gaomatlab

随笔 - 150, 文章 - 0, 评论 - 4, 引用 - 0

导航

博客园
首页
新随笔
联系
订阅
管理

< 2019年11月 >

日	一	二	三	四	五	六
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
1	2	3	4	5	6	7

搜索

找找看
谷歌搜索

常用链接

python中使用XPath

XPath在Python的爬虫学习中，起着举足轻重的地位，对比正则表达式 re两者可以完成同样的工作，实现的功能也差不多，但XPath明显比re具有优势，在网页分析上使re退居二线。

XPath介绍:

是什么？ 全称为**XML Path Language** 一种小型的**查询语言**

说道XPath是门语言，不得不说它所具备的优点：

1） 可在XML中查找信息

2） 支持HTML的查找

3） 通过元素和属性进行导航

python开发使用XPath条件:

由于XPath属于lxml库模块，所以首先要安装库lxml，具体的安装过程可以查看博客，包括easy_install 和 pip 的安装方法。

XPath的简单调用方法:

```
from lxml import etree

selector=etree.HTML(源码) #将源码转化为能被XPath匹配的格式

selector.xpath(表达式) #返回为一列表
```

XPath的使用方法:

首先讲一下XPath的基本语法知识：

我的随笔
我的评论
我的参与
最新评论
我的标签

我的标签

linux(19)
python(9)
hadoop(9)
mapreduce(7)
top(4)
vi(3)
shell(3)
hbase(3)
hive(2)
crontab(2)
更多

随笔分类

java学习总结(3)
jquery学习(3)
js学习
linux 学习(21)
linux学习笔记--常用命令篇(34)
linux学习笔记--交付运维篇(8)
linux学习笔记--开发编程篇
linux学习笔记--问题处理篇(1)
MySQL(3)
oracle数据库学习(9)
python学习笔记--基础知识篇(13)
python学习笔记--网络爬虫篇
sql学习(24)
大数据--flume系列
大数据--Hadoop系列(7)

四种标签的使用方法

1) **//** 双斜杠 定位根节点，会对全文进行扫描，在文档中选取所有符合条件的内容，以列表的形式返回。

2) **/** 单斜杠 寻找当前标签路径的下一层路径标签或者对当前路标签内容进行操作

3) **/text()** 获取当前路径下的文本内容

4) **/@xxxx** 提取当前路径下标签的属性值

5) **|** 可选符 使用|可选取若干个路径 如//p | //div 即在当前路径下选取所有符合条件的p标签和div标签。

6) **.** 点 用来选取当前节点

7) .. 双点 选取当前节点的父节点

另外还有starts-with(@属性名称,属性字符相同部分)，string(.)两种重要的特殊方法后面将重点讲。

利用实例讲解XPath的使用：

```
from lxml import etree
html="""
<!DOCTYPE html>
<html>
<head lang="en">
<title>测试</title>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
</head>
<body>
<div id="content">
<ul id="ul">
<li>NO.1</li>
<li>NO.2</li>
<li>NO.3</li>
</ul>
<ul id="ul2">
```

随笔档案

最新评论

1. Re:linux shell 字符串操作详解 （长度，读取，替换，截取，连接，对比，删除，位置 ）
太棒了，终于找到了，感谢博主

```
<li>one</li>
<li>two</li>
</ul>
</div>
<div id="url">
<a href="http:www.58.com" title="58">58</a>
<a href="http:www.csdn.net" title="CSDN">CSDN</a>
</div>
</body>
</html>
"""
selector=etree.HTML(html)
content=selector.xpath('//div[@id="content"]/ul[@id="ul"]/li/text()') #这里使用id属性来定位哪个div和ul被匹配 使用text()获取文本内容
for i in content:
print i
#输出为
NO.1
NO.2
NO.3

con=selector.xpath('//a/@href') #这里使用//从全文中定位符合条件的a标签，使用“@标签属性”获取a便签的href属性值
for each in con:
print each
#输出结果为:
http:www.58.com
http:www.csdn.net
```

con=selector.xpath('/html/body/div/a/@title') #使用绝对路径20 <a href="http:www.csdn.2Fa/@title') #使用相对路径定位 两者效果是一样的
print len(con)
print con[0]con[1]

