


BetaBin

BetaBin, a small program in the "world OS", will be executable.

个人资料



BetaBin

访问: 446902次

积分: 5173

等级: 

BLOG 5

排名: 第4242名

原创: 125篇

转载: 29篇

译文: 0篇

评论: 26条

文章搜索

文章分类

汇编 (8)

算法 (19)

Eclipse (1)

碎碎念 (3)

Python (8)

病毒 (9)

Windows (41)

安全 (37)

IT娱乐 (3)

资料 (9)

Sicily (17)

ClamAV (10)

C/C++ (5)

AI (3)

计算机网络 (6)

编译原理 (2)

Androi (0)

Android (2)

文章存档

2014年04月 (3)

2014年03月 (1)

2013年08月 (1)

2013年05月 (1)

2013年04月 (1)

2017直通软考，拿证无忧

程序员简历优化指南！

程序员1月书讯

云端应用征文大赛，秀绝招，赢无人机！

Python:用lxml解析HTML

2014-04-24 09:27

22162人阅读

评论(0)

收藏

举报

分类: Python (7)

转载自: <http://www.cnblogs.com/descusr/archive/2012/06/20/2557075.html>

=====

先演示一段获取页面链接代码示例:

#coding=utf-8

from lxml import etree

html = ""

<html>

<head>

<meta name="content-type" content="text/html; charset=utf-8" />

<title>友情链接查询 - 站长工具</title>

<!-- uRj0Ak8VLEPhjWhg3m9z4EjXJwc -->

<meta name="Keywords" content="友情链接查询" />

<meta name="Description" content="友情链接查询" />

</head>

<body>

<h1 class="heading">Top News</h1>

<p style="font-size: 200%">World News only on this page</p>

Ah, and here's some more text, by the way.

<p>... and this is a parsed fragment ...</p>

<a href="http://www.cydf.org.cn/" rel="nofollow" target="\_blank">青少年发展基金会</a>

<a href="http://www.4399.com/flash/32979.htm" target="\_blank">洛克王国</a>

<a href="http://www.4399.com/flash/35538.htm" target="\_blank">奥拉星</a>

<a href="http://game.3533.com/game/" target="\_blank">手机游戏</a>

<a href="http://game.3533.com/tupian/" target="\_blank">手机壁纸</a>

<a href="http://www.4399.com/" target="\_blank">4399小游戏</a>

<a href="http://www.91wan.com/" target="\_blank">91wan游戏</a>

</body>

</html>

""

page = etree.HTML(html.lower()).decode('utf-8'))

hrefs = page.xpath(u'//a')

for href in hrefs:

print href.attrib

打印出的结果为:

{'href': 'http://www.cydf.org.cn/', 'target': '\_blank', 'rel': 'nofollow'}

{'href': 'http://www.4399.com/flash/32979.htm', 'target': '\_blank'}

{'href': 'http://www.4399.com/flash/35538.htm', 'target': '\_blank'}

展开

阅读排行

- Python:用lxml解析HTML (22151)
- 编译原理中的正则表达式 (15656)
- 编译原理中正则表达式直 (8842)
- IDAPython插件安装 (8441)
- 解决问题：开启Wiresha (8219)
- 运输层可靠数据传输的原 (7071)
- vs2008加载dll深入 (6848)
- Win32 汇编 - 逻辑运算指 (6215)
- Python简单抓取新浪某网 (5815)
- Win32 汇编 - 乘除指令:l (5722)

评论排行

- ClamAV学习【9】——c (4)
- PEB及PEB\_LDR\_DATA (3)
- 链路层多路访问协议 (2)
- Python备份CSDN博客 (2)
- 算法学习【12】—— 11 (2)
- 编译原理中正则表达式直 (2)
- 编译原理中的正则表达式 (2)
- ClamAV学习【8】——6 (2)
- Dijkstra算法实现类—提 (2)
- 算法学习【14】—— 11 (2)

最新评论

- 碎碎念【1】- 新的不一定好——rtrtcc: August 03, 2011 - OllyDbg 2.01 alpha 4. Here is AI...
- Dijkstra算法实现类—提高，邻接艾尔杰弗森: 加上注释就好啦
- 算法学习【14】—— 1190. Redi 疯。不觉: TLE 了
- 链路层多路访问协议猪皮冻: 这是计算机网络（自上而下）里面的一段，哈哈，我正在读这本书
- ClamAV学习【8】——64位Win7 diggold: pthreadv2用最低的版本即可有个地方size\_t没有定义，需要定义一下
- 编译原理中正则表达式直接构造llyq374888272: 大牛，有木有代码
- ClamAV学习【8】——64位Win7 x3x2012: 博主，能留下你的QQ号码吗？我有很多Clamav编译问题向你请教，谢谢。
- 算法学习【12】—— 1155. Can 家泪: 我觉得这里的循环应该修改一下，for(i = 0; i < cityNum; i++) ...
- 算法学习【12】—— 1155. Can 家泪: 虽然这样写也能通过，不过感觉算法有问题。下面这个测试用例：530 33 11 4理论上这个用例...
- APK批量反编译到Java RLib: .....

```
{'href': 'http://game.3533.com/game/', 'target': '_blank'}
{'href': 'http://game.3533.com/tupian/', 'target': '_blank'}
{'href': 'http://www.4399.com/', 'target': '_blank'}
{'href': 'http://www.91wan.com/', 'target': '_blank'}
```

如果要取得<a></a>之间的内容，

```
for href in hrefs:
    print href.text
```

结果为：

青少年发展基金会  
洛克王国  
奥拉星  
手机游戏  
手机壁纸  
4399小游戏  
91wan游戏

使用lxml前注意事项：先确保html经过了utf-8解码，即code = html.decode('utf-8', 'ignore')，否则会出  
现解析出错情况。因为中文被编码成utf-8之后变成'/u2541' 之类的形式，lxml一遇到"/"就会认为其标签  
结束。

