

# 专栏：007：xpath使用及其实战

## 系列爬虫专栏

崇尚的学习思维是：输入，输出平衡，且平衡点不断攀升。

曾经有大神告诫说：没事别瞎写文章；所以，很认真的写的是能力范围内的，看客要是看不懂，不是你的问题，问题在我，得持续输入，再输出。

今天的主题是：xpath的使用及其当当心理学图书抓取

## 1：框架

| 序号 | 内容          | 说明 |
|----|-------------|----|
| 01 | 概念          | —  |
| 02 | xpath语法     | —  |
| 03 | 语法实例        | —  |
| 04 | 实战当当心理学图书抓取 | —  |
| 05 | 参考及总结       | —  |

## 2：概念

- Xpath  
XPath一门在 XML 文档中查找信息的语言。XPath即为XML路径语言（XML Path Language），它是一种用来确定XML文档中某部分位置的语言。  
XPath基于XML的树状结构，提供在数据结构树中找寻节点的能力。起初XPath的提出的初衷是将其作为一个通用的、介于XPointer与XSL间的语法模型。但是XPath很快的被开发者采用来当作小型查询语言。

复述：相当于定位地址，比如，我要找清华大学在哪：30 Shuangqing Rd, Haidian, Beijing, China。先定位在中国，再定位在北京，再定位在海淀区，继续定位具体的街道地址。  
那比如你定位到北京：那北京很多区。  
那再比如你定位到海淀区：那海淀区也分很多地方。

在 XPath 中，有七种类型的节点：元素、属性、文本、命名空间、处理指令、注释以及文档节点（或称为根节点）。(解析就是对这些节点进行定位提取需要的信息)

- lxml  
lxml 是一种使用Python 编写的库，可以迅速、灵活地处理XML。 它支持XML Path Language (XPath) 和Extensible Stylesheet Language Transformation (XSLT)，并且实现了常见的ElementTree API。

lxml是python第三方库，需要自己安装。安装会遇到很些问题，还是那句话：生命不息，折腾不止。

### 3：xpath语法

表格法：

| 序号 | 表达式      | 描述                           |
|----|----------|------------------------------|
| 01 | nodename | 选取此节点的所有子节点                  |
| 02 | /        | 从根节点选取                       |
| 03 | //       | 从匹配选择的当前节点选择文档中的节点，而不考虑它们的位置 |
| 04 | .        | 选取当前节点                       |
| 05 | ..       | 选取当前节点的父节点                   |
| 06 | @        | 选取属性                         |

看不懂？

推荐：chrome 插件：[XPath Helper](#)

chrome 浏览器也可以copy xpath.

打不开？那搜索引擎玩着吧.

效果显示：  
[w3school:xpath 教程](#)  
想要搜索到：路径表达式：如下表：中的nodename单词

下面列出了最有用的路径表达式：

| 表达式      | 描述                            |
|----------|-------------------------------|
| nodename | 选取此节点的所有子节点。                  |
| /        | 从根节点选取。                       |
| //       | 从匹配选择的当前节点选择文档中的节点，而不考虑它们的位置。 |
| .        | 选取当前节点。                       |
| ..       | 选取当前节点的父节点。                   |
| @        | 选取属性。                         |

层层查询下来：

W3School

QUERY

```
<book>
+title lang="eng" sleeping XM / (11)
///*[id="maincontent"]/div[4]/table[1]/tbody/tr[2]/td[1]
</book>
</bookstore>
```

RESULTS (1)

```
nodename
```

选取节点

XPath 使用路径表达式在 XML 文档中选取节点。节点是通过沿着路径或者 step 来选取的。

下面列出了最有用的路径表达式：

| 表达式      | 描述                            |
|----------|-------------------------------|
| nodename | 选取此节点的所有子节点。                  |
| /        | 从根节点选取。                       |
| //       | 从匹配选择的当前节点选择文档中的节点，而不考虑它们的位置。 |
| .        | 选取当前节点。                       |
| ..       | 选取当前节点的父节点。                   |
| @        | 选取属性。                         |

Elements Console

取节点。节点是通过沿着路径或者 step 来选取的。 </p> <h3>下面列出了最有用的路径表达式： </h3> <table class="dataintable"> <tbody class= <tr class=</tr> <tr class= <td class="xh-highlight" nodename </td> <td class= </td> </tr> <tr class=</tr> <tr class=</tr> <tr class=</tr> <tr class=</tr> <tr class=</tr> </tbody> </table> <h3>实例 </h3> <p>在下面的表格中，我们已列出了一些路径表达式以及表达式的结果： </p> <table class="dataintable"> <tbody> <tr> <td> </td> <td> </td> </tr> <tr> <td> </td> <td> </td> </tr> <tr> <td> </td> <td> </td> </tr> <tr> <td> </td> <td> </td> </tr> <tr> <td> </td> <td> </td> </tr> </tbody> </table> <div id="sidebar"> </div> <div id="footer"> </div> </div>