#输出结果为:
2
58 CSDN

介绍XPath的特殊用法:

1) **starts-with** 解决标签属性值以相同字符串开头的情况

举例说明

```
from lxml import etree
html="""
    <body>
        <div id="aa">aa</div>
        <div id="ab">ab</div>
        <div id="ac">ac</div>
    </body>
    """
selector=etree.HTML(html)
content=selector.xpath('//div[starts-with(@id,"a")]/text()') #这里使用starts-with方法提取div的id标签属性值开头为a的div标签
for each in content:
    print each
#输出结果为:
aa
ab
ac
```

推荐排行榜

1. linux shell 字符串操作详解 （长度，读取，替换，截取，连接，对比，删除，位置）(5)
2. python中使用XPath(2)

2） **string(.)** 标签套标签

```
html="""
    <div id="a">
    left
        <span id="b">
        right
            <ul>
            up
                <li>down</li>
            </ul>
        east
        </span>
        west
    </div>
"""
#下面是没有用string方法的输出
sel=etree.HTML(html)
con=sel.xpath('//div[@id="a"]/text()')
for i in con:
    print i    #输出内容为left west

data=sel.xpath('//div[@id="a"]')[0]
info=data.xpath('string(.)')
content=info.replace('\n','').replace(' ','')
for i in content:
    print i #输出为 全部内容
```

XPath提供的几个特殊的方法：

XPath中需要取的标签如果没有属性，可以使用text()，posision()来识别标签。

举两个简单的例子：

```
from lxml import etree
html="""
    <div>hello
        <p>H</p>
</div>
<div>hehe</div>
"""
sel=etree.HTML(html)
con=sel.xpath('//div[text()="hello"]/p/text()')
print con[0]
#H
```

这里使用**text()**的方法来判别是哪个div标签

```
from lxml import etree
html="""
    <div>hello
        <p>H</p>
        <p>J</p>
        <p>I</p>
</div>
<div>hehe</div>
"""
sel=etree.HTML(html)
con=sel.xpath('//div[text()="hello"]/p[posision()=2]/text()')
print con[0]
#J
```

另外，在XPath中可以使用多重过滤方法寻找标签，例如ul[3][@id=”a”] 这里使用【3】来寻找第三个ul标签 并且它的id属性值为a

获取XPath的方式有两种：

1） 使用以上等等的方法通过观察找规律的方式来获取XPath

2） 使用Chrome浏览器来获取 在网页中右击->选择审查元素（或者使用F12打开） 就可以在elements中查看网页的html标签了，找到你想要获取XPath的标签，右击->Copy XPath 就已经将XPath路径复制到了剪切板。

分类: python学习笔记--基础知识篇

标签: python, XPath.

gaomatlab
关注 - 18
粉丝 - 18
+加关注

2 0

« 上一篇： python学习笔记3---浅拷贝和深拷贝，file操作
» 下一篇： 常用的几个linux命令

posted on 2017-04-24 16:22 gaomatlab 阅读(32034) 评论(2) 编辑 收藏

评论

#1楼

感谢

支持(0) 反对(0)

2019-05-10 19:37 | 舵者

#2楼

最后的为 position

支持(0) 反对(0)

2019-05-10 20:02 | 舵者

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册， 访问 网站首页。

相关博文：
· python中使用XPath
· python中使用XPath
· python中使用XPath笔记
· XPath在python中的高级应用
· Python中利用xpath解析HTML
» 更多推荐...

最新 IT 新闻:
· 谷歌与Ascension达成云计算合作协议
· 阿迪达斯关闭欧美机器人智能工厂 还是中国、越南便宜
· 联发科5G芯片"官宣"：11月26日深圳正式发布
· 河豚毒素致命，竟也能缓解压力
· 官宣！荣耀V30将于11月26日发布：双模5G全国通、Matrix镜头加持
» 更多新闻...