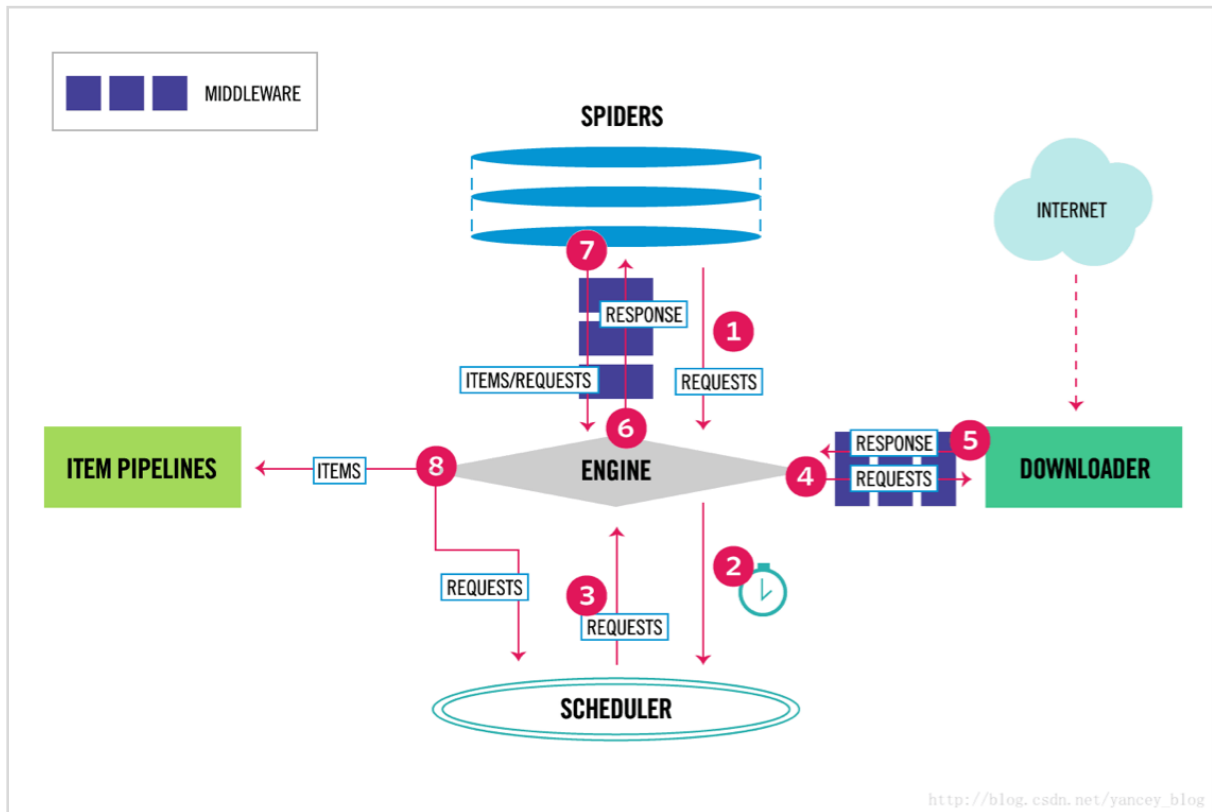


#爬虫/day8 Scrapy 爬虫框架#

特点：爬取效率高、扩展性强、Python编写跨平台运行

数据流程



Scrapy中的数据流由执行引擎控制，其过程如下：

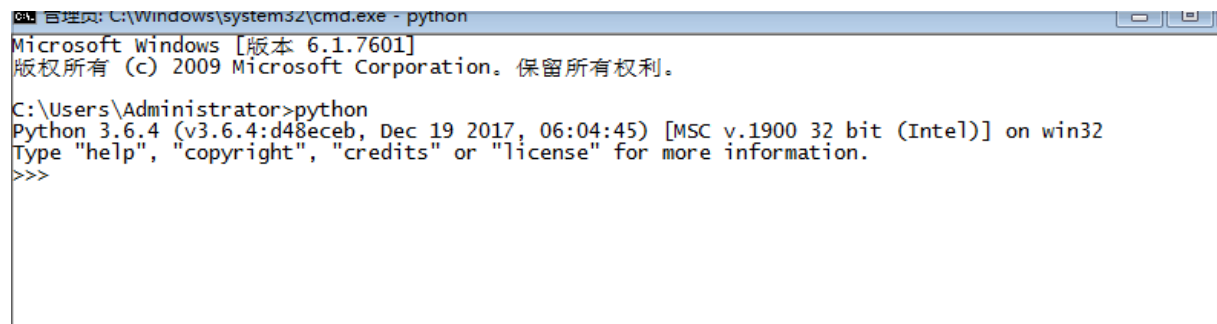
- 1、引擎打开一个网站(open a domain)，找到处理该网站的Spider并向该spider请求第一个要爬取的URL(s)。
- 2、引擎从Spider中获取到第一个要爬取的URL并在调度器(Scheduler)以Request调度。
- 3、引擎向调度器请求下一个要爬取的URL。
- 4、调度器返回下一个要爬取的URL给引擎，引擎将URL通过下载中间件(请求(request)方向)转发给下载器(Downloader)。
- 5、一旦页面下载完毕，下载器生成一个该页面的Response，并将其通过下载中间件(返回(response)方向)发送给引擎。
- 6、引擎从下载器中接收到Response并通过Spider中间件(输入方向)发送给Spider处理。
- 7、Spider处理Response并返回爬取到的Item及(跟进的)新的Request给引擎。
- 8、引擎将(Spider返回的)爬取到的Item给Item Pipeline，将(Spider返回的)Request给调度器。
- 9、(从第二步)重复直到调度器中没有更多地request，引擎关闭该网站。

安装Scrapy

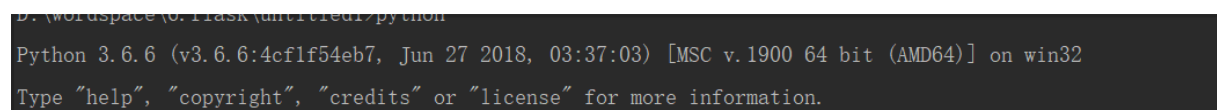
```
windows whl  
https://www.lfd.uci.edu/~gohlke/pythonlibs/  
  
pip install scrapy -i https://pypi.douban.com/simple
```

windows 安装

1.scrapy依赖twiste, 进入 <http://www.lfd.uci.edu/~gohlke/pythonlibs/> , 找到适合的版本, 下载Twisted-18.9.0-cp37-cp37m-win_amd64.whl



```
管理员: C:\Windows\system32\cmd.exe - python  
Microsoft Windows [版本 6.1.7601]  
版权所有 (c) 2009 Microsoft Corporation。保留所有权利。  
  