## 4：语法实例

xpath语法实例：chrome 插件：Xpath helper

目标：当当心理学图书[链接](#)

- 图书名称：Bookname `xpath = //li/div/a/@title`

W3School

QUERY

```
//li/div/a/@title
(理言词与治方(4894)
心理学理论与实践(6902)
心理学(111111111))
```

RESULTS (60)

```
自控力
天才在左 疯子
在右 (完整版)
(新增10个被封杀
章节！看高智商疯
子如何闹戏和羞辱
正常人！由陈
小春、应采儿倾力
演出的同名改编
剧！)
★★★★★ 189101条评论
凯利麦格尼格尔 著，王岑卉 译 / 2012-08-01 / 文化发展出版社
★《自控力2》，“自控力”实践应用版 如果你想让生活变得更美好，就从自控力入手吧。 自控力强的人能
绪和行为，更好地应对压力、解决冲突、战胜逆境，身体更健康，人际关系更和谐，恋情更长久，收入更
高，彻底告别拖延带来的恐惧和焦虑。学会时间管理，看
见更好的自己。
当当自营
¥22.80 定价：¥39.80 (5.73折)
```

当当网

图书

心理学

自控力

天才在左 疯子
在右 (完整版)
(新增10个被封杀
章节！看高智商疯
子如何闹戏和羞辱
正常人！由陈
小春、应采儿倾力
演出的同名改编
剧！)
★★★★★ 189101条评论
凯利麦格尼格尔 著，王岑卉 译 / 2012-08-01 / 文化发展出版社
★《自控力2》，“自控力”实践应用版 如果你想让生活变得更美好，就从自控力入手吧。 自控力强的人能
绪和行为，更好地应对压力、解决冲突、战胜逆境，身体更健康，人际关系更和谐，恋情更长久，收入更
高，彻底告别拖延带来的恐惧和焦虑。学会时间管理，看
见更好的自己。
当当自营
¥22.80 定价：¥39.80 (5.73折)

Elements Console

<a href="http://
comm.dangdang.com/review/
reviewlist.php?
pid=23705754&ddclick?ac=
405867486431&ref=&rcount=&
type=&t=1462083033000&search
api\_version=test\_new
target="\_blank" name="P\_pl"
class="4180条评论"> </a>
</p>
<div class="publisher\_info">
</div>
<div>
<p class="detail"> </p>
<p class="dang" style=
"display: block"> 当当自
营 </p>
<p class="buy\_button"> </p>
<p class="price"> </p>
<p class="subtitle"> 每个人都
拥有让自己成功快乐的能力！经典升
级版，国际NLP大师李中莹“经典著
作，国内“权威的NLP入门书！ </p>
<span class="tag\_box" style=
"background:url(http://
img4.ddimg.cn/00035/pic/
hg.png) no-repeat 0 0;

- 作者：Writer

`xpath=//div[@class="publisher_info"]/p[@class="author"]/a[1]//@title`

QUERY 经典著作(2400)

价格: 3元以下 3-7元 7-10元 10-30

RESULTS (60)

//div[@class="publisher\_info"]/p[@class="author"]/a[1]//@title

凯利麦格尼格尔 著, 王岑卉 译

高铭 著

(奥) 西格蒙德·弗洛伊德 著, 孙名之 等译

辰格 编著

(法) 勒庞 著, 冯克利 译

排序: 默认排序 销量 + 价格 + 好评 +

换购

自控力

(斯坦福大学“受欢迎心理学课程”, 提高自控力的“有效途径”, 在于弄清自己如何失控、为何失控)

★★★★★ 189101条评论

凯利麦格尼格尔 著, 王岑卉 译 / 2012-08-01 / 文化发展出版社

★《自控力2》, “自控力”实践应用版 如果你想让生活变得更美好, 就从自控力入手吧。自控力强的人能绪和行为, 更好地应对压力、解决冲突、战胜逆境, 身体更健康, 人际关系更和谐, 恋情更长久, 收入更

当当自营

¥22.80 定价: ¥39.80 (5.73折)

换购

天才在左 疯子在右 (完整版)

(新增10个被封杀篇章! 看高智商疯子如何闹戏和羞辱正常人! 由陈小春、应采儿倾力演出的同名改编剧)

★★★★★ 19226条评论

高铭 著 / 2016-01-01 / 北京联合出版公司

5年前, 一本默默无闻的书一经出版便迅速占据各大图书排行榜首。《天才在左 疯子在右》, 没有浮夸的简单的对话形式, 却在5年间以百万余册的畅销量级, 撼动了所有人自以为稳固的世界观。5年后, 这本

