**最简单的python爬虫制作**

看完这篇教程你可以完成一个爬取豆瓣电影top250网页的爬虫，其实比葫芦画瓢，你已经可以定向爬取很多静态网页了。

我们来看一下什么是网络爬虫：

**网络爬虫**（英语：web crawler），也叫网络蜘蛛（spider），是一种用来自动浏览万维网的网络机器人。 ---维基百科

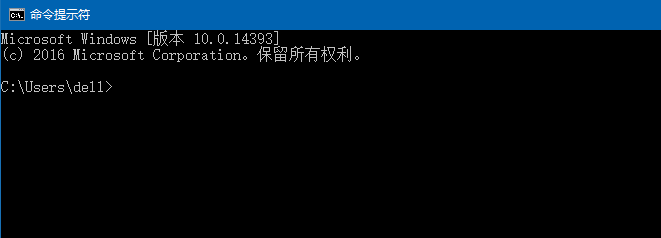
听起来很厉害，网络机器人，其实一个最最最简单的爬虫只需要四行python代码，不信请往下看。

1. **最最最简单的爬虫**

环境要求：python2.7，requests库

1、这个爬虫需要用的python的一个第三方库 requests,通过pip安装。

首先，打开命令提示符：（如何打开，每个windows版本略有不同，请自行百度）



输入 pip install requests 就会自动安装requests

2、最最最简单的爬虫

先看代码：



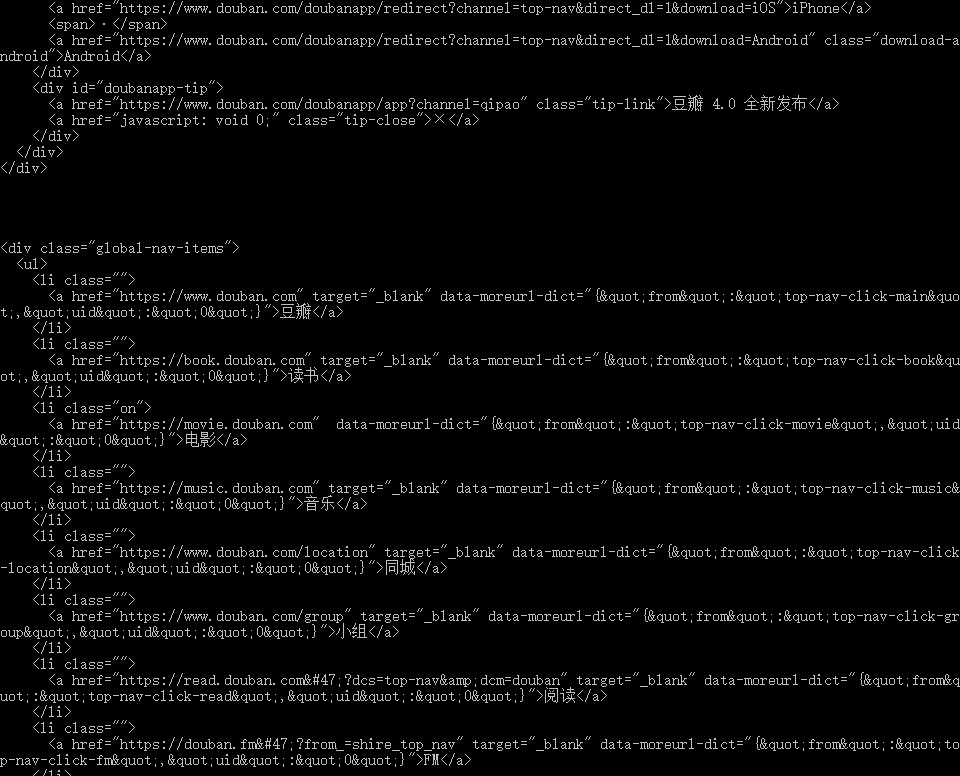
第1行，首先导入requests库

第3行，豆瓣电影top250的链接放入url

第4行，使用requests库的get方法获得目标网页的源代码

