.0.0.1:9050
splash_1       | 2017-07-11 16:10:13.796541 [-] Open files limit: 1048576
tor-privoxy_1  | Jul 11 16:10:13.000 [notice] Parsing GEOIP IPv4 file
/usr/share/tor/geoip.
splash_1       | 2017-07-11 16:10:13.796706 [-] Can't bump open files limit
tor-privoxy_1  | Jul 11 16:10:13.000 [notice] Parsing GEOIP IPv6 file
/usr/share/tor/geoip6.
splash_1       | 2017-07-11 16:10:13.903844 [-] Xvfb is started: ['Xvfb',
':1896918638', '-screen', '0', '1024x768x24', '-nolisten', 'tcp']
splash_1       | QStandardPaths: XDG_RUNTIME_DIR not set, defaulting to
'/tmp/runtime-root'
tor-privoxy_1  | Jul 11 16:10:13.000 [warn] You are running Tor as root. You
don't need to, and you probably shouldn't.
splash_1       | 2017-07-11 16:10:13.984515 [-] proxy profiles support is
enabled, proxy profiles path: /etc/splash/proxy-profiles
tor-privoxy_1  | Jul 11 16:10:13.000 [notice] Bootstrapped 0%: Starting
splash_1       | 2017-07-11 16:10:14.041562 [-] verbosity=1
splash_1       | 2017-07-11 16:10:14.041732 [-] slots=50
tor-privoxy_1  | Jul 11 16:10:13.000 [notice] Bootstrapped 5%: Connecting to
directory server
splash_1       | 2017-07-11 16:10:14.041806 [-] argument_cache_max_entries=500
tor-privoxy_1  | Jul 11 16:10:13.000 [notice] Bootstrapped 80%: Connecting to
the Tor network
splash_1       | 2017-07-11 16:10:14.043083 [-] Web UI: enabled, Lua: enabled
(sandbox: enabled)
splash_1       | 2017-07-11 16:10:14.044088 [-] Site starting on 8050
splash_1       | 2017-07-11 16:10:14.044240 [-] Starting factory
<twisted.web.server.Site object at 0x7f73a4e4b3c8>
tor-privoxy_1  | Jul 11 16:10:14.000 [notice] Bootstrapped 85%: Finishing
handshake with first hop
scraper_1      | 2017-07-11 16:10:15 [scrapy.utils.log] INFO: Scrapy 1.4.0
started (bot: scrashtest)
scraper_1      | 2017-07-11 16:10:15 [scrapy.utils.log] INFO: Overridden
settings: {'BOT_NAME': 'scrashtest', 'DUPEFILTER_CLASS':
'scrapy_splash.SplashAwareDupeFilter', 'HTTPCACHE_STORAGE':
'scrapy_splash.SplashAwareFSCacheStorage', 'NEWSPIDER_MODULE':
'scrashtest.spiders', 'SPIDER_MODULES': ['scrashtest.spiders']}
scraper_1      | 2017-07-11 16:10:15 [scrapy.middleware] INFO: Enabled
extensions:
scraper_1      | ['scrapy.extensions.corestats.CoreStats',
scraper_1      |  'scrapy.extensions.telnet.TelnetConsole',
scraper_1      |  'scrapy.extensions.memusage.MemoryUsage',
scraper_1      |  'scrapy.extensions.logstats.LogStats']
scraper_1      | 2017-07-11 16:10:15 [scrapy.middleware] INFO: Enabled
downloader middlewares:
scraper_1      | ['scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
scraper_1      |
'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
scraper_1      |
'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
scraper_1      |  'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
scraper_1      |  'scrapy.downloadermiddlewares.retry.RetryMiddleware',
scraper_1      |
'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
scraper_1      |  'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
scraper_1      |  'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
scraper_1      |  'scrapy_splash.SplashCookiesMiddleware',
scraper_1      |  'scrapy_splash.SplashMiddleware',
scraper_1      |  'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
scraper_1      |
'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
scraper_1      |  'scrapy.downloadermiddlewares.stats.DownloaderStats']
scraper_1      | 2017-07-11 16:10:15 [scrapy.middleware] INFO: Enabled spider
middlewares:
scraper_1      | ['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
scraper_1      |  'scrapy_splash.SplashDeduplicateArgsMiddleware',
scraper_1      |  'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
scraper_1      |  'scrapy.spidermiddlewares.referer.RefererMiddleware',
scraper_1      |  'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
scraper_1      |  'scrapy.spidermiddlewares.depth.DepthMiddleware']
scraper_1      | 2017-07-11 16:10:15 [scrapy.middleware] INFO: Enabled item
pipelines:
scraper_1      | []
scraper_1      | 2017-07-11 16:10:15 [scrapy.core.engine] INFO: Spider opened
scraper_1      | 2017-07-11 16:10:15 [scrapy.extensions.logstats] INFO: Crawled
0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
scraper_1      | 2017-07-11 16:10:15 [scrapy.extensions.telnet] DEBUG: Telnet
console listening on 127.0.0.1:6023
tor-privoxy_1  | Jul 11 16:10:16.000 [notice] Bootstrapped 90%: Establishing a
Tor circuit
tor-privoxy_1  | Jul 11 16:10:17.000 [notice] Tor has successfully opened a
circuit. Looks like client functionality is working.
tor-privoxy_1  | Jul 11 16:10:17.000 [notice] Bootstrapped 100%: Done
tor-privoxy_1  | Jul 11 16:10:17.000 [warn] Received http status code 404 ("Not
found") from server '216.218.222.10:443' while fetching
"/tor/keys/fp/585769C78764D58426B8B52B6651A5A71137189A+80550987E1D626E3EBA5E5E75A4
scraper_1      | 2017-07-11 16:10:29 [scrapy.core.engine] DEBUG: Crawled (200)
<GET http://quotes.toscrape.com/> (referer: None)
scraper_1      | 2017-07-11 16:10:29 [scrapy.spidermiddlewares.offsite] DEBUG:
Filtered offsite request to 'www.goodreads.com': <GET
https://www.goodreads.com/quotes>
scraper_1      | 2017-07-11 16:10:29 [scrapy.spidermiddlewares.offsite] DEBUG:
Filtered offsite request to 'scrapinghub.com': <GET https://scrapinghub.com>
tor-privoxy_1  | Jul 11 16:10:44.000 [notice] Have tried resolving or
connecting to address '[scrubbed]' at 3 different places. Giving up.
tor-privoxy_1  | Jul 11 16:10:44.000 [notice] Have tried resolving or
connecting to address '[scrubbed]' at 3 different places. Giving up.
```

