Common Practices

This section documents common practices when using Scrapy. These are things that cover many topics and don't often fall into any other specific section.

Run Scrapy from a script

You can use the :ref:`API <topics-api>` to run Scrapy from a script, instead of the typical way of running Scrapy via `scrapy crawl`.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 15); *backlink*

Unknown interpreted text role "ref".

Remember that Scrapy is built on top of the Twisted asynchronous networking library, so you need to run it inside the Twisted reactor.

The first utility you can use to run your spiders is :class:`scrapy.crawler.CrawlerProcess`. This class will start a Twisted reactor for you, configuring the logging and setting shutdown handlers. This class is the one used by all Scrapy commands.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 21); *backlink*

Unknown interpreted text role "class".

Here's an example showing how to run a single spider with it.

```
import scrapy
from scrapy.crawler import CrawlerProcess

class MySpider(scrapy.Spider):
    # Your spider definition
    ...

process = CrawlerProcess(settings={
    "FEEDS": {
        "items.json": {"format": "json"},
    },
})

process.crawl(MySpider)
process.start() # the script will block here until the crawling is finished
```

Define settings within dictionary in CrawlerProcess. Make sure to check :class:`~scrapy.crawler.CrawlerProcess` documentation to get acquainted with its usage details.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 46); *backlink*

Unknown interpreted text role "class".

If you are inside a Scrapy project there are some additional helpers you can use to import those components within the project. You can automatically import your spiders passing their name to :class:`~scrapy.crawler.CrawlerProcess`, and use `get_project_settings` to get a :class:`~scrapy.settings.Settings` instance with your project settings.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 49); *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 49); *backlink*

Unknown interpreted text role "class".

What follows is a working example of how to do that, using the testspiders project as example.

```
from scrapy.crawler import CrawlerProcess
from scrapy.utils.project import get_project_settings

process = CrawlerProcess(get_project_settings())

# 'followall' is the name of one of the spiders of the project.
process.crawl('followall', domain='scrapy.org')
process.start() # the script will block here until the crawling is finished
```

There's another Scrapy utility that provides more control over the crawling process: :class:`scrapy.crawler.CrawlerRunner`. This class is a thin wrapper that encapsulates some simple helpers to run multiple crawlers, but it won't start or interfere with existing reactors in any way.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 69); *backlink*

Unknown interpreted text role "class".

Using this class the reactor should be explicitly run after scheduling your spiders. It's recommended you use :class:`~scrapy.crawler.CrawlerRunner` instead of :class:`~scrapy.crawler.CrawlerProcess` if your application is already using Twisted and you want to run Scrapy in the same reactor.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 74); *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 74); *backlink*

Unknown interpreted text role "class".

Note that you will also have to shutdown the Twisted reactor yourself after the spider is finished. This can be achieved by adding callbacks to the deferred returned by the :meth:`CrawlerRunner.crawl <scrapy.crawler.CrawlerRunner.crawl>` method.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 79); *backlink*

Unknown interpreted text role "meth".

Here's an example of its usage, along with a callback to manually stop the reactor after `MySpider` has finished running.

```
from twisted.internet import reactor
import scrapy
from scrapy.crawler import CrawlerRunner
from scrapy.utils.log import configure_logging

class MySpider(scrapy.Spider):
    # Your spider definition
    ...

configure_logging({'LOG_FORMAT': '%(levelname)s: %(message)s'})
runner = CrawlerRunner()

d = runner.crawl(MySpider)
d.addBoth(lambda _: reactor.stop())
reactor.run() # the script will block here until the crawling is finished
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 105)

Unknown directive type "seealso".

```
.. seealso:: :doc:`twisted:core/howto/reactor-basics`
```

Running multiple spiders in the same process

By default, Scrapy runs a single spider per process when you run `scrapy crawl`. However, Scrapy supports running multiple spiders per process using the :ref:`internal API <topics-api>`.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 112); *backlink*

Unknown interpreted text role "ref".

Here is an example that runs multiple spiders simultaneously:

```
import scrapy
from scrapy.crawler import CrawlerProcess
from scrapy.utils.project import get_project_settings

class MySpider1(scrapy.Spider):
    # Your first spider definition
    ...

class MySpider2(scrapy.Spider):
    # Your second spider definition
    ...

settings = get_project_settings()
process = CrawlerProcess(settings)
process.crawl(MySpider1)
process.crawl(MySpider2)
process.start() # the script will block here until all crawling jobs are finished
```

Same example using :class:`~scrapy.crawler.CrawlerRunner`:

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 138); *backlink*

Unknown interpreted text role "class".