当当自营

¥26.60 定价: ¥39.80 (6.69折)

儿童心理学(1515)

社会心理学(1248)

教育与发展心理学(2203)

童书

小说

考试

中小学教辅

文学

青春文学

外语

成功/励志

管理

展开

推广商品

<p class="author">

<span></span>

<a href="http://search.dangdang.com/?key2=凯利麦格尼格尔&medium=01&category\_path=01.00.00.00.00.00" name="p\_zz" title="凯利麦格尼格尔">凯利麦格尼格尔</a>

" 著, "

<a href="http://search.dangdang.com/?key2=王岑卉&medium=01&category\_path=01.00.00.00.00.00" name="p\_zz" title="王岑卉">王岑卉</a>

" 译"

</p>

><p class="publishing\_time">

...</p>

><p class="publishing"></p>

</div>

><p class="detail"></p>

><p class="dang" style="display: block;">当当自

></p>

><p class="buy\_button"></p>

><p class="price"></p>

><p class="subtitle"> (斯坦福大学“受欢迎心理学课程”, 提高自控

- 出版时间：Time xpath =  
//div[@class="publisher\_info"]/p[@class="publishing\_time"]
- 评价数：Star xpath = //p[@class="star"]/a
- 简介：Detail xpath = //p[@class="detail"]
- 图书售价：Price\_n, xpath =  
//div[@class="inner"]/p[@class="price"]/span[@class="price\_n"]
- 图书定价：Price\_r xpath =  
//div[@class="inner"]/p[@class="price"]/span[@class="price\_r"]
- 网址链接：Url xpath = //div[@class="inner"]/a[@href]

## 5：实战心理学图书

在lxml下如何使用：

```
selector = etree.HTML(response) response为网页源代码
```

抓取：图书标题，评价人数实例：

```
# title属性是书名
self.Bookname_pattern_3 = r"//li/div/a/@title"
selector = etree.HTML(response)
booknames = selector.xpath(self.Bookname_pattern_3) # 返回一个list
# 评价人数
self.Star_pattern_3 = r'//p[@class="star"]/a/text()'
selector = etree.HTML(response)
stars = selector.xpath(self.Star_pattern_3) # 返回一个list
```

# 核心代码:

```
def contents_xpath(self, one_url):
    html = requests.get(one_url, headers=self.headers)
    if html.status_code != 200:
        return -1
    else:
        response = html.text
        selector = etree.HTML(response)
        booknames = selector.xpath(self.Bookname_pattern_3)
        writers = selector.xpath(self.Writer_pattern_3)
        time = selector.xpath(self.Time_pattern_3)
        stars = selector.xpath(self.Star_pattern_3)
        details = selector.xpath(self.Detail_pattern_3)
        price_n = selector.xpath(self.Price_n_pattern_3)
        price_r = selector.xpath(self.Price_r_pattern_3)
        urls = selector.xpath(self.Url_pattern_3)
        All_data = []
        for booknames, writers, time, stars, details, price_n, price_r, urls:
            data = {
                "bookname": booknames,
                "writers": writers,
                "stars": stars,
                "details": details,
                "price_n": price_n,
                "price_r": price_r,
                "urls": urls
            }
            All_data.append(data)
```

完整版代码：待重构

最后的数据存放在一个文本中：

如图：



|      |   |
|------|---|
| 1    | Book1   |
| 2    | 书名： 自控力   |
| 3    | 作者： 凯利麦格尼格尔 著，王岑卉 译   |
| 4    | 评价人数： 189125条评论   |
| 5    | 简介：   |
| 6    | ★《自控力2》，“自控力”实践应用版  |
| 1329 |   |
| 1330 | 售价： ¥24.90  |
| 1331 | 定价： ¥39.80  |
| 1332 | 书籍链接： <a href="http://product.dangdang.com/23484713.html#ddclick?act=click&amp;pos=23484713_59_1_p8">http://product.dangdang.com/23484713.html#ddclick?act=click&amp;pos=23484713_59_1_p8</a> |
| 1333 |   |

当然：代码还可以继续重构,比如，图书介绍好些空白行如何处理；比如：先抓大再进行xpath等等之类的代码优化...

心理学这个图书栏有100页。也可以尝试100页如何抓取，存取。会不会出现问题。  
核心代码，try ...except都没写...（差评！）；不写注释（差评！）

---

## 6：参考及总结

01: [w3school：xpath教程](#)

02: [lxml文档](#)

03: [练习版代码](#)

爬取思路还是和之前的系列专栏一致，解析方法变了而已。

Github:[github](#)

关于本人：

国内小硕，半途出家的IT学习者。

兴趣领域：爬虫，数据科学

本人正在构建一个共同成长爬虫小型社群。有兴趣私信。

文档及代码托管在Github上。

---