第5行，打印出来结果

如果不太清楚具体什么意思，没关系，先把这几行代码运行起来看看会得到什么结果吧。



哇，打印出来了好多东西，一段长长的文本，看起来里面的文字好像就是这个网页里的内容。

我们先用人脑来浏览一下这个网页，在浏览器打开这个网址：

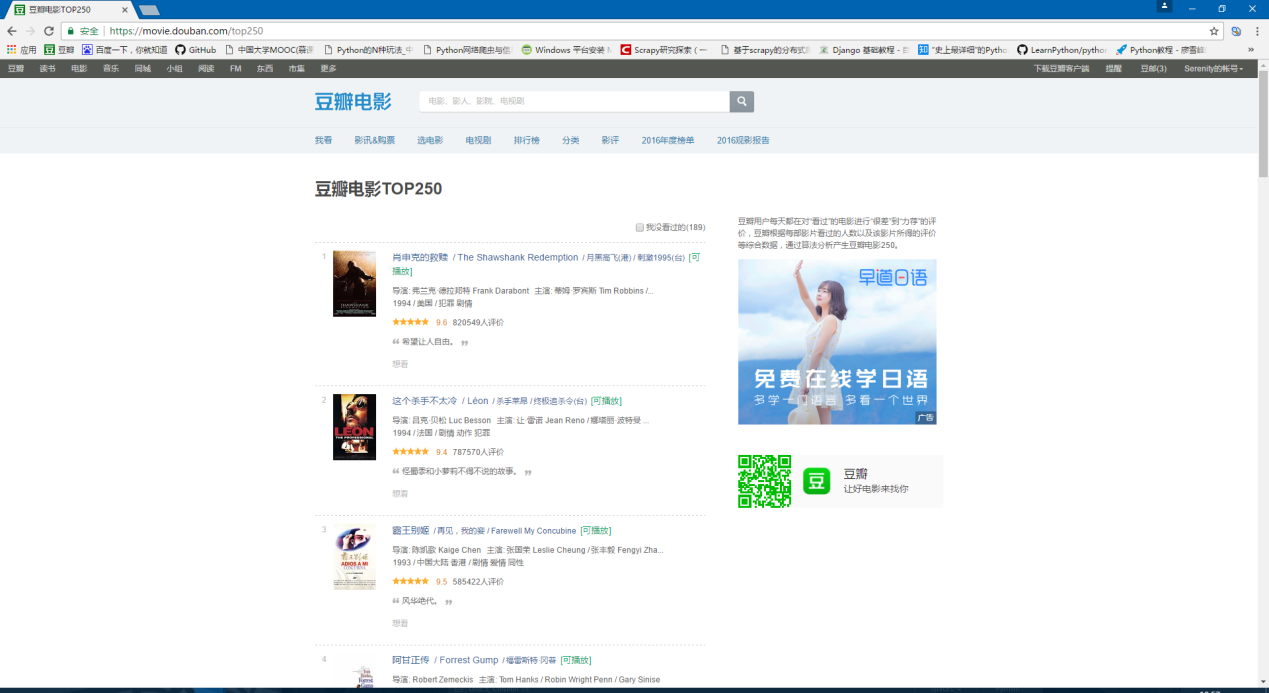


图1

对着网页点右键点击查看网页源代码：

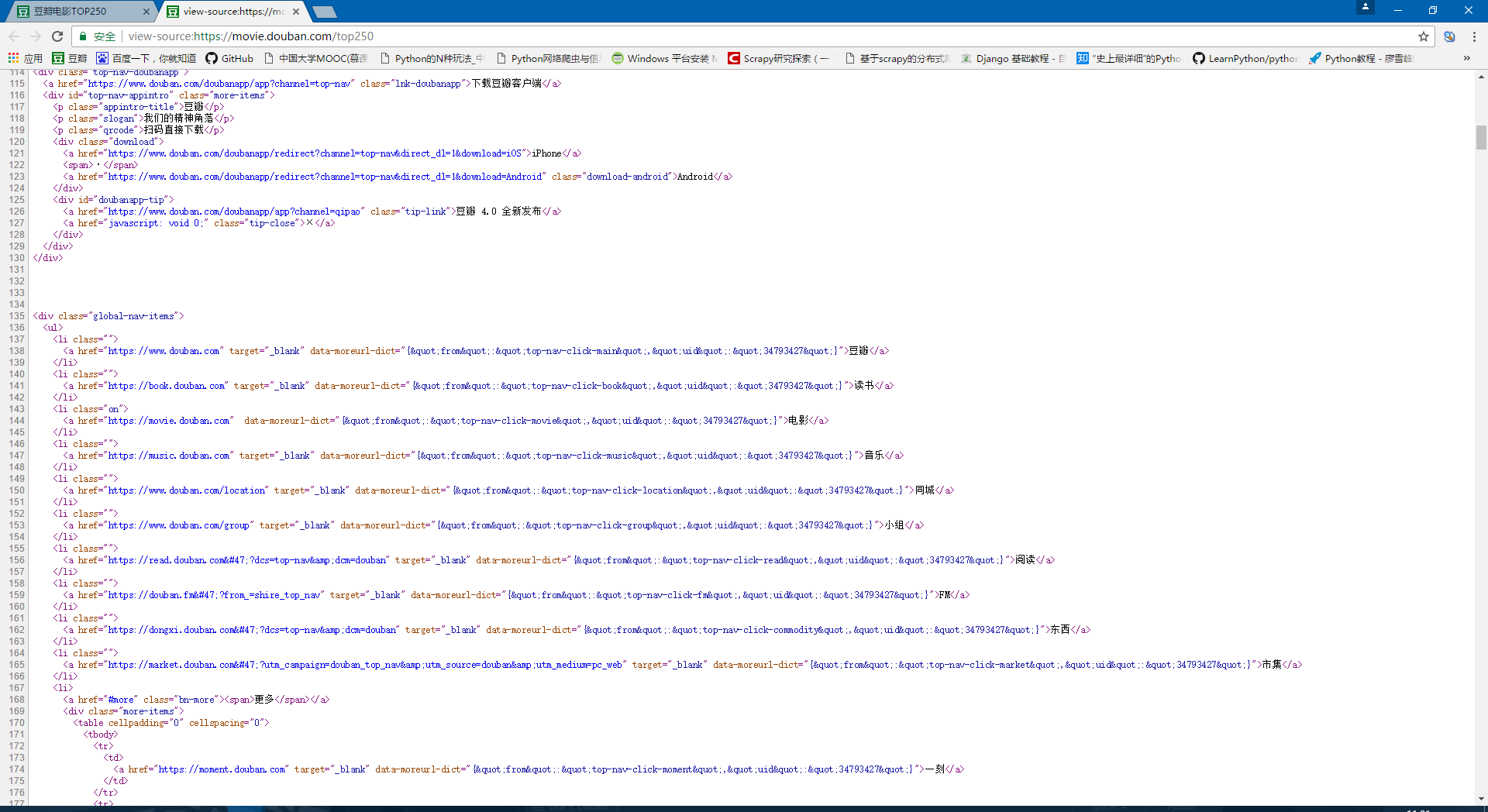
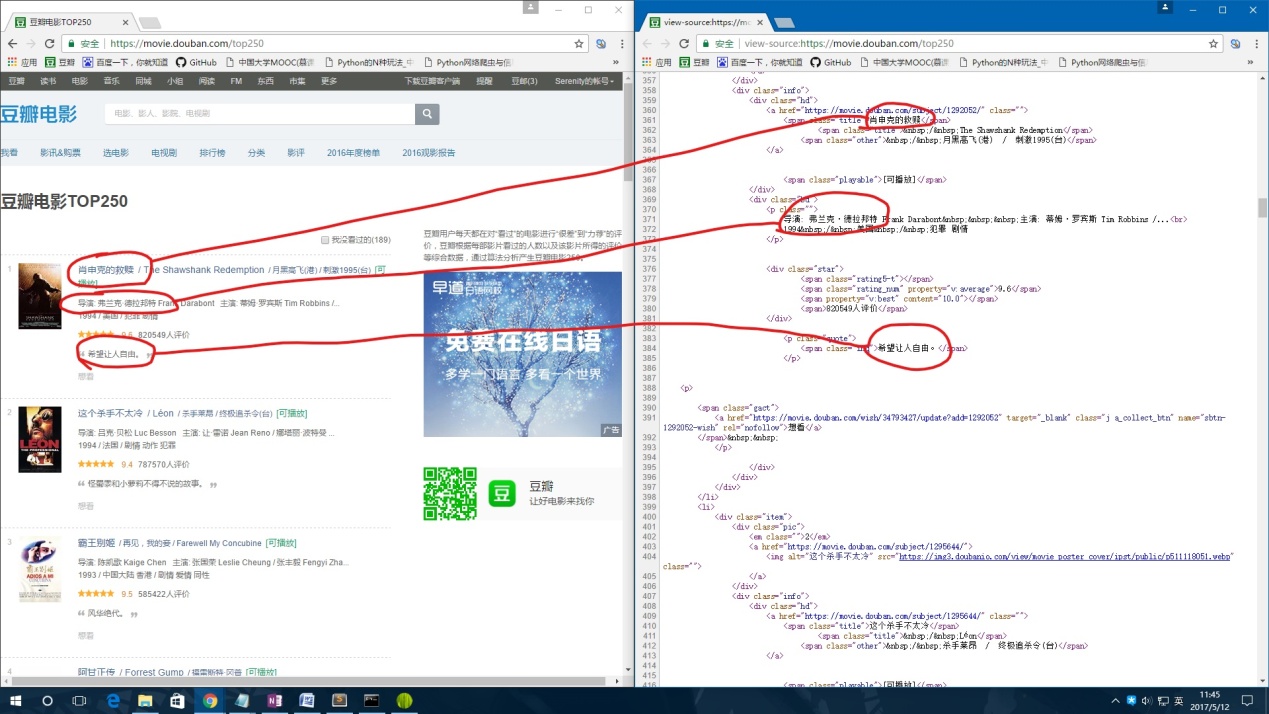


