常见问题

Scrapy与BeautifulSoup或lxml相比如何？

BeautifulSoup和lxml是用于解析HTML和XML的库。Scrapy是一个用于编写Web爬行器的应用程序框架，可以抓取Web站点并从中提取数据。

Scrapy提供了一种用于提取数据的内置机制（称为选择器），但如果您觉得使用它们更舒服，则可以轻松使用BeautifulSoup（或lxml）。毕竟，他们只是解析可以从任何Python代码导入和使用的库。

换句话说，将BeautifulSoup（或lxml）与Scrapy进行比较就像将jinja2与Django进行比较。

我可以和BeautifulSoup一起使用Scrapy吗？

是的你可以。如所提到的上面，BeautifulSoup可用于在Scrapy回调解析HTML响应。您只需将响应的主体提供给`BeautifulSoup`对象并从中提取所需的任何数据。

这是使用BeautifulSoup API的示例蜘蛛，`lxml`作为HTML解析器：

```
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
        }
```

注意

`BeautifulSoup`支持几个HTML / XML解析器。请参阅BeautifulSoup的官方文档，了解哪些可用。

Scrapy支持哪些Python版本？

在CPython（默认Python实现）和PyPy（从PyPy 5.9开始）下，Python 2.7和Python 3.4+支持Scrapy。从Scrapy 0.20开始，Python 2.6支持被删除。Scrapy 1.1中添加了Python 3支持。在Scrapy 1.4中添加了PyPy支持，在Scrapy 1.5中添加了PyPy3支持。

> **注意**
>
> 对于Windows上的Python 3支持，建议使用安装指南中概述的 Anaconda / Miniconda 。

Scrapy是否从Django“窃取”X？

可能，但我们不喜欢这个词。我们认为Django是一个很好的开源项目，也是一个值得关注的例子，因此我们将它作为Scrapy的灵感来使用。

我们相信，如果事情已经做好，就没有必要重新发明它。这个概念除了作为开源和自由软件的基础之外，不仅适用于软件，还适用于文档，程序，政策等。因此，我们不是自己解决每个问题，而是选择从这些项目中复制想法。已经妥善解决了这些问题，并专注于我们需要解决的实际问题。

如果Scrapy是其他项目的灵感，我们会感到自豪。随意偷我们！

Scrapy是否可以与HTTP代理一起使用？

是。通过HTTP代理下载器中间件提供对HTTP代理的支持（自Scrapy 0.8起）。见 `HttpProxyMiddleware` 。

如何在不同页面中使用属性刮取项目？

请参阅将其他数据传递给回调函数。

Scrapy崩溃：ImportError：没有名为win32api的模块

由于这个Twisted错误，你需要安装pywin32。

如何在蜘蛛中模拟用户登录？

请参阅使用FormRequest.from_response（）来模拟用户登录。

Scrapy是以广度优先还是深度优先的顺序爬行？

默认情况下，Scrapy使用LIFO队列来存储挂起的请求，这基本上意味着它以DFO顺序进行爬网。在大多数情况下，此订单更方便。如果您确实想要以真正的BFO顺序进行爬网，可以通过设置以下设置来执行此操作：

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

我的Scrapy爬虫有内存泄漏。我能做什么？

请参阅调试内存泄漏。

此外，Python有内置内存泄漏问题，泄漏中描述 没有泄漏。

如何让Scrapy消耗更少的内存？

见上一个问题。

我可以在蜘蛛中使用基本HTTP身份验证吗？

是的，看`HttpAuthMiddleware`。

为什么Scrapy用英语而不是我的母语下载页面？

尝试通过覆盖设置来更改默认的Accept-Language请求标头`DEFAULT_REQUEST_HEADERS`。

我在哪里可以找到一些示例Scrapy项目？

见例子。

我可以在没有创建项目的情况下运行蜘蛛吗？

是。您可以使用该`runspider`命令。例如，如果您在`my_spider.py`文件中编写了一个蜘蛛，则可以使用以下命令运行它：

```
scrapy runspider my_spider.py
```

