Python爬虫框架——Scrapy

2020年4月26日, 星期日　16:10

Scrapy入门

1.Scrapy是框架

2.采用异步框架，实现高效率的网络采集

3.最强大的爬虫框架

安装Scrapy

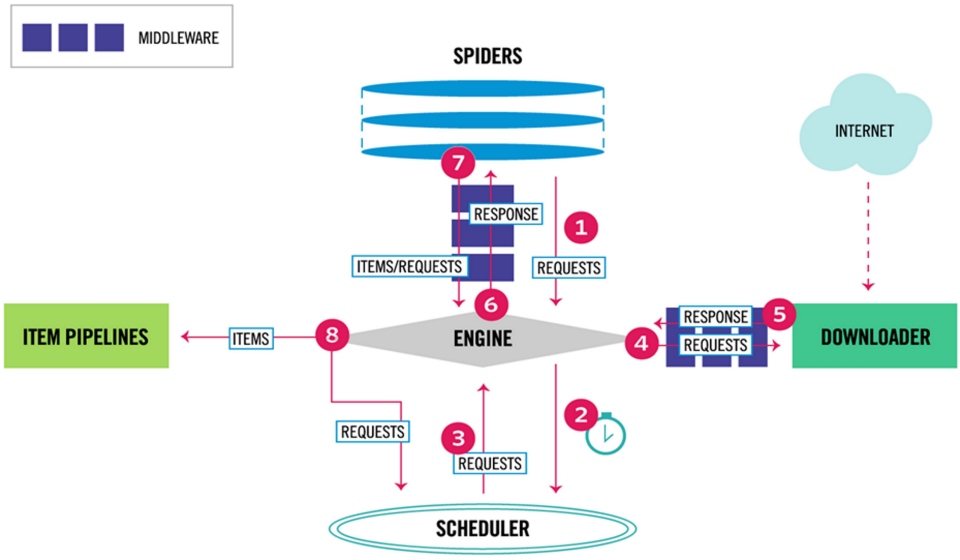
1.安装命令：pip install scrapy

2.成功标志：输入命令scrapy bench运行不会报错

3.常见问题：

- pip install scrapy -> VC++ 14.0 Twisted
  解决方法：离线安装pip install xxx.whl
- scrapy bench运行报错
  解决方法：pip install pywin32

Scrapy原理



示例程序

采集目标：西刺网的IP代理（包括IP和端口）

Step1 新建项目

在cmd中输入命令scrapy startproject xicidailiSpider

Step2 创建爬虫

1.在cmd中输入命令scrapy genspider xicidaili xicidaili.com

2.注意：

- 爬虫名称不要和项目名称一样
- 网站域名是允许爬虫采集的域名

3.解释爬虫文件

- 导入scrapy
- 创建爬虫类，并且继承自scrapy.Spider
  - 爬虫名称（必须唯一）
  - 允许采集的域名
  - 开始采集的网站
  - parse方法：解析响应数据，提取数据或者网址等

Step3 分析页面

1.提取数据

- 正则表达式：基础，必会，难掌握
- XPath：从HTML中提取数据
- CSS：从HTML中提取数据

2.yield scrapy.Request(next_url, callback=self.parse)

- Request：发起请求
- callback：回调函数，将请求得到的响应交给自己处理

Step4 运行爬虫

1.运行：在cmd中输入命令scrapy crawl xicidaili

2.执行程序并输出保存数据

- 在cmd中输入命令scrapy crawl xicidaili -o ip.json，生成ip.json文件
- 在cmd中输入命令scrapy crawl xicidaili -o ip.csv，生成ip.csv文件