图2

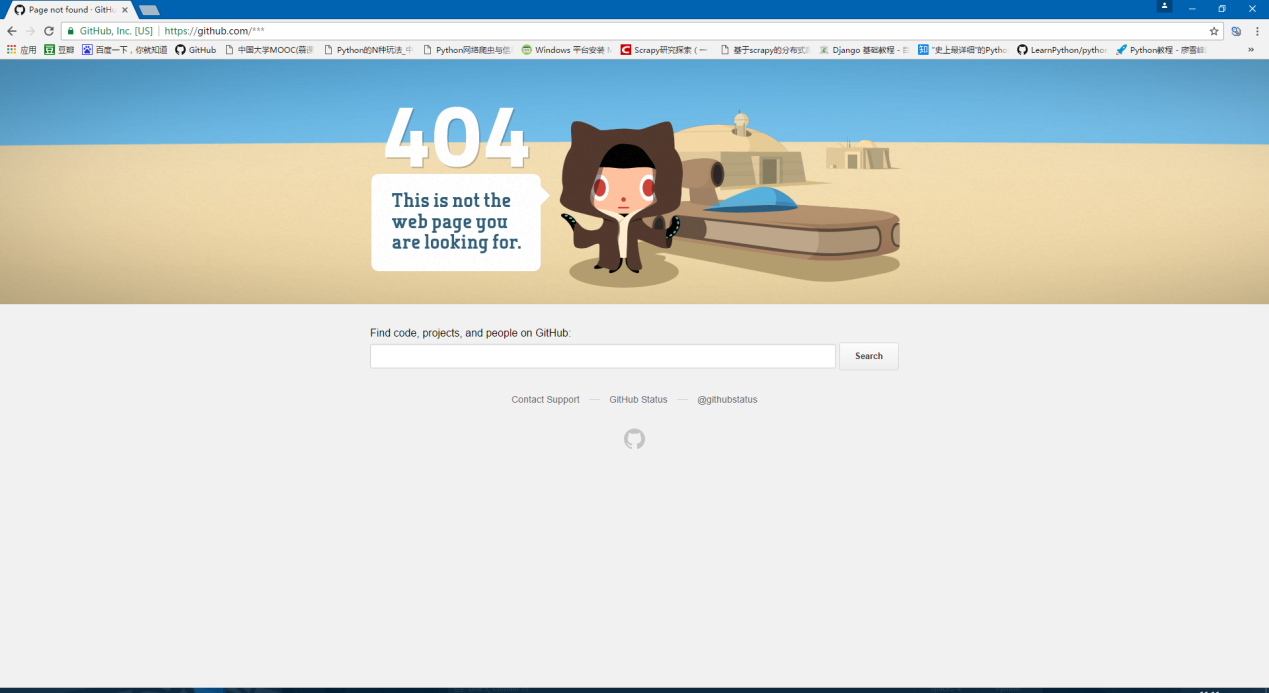
诶，这不就是刚刚用程序得到的东西吗。对的，我们人脑浏览的网页是图1这样的，而爬虫机器人浏览的网页是图2这样的，虽然看起来不同，但里面的内容却是相同的。浏览器将图2中的源代码解析成人类容易理解的图1的样子。



