基于 Python 的网络爬虫技术

钱 程 阳小兰 朱福喜
(武昌理工学院信息工程学院,湖北 武汉 430223)

摘 要:基于 Python 的网络爬虫可以方便地抓取网页信息,以豆瓣网站为例,实现了基于 Python 网络爬虫抓取豆瓣影视信息的过程。

关键词:Python;网络爬虫;源代码

1 基于 Python 的网络爬虫

网络爬虫又称网络蜘蛛,或网络机器人。网络爬虫通过网页的链接地址来查找网页内容,并直接返回给用户所需要的数据,不需要人工操纵浏览器获取。Python 是一个广泛使用的脚本语言,其自带了 urllib、urllib2 等爬虫最基本的库,Scrapy 网络爬虫是基于 Python 语言开发的开源爬虫软件,Scrapy 可在 Windows,Linux 等多个操作系统运行。如果待抓取网页的 HTML 源码很多,需要下载大量的内容,用户可在 Scrapy 爬虫框架上定制开发部分模块实现爬虫功能。

2 基于 Python 的网络爬虫实现

本文以豆瓣网站为例,探讨基于 Python 的网络爬虫实现过程。爬虫首先获得所有待抓取影视的链接地址,然后依次根据链接地址解析网页上对应影视的详细信息,保存到文件中。具体实现过程如下:

2.1 获得待爬取影视 URL

为了获得待抓取的所有影视 URL,必须找到一个“入口”URL,该“入口”URL 对应的网页上陈列着多个影视作品的简单信息。对于豆瓣网站,入口网址为:https://www.douban.com/tag/****/movie?start=**** ,该网页上陈列 15 部影视作品,其中的第一个“****”位置需要填入的是待抓取影视最早上映年份,第二个“****”位置需要填入的是该网页第一个影视作品的序号(可以是 0,15,30,…,表示从 0 开始每次递增 15)。通过改变序号实现“下一页”的功能,例如 start=0 表示的第一页,start=15 表示的是第二页,依次下去,每次递增 15。即:(i- 1)*15(i 表示当前页码,i=1,2,…)。所以只需要改变入口网址的两处位置的取值就能抓取不同年份的多个页面上的多部影视作品。例如,网址:https://www.douban.com/tag/2016/movie?start=0 页面上展示的是豆瓣网站 2016 年标签下第一页的 15 部影视,start=15 对应是第二页的 15 部影视,以此继续。

由以上分析可知,已经获得了链接地址的规律。为了得到影视作品的名称和 URL,可以直接使用 Python 自带的 urllib2 库和正则表达式解决问题。可以编写抓取更多信息的爬虫,并将返回的字符串类型变量的内容保存到不同的本地文件中。最终可以查看保存在本地磁盘上的一个文件中某个影视作品的主要 HTML 源代码。

实现“获得爬取影视 URL”爬虫的两个主要类为 doubanSpider 和 spiderUtil。实现“获得爬取影视 URL”功能的 doubanSpider 类的具体调用如下:在 main 函数内新建一个 doubanSpider 对象,通过这个对象调用 getContents 函数获得所有影视列表所在 URL 的源代码,利用 spiderUtil 类将这些文件保存,此时已经将影视列表所在 URL 的源代码信息按照年份分别存入到本地磁盘不同文件夹下的不同文件中。然后通过调用 readAll 函数遍历所有文件夹下的文件,在 readAll 函数中调用了 parseWeb 函数,用于解析每个源代码文件中的影视信息,以及调用了 spiderUtil 类中的 save 方法,用于保存影视信息。

2.2 解析影视详细信息

豆瓣网站上影视作品详情页面,包括影视作品的上下文信息以及豆瓣评分,可以使用 Scrapy 爬虫框架获得每个影视作品详情页面的重要上下文信息。下面介绍利用 Scrapy 爬虫框架获得豆瓣网站上影视作品的详细信息。

项目来源:湖北省教育厅科研计划项目(B2015281)。
作者简介:钱程(1980-),男,硕士,副教授,湖北公安人。

2.2.1 修改 items.py 文件

items.py 文件对应的类名为 TutorialItem,首先在该类中需要引入 scrapy.item 下的 Item 和 Field 类,然后根据需要定义影视作品的属性。

2.2.2 修改 pipelines.py 文件

pipelines.py 文件是管道处理文件,对应的类名为 TutorialPipeline。需要修改的地方是重写 process_item 方法。该方法的功能是处理 item(包括数据格式转换和其他复杂操作),可以通过这个方法保存 item。编写完 pipelines.py 的主要代码之后,需要在 settings.py 中配置 ITEM_PIPELINES,表示从 movieSpider.py 类中 parse()方法返回的 item 数据将依次被 TutorialPipeline 类处理,并对得到的 item 进行保存等其他操作。

2.2.3 编写 movieSpider.py 文件

spiders 文件夹下 movieSpider.py 文件是爬虫主体,开发人员的主要代码就是写在这个文件中,其主要功能是对下载下来的源代码进行解析,生成自定义的 item 对象或再次发起请求。该类的三个必要的元素为:name 属性、start_urls 列表、parse()方法。下面从定制爬虫流程的角度,依次说明这三个必要元素的作用。

首先需要给出该爬虫的唯一标识名 name 的取值,例如在 movieSpider 类中,赋值 name 属性为 doubanmoive,如果要运行该爬虫,可以直接用 doubanmovie 来标识。然后通过在 start_requests 方法中对 start_urls 属性进行赋值,并将 start_urls 中的 url 通过 make_requests_from_url 方法依次加入到待爬取的 URL 队列中等待被取出。最后重写 parse 方法,通过 HtmlXPathSelector 对象的选择器(Selector)定制需要的信息。在 parse 方法中同时可以将爬取到的信息存入 TutorialItem 对象中,这样就可以经由 TutorialPipeline 管道持久化到文件系统或者数据库。在 movieSpider 类中 parse 方法一定要返回当前的 item 对象,否则交给 Pipline 处理的对象只有最后一个。

2.2.4 修改 settings.py 配置文件

大部分的网站都会有“防爬取”机制。需要设置一些属性,来防止网站轻易封 ip。例如可以设置 settings.py 文件中的 DOWNLOAD_DELAY 属性。

修改上述四个文件之后,接下来打开命令提示符窗口,将当前路径切换到爬虫所在的路径,并执行 scrapy crawl doubanmovie 命令,这样爬虫就开始爬取网页的动作,最后保存影视数据到本地磁盘上。

结束语

利用 Python 自带的库可以得到网页的内容,并通过正则表达式提取分析所需要的信息,利用基于 Python 的开源爬虫软件 Scrapy 可以有效实现对 Web 信息的抓取。

参考文献

[1]刘娜.Python 正则表达式高级特性研究[J].电脑编程技巧与维护,2015,22:12-13.
[2]于娟,刘强.主题网络爬虫研究综述[J].计算机工程与科学,2015,02:231-237.