有关`runspider`详细信息，请参阅命令

我收到“Filtered offsite request”消息。我该如何解决它们？

这些消息（以`DEBUG`级别记录）并不一定意味着存在问题，因此您可能不需要修复它们。

这些消息由Offsite Spider Middleware抛出，它是一个蜘蛛中间件（默认启用），其目的是过滤掉蜘蛛所覆盖的域之外的域的请求。

有关更多信息，请参阅：`OffsiteMiddleware`。

在生产中部署Scrapy搜寻器的推荐方法是什么？

请参阅部署Spider。

我可以将JSON用于大型出口吗？

这取决于你的输出有多大。请参阅此警告的`JsonItemExporter`文档。

我可以从信号处理程序返回（扭曲）延迟吗？

有些信号支持从处理程序返回延迟，其他信号则不支持。请参阅内置信号参考以了解哪些参考。

响应状态代码999的含义是什么？

999是雅虎网站用来限制请求的自定义响应状态代码。尝试使用`2`蜘蛛中的下载延迟（或更高）来降低爬行速度：

```
class MySpider(CrawlSpider):

    name = 'myspider'

    download_delay = 2

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

或者通过设置在项目中设置全局下载延迟`DOWNLOAD_DELAY`。

我可以打电话`pdb.set_trace()`给我的蜘蛛进行调试吗？

是的，但您也可以使用Scrapy shell，它允许您快速分析（甚至修改）蜘蛛处理的响应，这通常比普通的更有用`pdb.set_trace()`。

有关更多信息，请参阅从spiders调用shell以检查响应。

将所有已删除项目转储到JSON / CSV / XML文件的最简单方法是什么？

要转储到JSON文件中：

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

要转储到CSV文件：

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

要转储到XML文件中：

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

有关更多信息，请参阅Feed导出

`__VIEWSTATE` 在某些形式中使用的这个巨大的神秘参数是什么？

该 `__VIEWSTATE` 参数用于使用ASP.NET / VB.NET构建的站点。有关其工作原理的详细信息，请参阅此页面。此外，这是一个蜘蛛的示例 ，它刮擦其中一个站点。

解析大型XML / CSV数据源的最佳方法是什么？

使用XPath选择器解析大型提要可能会有问题，因为它们需要在内存中构建整个提要的DOM，这可能非常慢并且占用大量内存。

为了避免在内存中一次解析所有的全部饲料原料，可以使用的功能 `xmliter` 和 `csviter` 从 `scrapy.utils.iterators` 模块。事实上，这就是饲料蜘蛛（见蜘蛛）在封面下使用的东西。

Scrapy是否自动管理cookie？

是的，Scrapy接收并跟踪服务器发送的cookie，并将其发送回后续请求，就像任何常规Web浏览器一样。

有关更多信息，请参阅请求和响应以及CookiesMiddleware。

如何查看Scrapy发送和接收的cookie？

启用 `COOKIES_DEBUG` 设置。

如何指示蜘蛛自行停止？

`CloseSpider`从回调中提出异常。有关更多信息，请参阅：`CloseSpider`。

如何防止Scrapy bot被禁止？

请参阅避免被禁止。

我应该使用蜘蛛参数或设置来配置我的蜘蛛吗？

这两种蜘蛛的参数和设置， 可以用于配置您的蜘蛛。没有严格的规则要求使用其中一个，但设置更适合参数，一旦设置，变化不大，而蜘蛛参数意味着更频繁地更改，即使在每次蜘蛛运行时，有时候蜘蛛根本需要运行（例如，设置蜘蛛的起始URL）。

举一个例子来说明，假设您有一个需要登录站点来抓取数据的蜘蛛，并且您只想从站点的某个部分（每次都有所不同）中抓取数据。在这种情况下，登录的凭据将是设置，而要刮取的部分的URL将是蜘蛛参数。

我正在抓取XML文档，我的XPath选择器不返回任何项目

您可能需要删除命名空间。请参阅删除命名空间。