```
import scrapy
from twisted.internet import reactor
from scrapy.crawler import CrawlerRunner
from scrapy.utils.log import configure_logging
from scrapy.utils.project import get_project_settings

class MySpider1(scrapy.Spider):
    # Your first spider definition
    ...

class MySpider2(scrapy.Spider):
    # Your second spider definition
    ...

configure_logging()
settings = get_project_settings()
runner = CrawlerRunner(settings)
runner.crawl(MySpider1)
runner.crawl(MySpider2)
d = runner.join()
d.addBoth(lambda _: reactor.stop())

reactor.run() # the script will block here until all crawling jobs are finished
```

Same example but running the spiders sequentially by chaining the deferreds:

```
from twisted.internet import reactor, defer
from scrapy.crawler import CrawlerRunner
from scrapy.utils.log import configure_logging
from scrapy.utils.project import get_project_settings

class MySpider1(scrapy.Spider):
    # Your first spider definition
    ...

class MySpider2(scrapy.Spider):
    # Your second spider definition
    ...

configure_logging()
settings = get_project_settings()
runner = CrawlerRunner(settings)

@defer.inlineCallbacks
def crawl():
    yield runner.crawl(MySpider1)
    yield runner.crawl(MySpider2)
    reactor.stop()

crawl()
reactor.run() # the script will block here until the last crawl call is finished
```

Different spiders can set different values for the same setting, but when they run in the same process it may be impossible, by design or because of some limitations, to use these different values. What happens in practice is different for different settings:

- :setting:`SPIDER_LOADER_CLASS` and the ones used by its value (:setting:`SPIDER_MODULES`, :setting:`SPIDER_LOADER_WARN_ONLY` for the default one) cannot be read from the per-spider settings. These are applied when the :class:`~scrapy.crawler.CrawlerRunner` or :class:`~scrapy.crawler.CrawlerProcess` object is created.

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 201); *backlink*

  Unknown interpreted text role "setting".

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 201); *backlink*

  Unknown interpreted text role "setting".

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 201); *backlink*

  Unknown interpreted text role "setting".

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 201); *backlink*

  Unknown interpreted text role "class".

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 201); *backlink*

  Unknown interpreted text role "class".

- For :setting:`TWISTED_REACTOR` and :setting:`ASYNCIO_EVENT_LOOP` the first available value is used, and if a spider requests a different reactor an exception will be raised. These are applied when the reactor is installed.

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 206); *backlink*

  Unknown interpreted text role "setting".

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 206); *backlink*

  Unknown interpreted text role "setting".

- For :setting:`REACTOR_THREADPOOL_MAXSIZE`, :setting:`DNS_RESOLVER` and the ones used by the resolver (:setting:`DNSCACHE_ENABLED`, :setting:`DNSCACHE_SIZE`, :setting:`DNS_TIMEOUT` for ones included in Scrapy) the first available value is used. These are applied when the reactor is started.

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 209); *backlink*

  Unknown interpreted text role "setting".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 209); *backlink*

Unknown interpreted text role "setting".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 209); *backlink*

Unknown interpreted text role "setting".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 209); *backlink*

Unknown interpreted text role "setting".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 209); *backlink*

Unknown interpreted text role "setting".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 215)

Unknown directive type "seealso".

```
.. seealso:: :ref:`run-from-script`.
```

Distributed crawls

Scrapy doesn't provide any built-in facility for running crawls in a distribute (multi-server) manner. However, there are some ways to distribute crawls, which vary depending on how you plan to distribute them.

If you have many spiders, the obvious way to distribute the load is to setup many Scrapyd instances and distribute spider runs among those.

If you instead want to run a single (big) spider through many machines, what you usually do is partition the urls to crawl and send them to each separate spider. Here is a concrete example:

First, you prepare the list of urls to crawl and put them into separate files/urls:

```
http://somedomain.com/urls-to-crawl/spider1/part1.list
http://somedomain.com/urls-to-crawl/spider1/part2.list
http://somedomain.com/urls-to-crawl/spider1/part3.list
```

Then you fire a spider run on 3 different Scrapyd servers. The spider would receive a (spider) argument `part` with the number of the partition to crawl:

```
curl http://scrapy1.mycompany.com:6800/schedule.json -d project=myproject -d spider=spider1 -d part=1
curl http://scrapy2.mycompany.com:6800/schedule.json -d project=myproject -d spider=spider1 -d part=2
curl http://scrapy3.mycompany.com:6800/schedule.json -d project=myproject -d spider=spider1 -d part=3
```

Avoiding getting banned

Some websites implement certain measures to prevent bots from crawling them, with varying degrees of sophistication. Getting around those measures can be difficult and tricky, and may sometimes require special infrastructure. Please consider contacting commercial support if in doubt.

Here are some tips to keep in mind when dealing with these kinds of sites:

- rotate your user agent from a pool of well-known ones from browsers (google around to get a list of them)
- disable cookies (see :setting:`COOKIES_ENABLED`) as some sites may use cookies to spot bot behaviour

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 262); *backlink*

Unknown interpreted text role "setting".

- use download delays (2 or higher). See :setting:`DOWNLOAD_DELAY` setting.

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\(scrapy-master)(docs)(topics)practices.rst`, line 264); *backlink*

  Unknown interpreted text role "setting".

- if possible, use Common Crawl to fetch pages, instead of hitting the sites directly
- use a pool of rotating IPs. For example, the free Tor project or paid services like ProxyMesh. An open source alternative is scrapoxy, a super proxy that you can attach your own proxies to.
- use a highly distributed downloader that circumvents bans internally, so you can just focus on parsing clean pages. One example of such downloaders is Zyte Smart Proxy Manager

If you are still unable to prevent your bot getting banned, consider contacting commercial support.