Frequently Asked Questions

How does Scrapy compare to BeautifulSoup or lxml?

BeautifulSoup and lxml are libraries for parsing HTML and XML. Scrapy is an application framework for writing web spiders that crawl web sites and extract data from them.

Scrapy provides a built-in mechanism for extracting data (called :ref:`selectors <topics-selectors>`) but you can easily use BeautifulSoup (or lxml) instead, if you feel more comfortable working with them. After all, they're just parsing libraries which can be imported and used from any Python code.

> **System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 15); *backlink***
>
> Unknown interpreted text role "ref".

In other words, comparing BeautifulSoup (or lxml) to Scrapy is like comparing jinja2 to Django.

Can I use Scrapy with BeautifulSoup?

Yes, you can. As mentioned :ref:`above <faq-scrapy-bs-cmp>`, BeautifulSoup can be used for parsing HTML responses in Scrapy callbacks. You just have to feed the response's body into a `BeautifulSoup` object and extract whatever data you need from it.

> **System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 32); *backlink***
>
> Unknown interpreted text role "ref".

Here's an example spider using BeautifulSoup API, with `lxml` as the HTML parser:

```
from bs4 import BeautifulSoup
import scrapy


class ExampleSpider(scrapy.Spider):
    name = "example"
    allowed_domains = ["example.com"]
    start_urls = (
        'http://www.example.com/',
    )

    def parse(self, response):
        # use lxml to get decent HTML parsing speed
        soup = BeautifulSoup(response.text, 'lxml')
        yield {
            "url": response.url,
            "title": soup.h1.string
        }
```

> **Note**
>
> `BeautifulSoup` supports several HTML/XML parsers. See BeautifulSoup's official documentation on which ones are available.

Did Scrapy "steal" X from Django?

Probably, but we don't like that word. We think Django is a great open source project and an example to follow, so we've used it as an inspiration for Scrapy.

We believe that, if something is already done well, there's no need to reinvent it. This concept, besides being one of the foundations for open source and free software, not only applies to software but also to documentation, procedures, policies, etc. So, instead of going through each problem ourselves, we choose to copy ideas from those projects that have already solved them properly, and focus on the real problems we need to solve.

We'd be proud if Scrapy serves as an inspiration for other projects. Feel free to steal from us!

Does Scrapy work with HTTP proxies?

Yes. Support for HTTP proxies is provided (since Scrapy 0.8) through the HTTP Proxy downloader middleware. See :class:`~scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 88); *backlink*

Unknown interpreted text role "class".

How can I scrape an item with attributes in different pages?

See :ref:`topics-request-response-ref-request-callback-arguments`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 95); *backlink*

Unknown interpreted text role "ref".

How can I simulate a user login in my spider?

See :ref:`topics-request-response-ref-request-userlogin`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 100); *backlink*

Unknown interpreted text role "ref".

Does Scrapy crawl in breadth-first or depth-first order?

By default, Scrapy uses a LIFO queue for storing pending requests, which basically means that it crawls in DFO order. This order is more convenient in most cases.

If you do want to crawl in true BFO order, you can do it by setting the following settings:

```
DEPTH_PRIORITY = 1
SCHEDULER_DISK_QUEUE = 'scrapy.squeues.PickleFifoDiskQueue'
SCHEDULER_MEMORY_QUEUE = 'scrapy.squeues.FifoMemoryQueue'
```

While pending requests are below the configured values of :setting:`CONCURRENT_REQUESTS`, :setting:`CONCURRENT_REQUESTS_PER_DOMAIN` or :setting:`CONCURRENT_REQUESTS_PER_IP`, those requests are sent concurrently. As a result, the first few requests of a crawl rarely follow the desired order. Lowering those settings to `1` enforces the desired order, but it significantly slows down the crawl as a whole.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 118); *backlink*

Unknown interpreted text role "setting".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 118); *backlink*

Unknown interpreted text role "setting".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 118); *backlink*

Unknown interpreted text role "setting".

My Scrapy crawler has memory leaks. What can I do?

See :ref:`topics-leaks`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 129); *backlink*

Unknown interpreted text role "ref".

Also, Python has a builtin memory leak issue which is described in :ref:`topics-leaks-without-leaks`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 131); *backlink*

Unknown interpreted text role "ref".

How can I make Scrapy consume less memory?

See previous question.

How can I prevent memory errors due to many allowed domains?

If you have a spider with a long list of :attr:`~scrapy.Spider.allowed_domains` (e.g. 50,000+), consider replacing the default :class:`~scrapy.spidermiddlewares.offsite.OffsiteMiddleware` spider middleware with a :ref:`custom spider middleware <custom-spider-middleware>` that requires less memory. For example:

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 142); *backlink*

Unknown interpreted text role "attr".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 142); *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 142); *backlink*

Unknown interpreted text role "ref".