XPATH基本上是用一种类似目录树的方法来描述在XML文档中的路径。比如用"/"来作为上下层级间的分  
隔。第一个"/"表示文档的根节点（注意，不是指文档最外层的tag节点，而是指文档本身）。比如对于一个  
HTML文件来说，最外层的节点应该是"/html"。

定位某一个HTML标签，可以使用类似文件路径里的绝对路径，如page.xpath(u"/html/body/p")，它会找到  
body这个节点下所有的p标签；也可以使用类似文件路径里的相对路径，可以这样使用：  
page.xpath(u"//p"),它会找到整个html代码里的所有p标签：

```
<p style="font-size: 200%">World News only on this page</p>
Ah, and here's some more text, by the way.
<p>... and this is a parsed fragment ...</p>
```

注意：XPATH返回的不一定就是唯一的节点，而是符合条件的所有节点。如上所示，只要是body里的p标  
签，不管是body的第一级节点，还是第二级，第三级节点，都会被取出来。

如果想进一步缩小范围，直接定位到"<p style="font-size: 200%">World News only on this  
page</p>"要怎么做呢？这就需要增加过滤条件。过滤的方法就是用"[]"把过滤条件加上。lxml里有个过滤  
语法：

```
p = page.xpath(u"/html/body/p[@style='font-size: 200%']")
或者：p = page.xpath(u"//p[@style='font-size:200%']")
```

这样就取出了body里style为font-size:200%的p节点，注意：这个p变量是一个lxml.etree.\_Element对象  
列表，p[0].text结果为World News only on this page，即标签之间的值；p[0].values()结果为font-size:  
200%，即所有属性值。其中 @style表示属性style，类似地还可以使用如@name, @id, @value, @href,  
@src, @class....

如果标签里面没有属性怎么办？那就可以用text(), position()等函数来过滤，函数text()的意思则是取得  
节点包含的文本。比如：<div>hello<p>world</p></div>中，用"div[text()='hello']"即可取得这个div，而  
world则是p的text()。函数position()的意思是取得节点的位置。比如"l[position()=2]"表示取得第二个l节点，  
它也可以被省略为"l[2]"。

不过要注意的是数字定位和过滤 条件的顺序。比如"u/l[5][@name='hello']"表示取ul下第五项l，并且其  
name必须是hello，否则返回空。而如果用"u/l[@name='hello'][5]"的意思就不同，它表示寻找ul下第五个  
name为"hello"的l节点。

此外，"\*"可以代替所有的节点名，比如用"/html/body/\*/span"可以取出body下第二级的所有span，而  
不管它上一级是div还是p或是其它什么东东。

而"descendant::"前缀可以指代任意多层的中间节点，它也可以被省略成一个"/"。比如在整個HTML文档中  
查找id为"leftmenu"的 div，可以用"/descendant::div[@id='leftmenu']"，也可以简单地使用"  
//div[@id='leftmenu']"。

text = page.xpath(u"/descendant::\*[text()='']")表示任意多层的中间节点下任意标签之间的内容，也即实现蜘蛛  
抓取页面内容功能。以下内容使用text属性是取不到的：





晚上十点开始的副业收入 - 一份特别收入，每晚十点准时开始  
每周都有至少一笔收入自动打入你的账户，无需工作一天，睡觉时都在赚钱



猜你在找

- WEB前端整套教程html+divcss+javascript+jquery+htm
- Python 开源项目
- HTML 入门视频课程
- Python渗透测试开源项目
- html+div+css零基础快速入门到制作企业站视频课程
- 搜索引擎 - Python下开源爬虫spider框架scrapy的使用
- 在HTML5画布上绘制炫酷太阳系
- python以gzip header请求html数据时response内容乱码
- 零基础学习HTML5—html+css基础
- Python实战之自动化评论

查看评论

暂无评论

您还没有登录,请[\[登录\]](#)或[\[注册\]](#)

\* 以上用户言论只代表其个人观点，不代表CSDN网站的观点或立场

核心技术类目

- 全部主题
- Hadoop
- AWS
- 移动游戏
- Java
- Android
- iOS
- Swift
- 智能硬件
- Docker
- OpenStack
- VPN
- Spark
- ERP
- IE10
- Eclipse
- CRM
- JavaScript
- 数据库
- Ubuntu
- NFC
- WAP
- jQuery
- BI
- HTML5
- Spring
- Apache
- .NET
- API
- HTML
- SDK
- IIS
- Fedora
- XML
- LBS
- Unity
- Splashtop
- UML
- components
- Windows Mobile
- Rails
- QEMU
- KDE
- Cassandra
- CloudStack
- FTC
- coremail
- OPhone
- CouchBase
- 云计算
- iOS6
- Rackspace
- Web App
- SpringSide
- Maemo
- Compuware
- 大数据
- aptech
- Perl
- Tornado
- Ruby
- Hibernate
- ThinkPHP
- HBase
- Pure
- Solr
- Angular
- Cloud Foundry
- Redis
- Scala
- Django
- Bootstrap

公司简介 | 招贤纳士 | 广告服务 | 联系方式 | 版权声明 | 法律顾问 | 问题报告 | 合作伙伴 | 论坛反馈

网站客服 杂志客服 微博客服 webmaster@csdn.net 400-600-2320 | 北京创新乐知信息技术有限公司 版权所有 | 江苏知之为计算机有限公司 |

江苏乐知网络技术有限公司

京 ICP 证 09002463 号 | Copyright © 1999-2016, CSDN.NET, All Rights Reserved