```
scraper_1      | 2017-07-11 16:10:44 [scrapy.downloadermiddlewares.retry]
DEBUG: Retrying <GET http://quotes.toscrape.com/tag/adulthood/page/1/ via
http://splash:8050/render.json> (failed 1 times): 500 Internal Server Error
scraper_1      | 2017-07-11 16:10:44 [scrapy.downloadermiddlewares.retry]
DEBUG: Retrying <GET http://quotes.toscrape.com/tag/be-yourself/page/1/ via
http://splash:8050/render.json> (failed 1 times): 500 Internal Server Error
tor-privoxy_1  | Jul 11 16:10:55.000 [notice] Have tried resolving or
connecting to address '[scrubbed]' at 3 different places. Giving up.
tor-privoxy_1  | Jul 11 16:10:55.000 [notice] Have tried resolving or
connecting to address '[scrubbed]' at 3 different places. Giving up.
scraper_1      | 2017-07-11 16:10:55 [scrapy.downloadermiddlewares.retry]
DEBUG: Retrying <GET http://quotes.toscrape.com/tag/success/page/1/ via
http://splash:8050/render.json> (failed 1 times): 500 Internal Server Error
scraper_1      | 2017-07-11 16:10:55 [scrapy.downloadermiddlewares.retry]
DEBUG: Retrying <GET http://quotes.toscrape.com/tag/books/page/1/ via
http://splash:8050/render.json> (failed 1 times): 500 Internal Server Error
tor-privoxy_1  | Jul 11 16:10:56.000 [notice] Have tried resolving or
connecting to address '[scrubbed]' at 3 different places. Giving up.
scraper_1      | 2017-07-11 16:10:56 [scrapy.downloadermiddlewares.retry]
DEBUG: Retrying <GET http://quotes.toscrape.com/ via
http://splash:8050/render.json> (failed 1 times): 500 Internal Server Error
tor-privoxy_1  | Jul 11 16:10:57.000 [notice] Have tried resolving or
connecting to address '[scrubbed]' at 3 different places. Giving up.
tor-privoxy_1  | Jul 11 16:10:57.000 [notice] Have tried resolving or
connecting to address '[scrubbed]' at 3 different places. Giving up.
scraper_1      | 2017-07-11 16:10:57 [scrapy.downloadermiddlewares.retry]
DEBUG: Retrying <GET http://quotes.toscrape.com/tag/classic/page/1/ via
http://splash:8050/render.json> (failed 1 times): 500 Internal Server Error
scraper_1      | 2017-07-11 16:10:57 [scrapy.downloadermiddlewares.retry]
DEBUG: Retrying <GET http://quotes.toscrape.com/tag/aliteracy/page/1/ via
http://splash:8050/render.json> (failed 1 times): 500 Internal Server Error
```

where I've interrupted the process for brevity. It seems like the `scraper` and `tor-privoxy` services are alternately complaining about a `500 Internal Service Error` and not being able to 'resolve or connect to address', respectively.

I'm struggling to figure out why the `http_proxy` and Splash don't 'work together'. Can anyone point me in the right direction?

python **docker** **scrapy** **docker-compose** **http-status-code-500**

Share Improve this question Follow

edited Jul 11, 2017 at 16:43

asked Jul 11, 2017 at 16:20

Kurt Peek
57.8k 102 345 564

1 Answer

Sorted by: Highest score (default)

5

Following the Aquarium template project (https://github.com/TeamHG-Memex/aquarium), I found that the trick is to make Splash use Tor, not the spider directly.

My adapted project has the following structure:

```
.
├── docker-compose.yml
├── example
│   ├── Dockerfile
│   ├── scrapy.cfg
│   └── scrashtest
│       ├── __init__.py
│       ├── settings.py
│       └── spiders
│           ├── __init__.py
│           └── quotes.py
└── splash
    └── proxy-profiles
        └── default.ini
```

and the `docker-compose.yml` is

```
version: '3'

services:
  scraper:
    build: ./example
    links:
      - splash

  tor-privoxy:
    image: rdsubhas/tor-privoxy-alpine

  splash:
    image: scrapinghub/splash
    volumes:
      - ./splash/proxy-profiles:/etc/splash/proxy-profiles:ro
```

```
    links:
      - tor-privoxy
```

where I've mounted the `proxy-profiles` directory as a volume into the `splash` container following http://splash.readthedocs.io/en/stable/api.html#proxy-profiles. The `default.ini` reads

```
[proxy]

host=tor-privoxy
port=8118
```

(I also noticed it is essential to call it `default.ini`).

With this setup, upon `docker-compose build` and `docker-compose up` the scraper runs successfully using Splash.

Share Improve this answer Follow

answered Jul 14, 2017 at 14:13

Kurt Peek
57.8k 102 345 564

Thanks for this, it really helped me! I have just started with docker and I don't understand what you mean by "I've mounted the proxy-profiles directory as a volume into the splash container". Do you mount proxy-profile before you run `docker-compose build` , `docker-compose up` ? How do you mount it ? I tried, as indicated in the docs `docker run -p 8050:8050 -v /splash/proxy-profiles:/etc/splash/filters scrapinghub/splash` but this creates another container than the one `docker-compose build` and `docker-compose up` create. On httpbin.org/ip, I can see that the proxy isn't used. – J. Does Dec 19, 2017 at 16:21