所以，我们这个简单的爬虫程序将这个网页的所有内容都给打印了出来。

**二、更为完善的简单爬虫**

我们有时候打开网页的时候是不是都遇到过这种情况：



可能图片文字会有不同，但是那个大大的404一定让你印象深刻。

404是一个HTTP状态码，它代表服务器找不到请求的网页。可能是你的链接输错了，也可能是原本的网页已经被删除了。

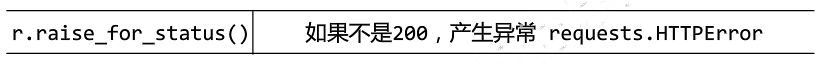
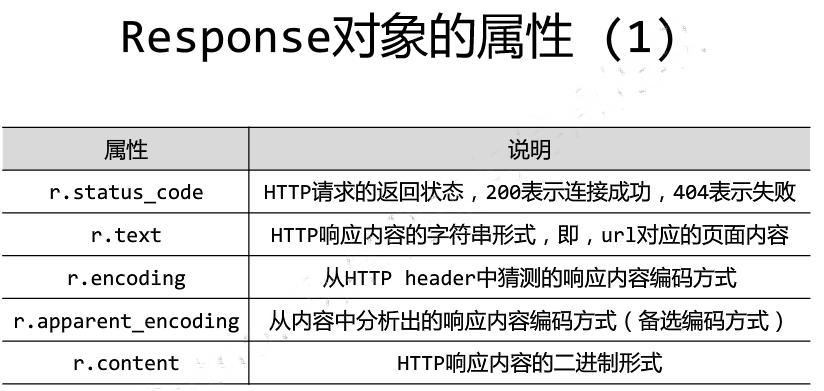
200也是一个HTTP状态码，它代表服务器成功处理了请求，并返回了网页的内容。只有服务器返回200的时候才说明网页访问正常。

（除了404，200，HTTP状态码还有很多，感兴趣可以了解一下）

所以说，我们要爬取的目标网站也有可能会出现无法访问的情况，这时候再运行我们之前那个程序的时候就会报错，所以我们升级一下代码，添加异常处理功能，并将爬取网页的功能写成函数，使代码更健壮。这样，遇到网页无法访问的情况后，程序就知道怎么处理了。

代码如下：





这个程序实现的功能和上一个一样，但是遇到网页无法访问的情况可以正常处理（打印出一个 产生异常）。

**三、提取信息**

光获取网页的源代码远远不够，我们要在源代码中获取我们想要的东西，比如一段文本、链接、图片等。

这里，我们将实现一个可以获得豆瓣电影TOP250电影名字的爬虫程序。

我们已经看到我们在浏览器看到的内容，源代码中都存在。而且我们已经可以获得完整的源代码了，接下来就是在源代码中提取我们想要的内容。

提取信息的方法，这里使用正则表达式。

正则表达式通常被用来检索、替换那些符合某个模式(规则)的文本。

我们现在要在豆瓣电影top250的页面获取电影的名字

我们在网页的源代码中找到带有电影名字的这一行



那么匹配电影名字的正则表达式就是：



聪明的你一定看出来了(.\*?)就是我们要提取的电影的名字。为什么多了两个\呢，因为需要反转义 "。

再举个例子，我们想匹配每个电影的评分，先看一下源代码：



那么正则表达式就是这样：

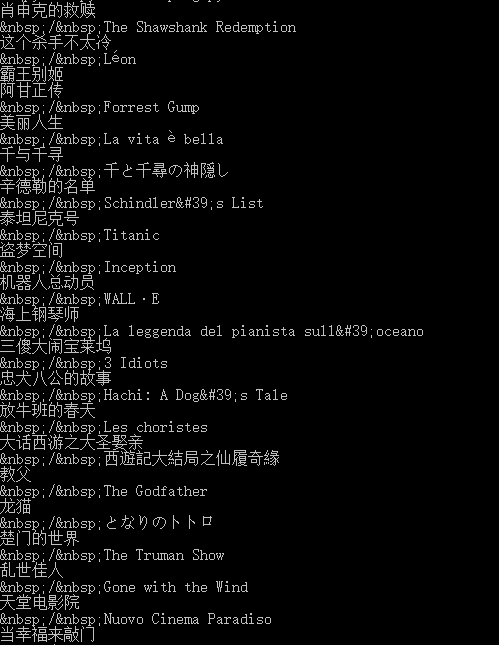


然后使用re模块中的findall方法

findall方法将html中符合传入的正则表达式的文本全部提取出来放到一个列表中。



运行一下，看一下结果：



果然都是电影的名字，而不是全部的源代码，我们的成功了。

等等为什么英文名字前会有&nbsp;/&nbsp;

我们再来看一下源代码：



这说明我们提取的信息并不一定都是纯净的，里面会有我们不想要的杂质。大家可以想一想如何处理。

提示：

如果只想要中文电影名字，可以观察一下是不是可以遇到字符串开头为&的不输出。

如果想同时得到中英文电影名，可以使用python中的split()方法

方法不是唯一的，能解决问题的方法就是好方法。

说明：

爬虫相关的包：本程序使用的是第三方库requests,此外还有urllib,urllib2,bs4,scrappy,pyspider。

提取信息的方法：正则表达式，Beautiful Soup, Lxml等。

参考材料：

中国大学MOOC Python网络爬虫与信息提取 北京理工大学

<http://www.icourse163.org/course/BIT-1001870001>

Python爬虫学习系列教程

<http://cuiqingcai.com/1052.html>

书籍：《用Python写网络爬虫》