第三课 网络数据的获取与表示
第六节 Scrapy简介
主　讲:　Robin

Scrapy简介

- 开源的Python爬虫框架，用于抓取web站点并从页面中提取结构化的数据
- 用途广泛，可用于数据挖掘、监测和自动化测试
- 快速强大，只需编写少量代码即可完成爬取任务
- 易扩展，可添加新的功能模块
- 用户 https://scrapy.org/companies/

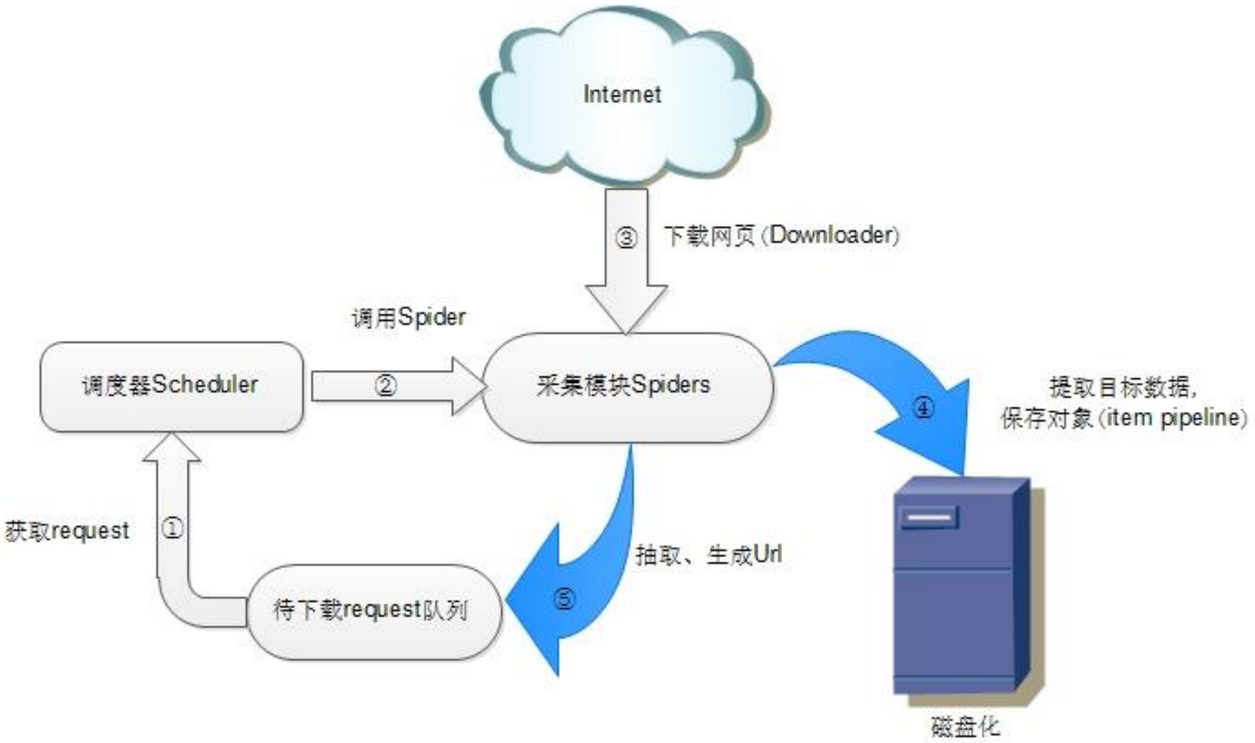
Scrapy

An open source and collaborative framework
for extracting the data you need from websites.
In a fast, simple, yet extensible way.

Scrapy高级特性

- 内置数据抽取器CSS/XPath/re
- 交互式控制台用于调试
- 结果输出的格式支持，JSON, CSV, XML等
- 自动处理编码
- 支持自定义扩展

Scrapy架构



1. 调度器从待下载URL中取出一个URL
2. 调度器启动采集模块Spiders模块
3. 采集模块把URL传给下载器，下载器把资源下载下来自动处理编码
4. 提取目标数据，抽取出目标对象，由管道进行进一步的处理；比如存入数据库、文本
5. 若是解析出的是URL，则把URL插入到待爬取队列当中

只为遇见明天更优秀的你！