 **证书 + 能力** **安全工程师 软件工程师 网站工程师 网络工程师 电脑工程师** **不断增加新科目**
为新手量身定做的课程，让菜鸟快速变身高手 正规公司助您腾飞 **立即加入**

 **全场5折起** 快云VPS、快云服务器、SSL证书、虚拟主机 

红黑联盟13年来最大优惠回馈  **学习套餐最低 6 折起**

首页 > 程序开发 > Web开发 > Python > 正文

搜索

Python与简单网络爬虫的编写

2012-11-30  0 条评论

收藏  我要投稿

回到用Python写爬虫的话题。

Python一直是我主要使用的脚本语言，没有之一。Python的语言简洁灵活，标准库功能强大，平常可以用作计算器，文本编码转换，图片处理，批量下载，批量处理文本等。总之我很喜欢，也越用越上手，这么好用的一个工具，一般人我不告诉他。。。

因为其强大的字符串处理能力，以及urllib2, cookielib, re, threading这些模块的存在，用Python来写爬虫就简直易于反掌了。简单到什么程度呢。我当时跟某同学说，我写电影来了用到的几个爬虫以及数据整理的一堆零零散散的脚本代码行数总共不超过1000行，写电影来了这个网站也只有150来行代码。因为爬虫的代码在另外一台64位的黑苹果上，所以就不列出来，只列一下VPS上网站的代码，tornadoweb框架写的

```
[xiaoxia@307232 movie_site]$ wc -l *.py template/*
```

```
156 msite.py
```

```
92 template/base.html
```

```
79 template/category.html
```

```
94 template/id.html
```

```
47 template/index.html
```

```
77 template/search.html
```

下面直接show一下爬虫的编写流程。以下内容仅供交流学习使用，没有别的意思。

以某湾的最新视频下载资源为例，其网址是

```
http://某piratebay.se/browse/200
```

因为该网页里有大量广告，只贴一下正文部分内容：

暂无图片

对于一个python爬虫，下载这个页面的源代码，一行代码足以。这里用到urllib2库。

```
>>> import urllib2
```

```
>>> html = urllib2.urlopen('http://某piratebay.se/browse/200').read()
```

```
>>> print 'size is', len(html)
```

```
size is 52977
```

当然，也可以用os模块里的system函数调用wget命令来下载网页内容，对于掌握了wget或者curl工具的同学是很方便的。

使用Firebug观察网页结构，可以知道正文部分html是一个table。每一个资源就是一个tr标签。

暂无图片

文章

读书

- Win2000下关闭无用端口
- 禁止非法用户登录综合设置 [win9x篇]
- 关上可恶的后门——消除NetBIOS隐患
- 网络入侵检测系统
- 潜伏在Windows默认设置中的陷阱
- 调制解调器的不安全
- 构建Windows 2000服务器的安全防护林
- SQL Server 2000的安全配置

重金招募VIP讲师

动画教程制作者(各类电脑技术)



邮箱订阅 红黑联盟 精彩内容

立即订阅

点击排行

- python urllib2详解及实例
- python join 和 split的常用使用方法
- 使用Python读取和写入CSV文件
- AttributeError: 'module' object
- python各种类型转换-int,str,char,flo
- Python列表list 数组array常用操作集锦
- python模块介绍- binascii: 二进制和
- python中文decode和encode转码

而对于每一个资源，需要提取的信息有：

- 1、视频分类
- 2、资源名称
- 3、资源链接
- 4、资源大小
- 5、上传时间

就这么多就够了，如果有需要，还可以增加。

首先提取一段tr标签里的代码来观察一下。

```
<tr>
<td class="vertTh">
<center>
<a href="/browse/200" title="此目录中更多">视频</a><br />
(<a href="/browse/205" title="此目录中更多">电视</a>)
</center>
</td>
<td>
<div class="detName"> <a href="/torrent/7782194/The_Walking_Dead_Season_3_Episodes_1-3_HDTV-
x264" class="detLink" title="细节 The Walking Dead Season 3 Episodes 1-3 HDTV-x264">The Walking Dead
Season 3 Episodes 1-3 HDTV-x264</a>
</div>
<a href="magnet:?
xt=urn:btih:4f63d58e51c1a4a997c6f099b2b529bdbba72741&dn=The+Walking+Dead+Season+3+Episodes+1-
3+HDTV-
x264&tr=udp%3A%2F%2Ftracker.openbittorrent.com%3A80&tr=udp%3A%2F%2Ftracker.publicbt.com%3A80&
title="Download this torrent using magnet"></a> <a href="//torrents.某
piratebay.se/7782194/The_Walking_Dead_Season_3_Episodes_1-3_HDTV-x264.7782194.TPB.torrent"
title="下载种子"></a>
<font class="detDesc">已上传 <b>3</b>分钟前</b>, 大小 2<math>\text{GiB}</math>, 上传者 <a class="detDesc"
href="/user/paridha/" title="浏览 paridha">paridha</a></font>
</td>
<td align="right">0</td>
<td align="right">0</td>
</tr>
```

