scrapy原理

Ringo Yungpo Kao

December 19, 2017

# Python网络爬虫笔记

## Scrapy简介

Scrapy是一个为了爬取网站数据，提取结构性数据而编写的应用框架。 可以应用在包括数据挖掘，信息处理或存储历史数据等一系列的程序中。 其最初是为了 页面抓取 (更确切来说, 网络抓取 )所设计的， 也可以应用在获取API所返回的数据(例如 Amazon Associates Web Services ) 或者通用的网络爬虫。

## 各组件作用

### Scrapy Engine

引擎负责控制数据流在系统中所有组件中流动，并在相应动作发生时触发事件。 详细内容查看下面的数据流(Data Flow)部分。 此组件相当于爬虫的“大脑”，是整个爬虫的调度中心。

### 调度器(Scheduler)

调度器从引擎接受request并将他们入队，以便之后引擎请求他们时提供给引擎。

初始的爬取URL和后续在页面中获取的待爬取的URL将放入调度器中，等待爬取。同时调度器会自动去除重复的URL（如果特定的URL不需要去重也可以通过设置实现，如post请求的URL）

### 下载器(Downloader)

下载器负责获取页面数据并提供给引擎，而后提供给spider。

### Spiders

Spider是Scrapy用户编写用于分析response并提取item(即获取到的item)或额外跟进的URL的类。 每个spider负责处理一个特定(或一些)网站。

### Item Pipeline

Item Pipeline负责处理被spider提取出来的item。典型的处理有清理、 验证及持久化(例如存取到数据库中)。

当页面被爬虫解析所需的数据存入Item后，将被发送到项目管道(Pipeline)，并经过几个特定的次序处理数据，最后存入本地文件或存入数据库。

### 下载器中间件(Downloader middlewares)

下载器中间件是在引擎及下载器之间的特定钩子(specific hook)，处理Downloader传递给引擎的response。 其提供了一个简便的机制，通过插入自定义代码来扩展Scrapy功能。

通过设置下载器中间件可以实现爬虫自动更换user-agent、IP等功能。

### Spider中间件(Spider middlewares)

Spider中间件是在引擎及Spider之间的特定钩子(specific hook)，处理spider的输入(response)和输出(items及requests)。 其提供了一个简便的机制，通过插入自定义代码来扩展Scrapy功能。

### 数据流(Data flow)

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。

## 建立Scrapy爬虫项目流程

### 创建项目

在开始爬取之前，首先要创建一个新的Scrapy项目。这里以爬取我的博客为例，进入你打算存储代码的目录中，运行下列命令:

scrapy startproject scrapyspider

该命令将会创建包含下列内容的scrapyspider目录:

scrapyspider/  
 scrapy.cfg  
 scrapyspider/  
 \_\_init\_\_.py  
 items.py  
 pipelines.py  
 settings.py  
 spiders/  
 \_\_init\_\_.py  
 ...

这些文件分别是: - scrapy.cfg: 项目的配置文件。 - scrapyspider/: 该项目的python模块。之后您将在此加入代码。 - scrapyspider/items.py: 项目中的item文件。 - scrapyspider/pipelines.py: 项目中的pipelines文件。 - scrapyspider/settings.py: 项目的设置文件。 - scrapyspider/spiders/: 放置spider代码的目录。