xpath解析

XPath是一门在 XML 文档中查找信息的语言. XPath可用来在 XML 文档中对元素和属性进行遍历. 而我们熟知的HTML恰巧属于XML的一个子集. 所以完全可以用xpath去查找html中的内容.

首先, 先了解几个概念.

```
<book>
    <id>1</id>
    <name>野花遍地香</name>
    <price>1.23</price>
    <author>
        <nick>周大强</nick>
        <nick>周芷若</nick>
    </author>
</book>
```

在上述html中,

1. book, id, name, price....都被称为节点.
2. Id, name, price, author被称为book的子节点
3. book被称为id, name, price, author的父节点

4. id, name, price,author被称为同胞节点

OK~ 有了这些基础知识后, 我们就可以开始了解xpath的基本语法了
在python中想要使用xpath, 需要安装lxml模块.

```
1 pip install lxml
```

用法:

1. 将要解析的html内容构造出etree对象.
2. 使用etree对象的xpath()方法配合xpath表达式来完成对数据的提取

```
 1 from lxml import etree
 2
 3 html = """
 4 <book>
 5     <id>1</id>
 6     <name>野花遍地香</name>
 7     <price>1.23</price>
 8     <nick>臭豆腐</nick>
 9     <author>
10         <nick id="10086">周大强</nick>
11         <nick id="10010">周芷若</nick>
12         <nick class="joy">周杰伦</nick>
13         <nick class="jolin">蔡依林</nick>
```

```
14          <div>
15              <nick>惹了</nick>
16          </div>
17      </author>
18
19      <partner>
20          <nick id="ppc">胖胖陈</nick>
21          <nick id="ppbc">胖胖不陈</nick>
22      </partner>
23  </book>
24  """
25
26  et = etree.XML(html)
27  # 根据节点进行搜索
28  # result = et.xpath("/book")
29  # result = et.xpath("/book/id")  # /在开头表示文档最
    开始, /在中间表示儿子
30  # result = et.xpath("/book//nick")  # //表示后代
31  result = et.xpath("/book/*/nick")  # *表示通配符
32
33  print(result)
34
```

xpath如何提取属性信息. 我们上一段真实的HTML来给各位讲解一下

准备HTML:

```html
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8" />
    <title>Title</title>
</head>
<body>
    <ul>
        <li><a href="http://www.baidu.com">百度</a></li>
        <li><a href="http://www.google.com">谷歌</a></li>
        <li><a href="http://www.sogou.com">搜狗</a></li>
    </ul>
    <ol>
        <li><a href="feiji">飞机</a></li>
        <li><a href="dapao">大炮</a></li>
        <li><a href="huoche">火车</a></li>
    </ol>
    <div class="job">李嘉诚</div>
    <div class="common">胡辣汤</div>
</body>
</html>
```

xpath解析

```python
from lxml import etree

tree = etree.parse("1.html")
result = tree.xpath("/html/body/ul/li/a/@href")
print(result)

result = tree.xpath("/html/body/ul/li")
for li in result:
    print(li.xpath("./a/@href"))  # 局部解析

result = tree.xpath("//div[@class='job']/text()")
 # [@class='xxx']属性选取  text()获取文本
print(result)
```