- If your domain names are similar enough, use your own regular expression instead joining the strings in :attr:`~scrapy.Spider.allowed_domains` into a complex regular expression.

  System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 149); *backlink*

  Unknown interpreted text role "attr".

- If you can meet the installation requirements, use pyre2 instead of Pythonâ€™s re to compile your URL-filtering regular expression. See :issue:`1908`.

  System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 154); *backlink*

  Unknown interpreted text role "issue".

See also other suggestions at StackOverflow.

Note

Remember to disable :class:`scrapy.spidermiddlewares.offsite.OffsiteMiddleware` when you enable your custom implementation:

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 160); *backlink*

Unknown interpreted text role "class".

```
SPIDER_MIDDLEWARES = {
    'scrapy.spidermiddlewares.offsite.OffsiteMiddleware': None,
    'myproject.middlewares.CustomOffsiteMiddleware': 500,
}
```

Can I use Basic HTTP Authentication in my spiders?

Yes, see :class:`~scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 177); *backlink*

Unknown interpreted text role "class".

Why does Scrapy download pages in English instead of my native language?

Try changing the default Accept-Language request header by overriding the :setting:`DEFAULT_REQUEST_HEADERS` setting.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 182); *backlink*

Unknown interpreted text role "setting".

Where can I find some example Scrapy projects?

See :ref:`intro-examples`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 190); *backlink*

Unknown interpreted text role "ref".

Can I run a spider without creating a project?

Yes. You can use the :command:`runspider` command. For example, if you have a spider written in a `my_spider.py` file you can run it with:

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 195); *backlink*

Unknown interpreted text role "command".

```
scrapy runspider my_spider.py
```

See :command:`runspider` command for more info.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 200); *backlink*

Unknown interpreted text role "command".

I get "Filtered offsite request" messages. How can I fix them?

Those messages (logged with `DEBUG` level) don't necessarily mean there is a problem, so you may not need to fix them.

Those messages are thrown by the Offsite Spider Middleware, which is a spider middleware (enabled by default) whose purpose is to filter out requests to domains outside the ones covered by the spider.

For more info see: :class:`~scrapy.spidermiddlewares.offsite.OffsiteMiddleware`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 212); *backlink*

Unknown interpreted text role "class".

What is the recommended way to deploy a Scrapy crawler in production?

See :ref:`topics-deploy`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 218); *backlink*

Unknown interpreted text role "ref".

Can I use JSON for large exports?

It'll depend on how large your output is. See :ref:`this warning <json-with-large-data>` in :class:`~scrapy.exporters.JsonItemExporter` documentation.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 223); *backlink*

Unknown interpreted text role "ref".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 223); *backlink*

Unknown interpreted text role "class".

Can I return (Twisted) deferreds from signal handlers?

Some signals support returning deferreds from their handlers, others don't. See the :ref:`topics-signals-ref` to know which ones.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 230); *backlink*

Unknown interpreted text role "ref".

What does the response status code 999 means?

999 is a custom response status code used by Yahoo sites to throttle requests. Try slowing down the crawling speed by using a download delay of `2` (or higher) in your spider:

```
class MySpider(CrawlSpider):

    name = 'myspider'

    download_delay = 2

    # [ ... rest of the spider code ... ]
```

Or by setting a global download delay in your project with the :setting:`DOWNLOAD_DELAY` setting.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 248); *backlink*

Unknown interpreted text role "setting".

Can I call `pdb.set_trace()` from my spiders to debug them?

Yes, but you can also use the Scrapy shell which allows you to quickly analyze (and even modify) the response being processed by your spider, which is, quite often, more useful than plain old `pdb.set_trace()`.

For more info see :ref:`topics-shell-inspect-response`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 258); *backlink*

Unknown interpreted text role "ref".

Simplest way to dump all my scraped items into a JSON/CSV/XML file?

To dump into a JSON file:

```
scrapy crawl myspider -O items.json
```

To dump into a CSV file:

```
scrapy crawl myspider -O items.csv
```

To dump into a XML file:

```
scrapy crawl myspider -O items.xml
```

For more information see :ref:`topics-feed-exports`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 275); *backlink*

Unknown interpreted text role "ref".

What's this huge cryptic `__VIEWSTATE` parameter used in some forms?

The `__VIEWSTATE` parameter is used in sites built with ASP.NET/VB.NET. For more info on how it works see this page. Also, here's an example spider which scrapes one of these sites.

What's the best way to parse big XML/CSV data feeds?

Parsing big feeds with XPath selectors can be problematic since they need to build the DOM of the entire feed in memory, and this can be quite slow and consume a lot of memory.