为何要用正则表达式而不用其他一些解析HTML或者DOM树的工具是有原因的。我之前试过用BeautifulSoup3来提取内容，后来发觉速度实在是慢死了啊，一秒钟能够处理100个内容，已经是我电脑的极限了。。。而换了正则表达式，编译后处理内容，速度上直接把它秒杀了！

提取这么多内容，我的正则表达式要如何写呢？

根据我以往的经验，“.*?”或者“.+?”这个东西是很好使的。不过也要注意一些小问题，实际用到的时候就会知道

对于上面的tr标签代码，我首先需要让我的表达式匹配到的符号是

```
<tr>
```

表示内容的开始，当然也可以是别的，只要不要错过需要的内容即可。然后我要匹配的内容是下面这个，获取视频分类。

新闻排行榜

- 1 Unity3d与iOS交互开发——接
- 2 学习Android之SimpleAd
- 3 申通被曝13个安全漏洞 黑客窃取3万
- 4 JAVA垃圾收集器之ParNew收集
- 5 ios之清除cell缓存，解决cel
- 6 电脑上怎样查看微信聊天记录
- 7 Qt on Android：添加分享
- 8 小米开源文件管理器MiCodeFil
- 9 在Linux下安装Oracle数据库
- 10 迹和我之间划下的河

所以说，电影来了网站用到的爬虫不难写，难的是获得数据后如何整理获取有用信息。例如，如何匹配一个影片信息跟一个资源，如何在影片信息库和视频链接之间建立关联，这些都需要不断尝试各种方法，最后选出比较靠谱的。

曾有某同学发邮件想花钱也要得到我的爬虫的源代码。

要是我真的给了，我的爬虫就几百来行代码，一张A4纸，他不会说，坑爹啊！！！.....

都说现在是信息爆炸的时代，所以比的还是谁的数据挖掘能力强

	<ul style="list-style-type: none">■ 培训学校加盟■ 培训机构注册■ 中国校外教育		<ul style="list-style-type: none">■ 韩语培训班■ ui界面设计■ 商标查询
--	--	--	---

点击复制链接 与好友分享!

回本站首页



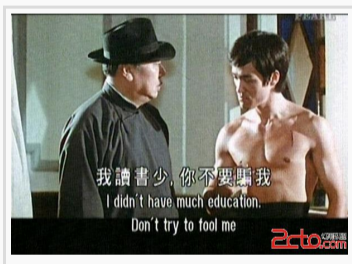
上一篇: [Python学习笔记 \(二\)](#)

下一篇: [关于python 字符编码的一些认识](#)

相关文章

- [用python修改注册表干掉360safe](#)
- [python版本的Access溢出利用程序](#)
- [用python写windows code inject的一](#)
- [python模块Nmap-Parser](#)
- [PHP webshell检查工具 python版](#)
- [Python的url编码函数使用的一个小问题](#)
- [Python批量修改文件后缀脚本](#)
- [python backconnect door](#)
- [python的一个字典创建程序](#)
- [PortScanner in Python 3.1](#)

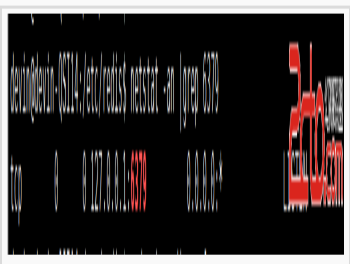
图文推荐



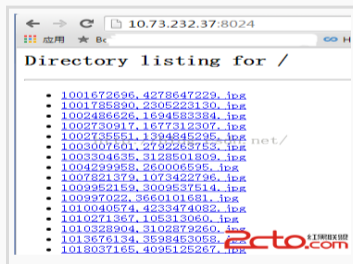
一小时学会用Python



python中遍历dict的v



Ubuntu 14.04 下安



搭建Python HTTP服务

我有话说(0条评论)



发布


还没有评论, 快来抢沙发吧!



热评话题

- 1 android保存图片到本地并可以在相册中显...
- 2 AngularJS实现数据可视化
- 3 光棍节“单身大逃亡” 网易花田新玩法出炉
- 4 【微商必学，自己原创】护肤品微商怎样...
- 5 美剧讲述硅谷创业者故事：远不及现实荒...
- 6 “匿名”黑客组织黑掉五个冰岛政府网站以...
- 7 联想系统更新中发现的两枚提权漏洞原理...

畅言

 **红黑联盟** 2cto.com **资源分享活动** 开始了，上传你的文件 **赚积分** 赢取各种奖品

[关于我们](#) | [联系我们](#) | [广告服务](#) | [投资合作](#) | [版权申明](#) | [在线帮助](#) | [网站地图](#) | [作品发布](#) | [Vip技术培训](#)

版权所有: 红黑联盟--致力于做最好的IT技术学习网站