C:\Users\Administrator>python  
Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit (Intel)] on win32  
Type "help", "copyright", "credits" or "license" for more information.  
>>>
```



```
D:\workspace\0. Flask (untitled1)>python  
Python 3.6.6 (v3.6.6:4cflf54eb7, Jun 27 2018, 03:37:03) [MSC v.1900 64 bit (AMD64)] on win32  
Type "help", "copyright", "credits" or "license" for more information.
```

2.在下载存放的目录下安装

```
pip install Twisted-18.9.0-cp37-cp37m-win_amd64.whl
```

如遇到需要下载 pywin32, 请下载安装

[Python for Windows Extensions - Browse /pywin32/Build 221 at SourceForge.net](#)

3. pip install pypiwin32

4. pip install scrapy

```
### 创建项目
```

scrapy startproject myscrapy

```
防止被爬网站的robots.txt起作用
settings.py
修改 ROBOTSTXT_OBEY = False

### 创建Spider
```

scrapy genspider football sports.sina.com.cn

Spider 属性

name: Spider名字

allowed_domains: 允许爬取的域名

start_urls: Spider启动时爬取的url列表

parse: 负责解析返回的响应，提取数据或进一步处理

创建Item

Item是保存爬取数据的容器

解析Response

使用Item

后续Request

运行

scrapy crawl football

保存到文件 (csv, xml, pickle, marshal)

scrapy crawl football -o football.json

FEED_EXPORT_ENCODING = 'utf-8'