In order to avoid parsing all the entire feed at once in memory, you can use the functions `xmliter` and `csviter` from `scrapy.utils.iterators` module. In fact, this is what the feed spiders (see :ref:`topics-spiders`) use under the cover.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 294); *backlink*

Unknown interpreted text role "ref".

Does Scrapy manage cookies automatically?

Yes, Scrapy receives and keeps track of cookies sent by servers, and sends them back on subsequent requests, like any regular web browser does.

For more info see :ref:`topics-request-response` and :ref:`cookies-mw`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 305); *backlink*

Unknown interpreted text role "ref".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 305); *backlink*

Unknown interpreted text role "ref".

How can I see the cookies being sent and received from Scrapy?

Enable the :setting:`COOKIES_DEBUG` setting.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 310); *backlink*

Unknown interpreted text role "setting".

How can I instruct a spider to stop itself?

Raise the :exc:`~scrapy.exceptions.CloseSpider` exception from a callback. For more info see: :exc:`~scrapy.exceptions.CloseSpider`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 315); *backlink*

Unknown interpreted text role "exc".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 315); *backlink*

Unknown interpreted text role "exc".

How can I prevent my Scrapy bot from getting banned?

See :ref:`bans`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 321); *backlink*

Unknown interpreted text role "ref".

Should I use spider arguments or settings to configure my spider?

Both :ref:`spider arguments <spiderargs>` and :ref:`settings <topics-settings>` can be used to configure your spider. There is no strict rule that mandates to use one or the other, but settings are more suited for parameters that, once set, don't change much, while spider arguments are meant to change more often, even on each spider run and sometimes are required for the spider to run at all (for example, to set the start url of a spider).

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 326); *backlink*

Unknown interpreted text role "ref".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 326); *backlink*

Unknown interpreted text role "ref".

To illustrate with an example, assuming you have a spider that needs to log into a site to scrape data, and you only want to scrape data from a certain section of the site (which varies each time). In that case, the credentials to log in would be settings, while the url of the section to scrape would be a spider argument.

I'm scraping a XML document and my XPath selector doesn't return any items

You may need to remove namespaces. See :ref:`removing-namespaces`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 342); *backlink*

Unknown interpreted text role "ref".

How to split an item into multiple items in an item pipeline?

:ref:`Item pipelines <topics-item-pipeline>` cannot yield multiple items per input item. :ref:`Create a spider middleware <custom-spider-middleware>` instead, and use its :meth:`~scrapy.spidermiddlewares.SpiderMiddleware.process_spider_output` method for this purpose. For example:

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 350); *backlink*

Unknown interpreted text role "ref".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 350); *backlink*

Unknown interpreted text role "ref".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 350); *backlink*

Unknown interpreted text role "meth".

```
from copy import deepcopy

from itemadapter import is_item, ItemAdapter

class MultiplyItemsMiddleware:

    def process_spider_output(self, response, result, spider):
        for item in result:
            if is_item(item):
                adapter = ItemAdapter(item)
                for _ in range(adapter['multiply_by']):
                    yield deepcopy(item)
```

Does Scrapy support IPv6 addresses?

Yes, by setting :setting:`DNS_RESOLVER` to `scrapy.resolver.CachingHostnameResolver`. Note that by doing so, you lose the ability to set a specific timeout for DNS requests (the value of the :setting:`DNS_TIMEOUT` setting is ignored).

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 372); *backlink*

Unknown interpreted text role "setting".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 372); *backlink*

Unknown interpreted text role "setting".

How to deal with `<class 'ValueError'>: filedescriptor out of range in select()` exceptions?

This issue has been reported to appear when running broad crawls in macOS, where the default Twisted reactor is :class:`twisted.internet.selectreactor.SelectReactor`. Switching to a different reactor is possible by using the :setting:`TWISTED_REACTOR` setting.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 382); *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 382); *backlink*

Unknown interpreted text role "setting".

How can I cancel the download of a given response?

In some situations, it might be useful to stop the download of a certain response. For instance, sometimes you can determine whether or not you need the full contents of a response by inspecting its headers or the first bytes of its body. In that case, you could save resources by attaching a handler to the :class:`~scrapy.signals.bytes_received` or :class:`~scrapy.signals.headers_received` signals and raising a :exc:`~scrapy.exceptions.StopDownload` exception. Please refer to the :ref:`topics-stop-response-download` topic for additional information and examples.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 392); *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 392); *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 392); *backlink*

Unknown interpreted text role "exc".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 392); *backlink*

Unknown interpreted text role "ref".

Running `runspider` I get `error: No spider found in file: <filename>`

This may happen if your Scrapy project has a spider module with a name that conflicts with the name of one of the Python standard library modules, such as `csv.py` or `os.py`, or any Python package that you have installed. See :issue:`2680`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\(scrapy-master)(docs)faq.rst, line 404); *backlink*

Unknown interpreted text role "issue".