Xpath语法

- /表示根节点
- a/text()取**a**下的文本，a//text()a下的所有文本
- a/@href**“@”符号取属性**
- .当前路径，..上一级路径
- //li选取任意位置/某元素任意位置的元素
- //input[@type="submit"]**对元素的属性进行限制**（如果@type则选中所有具有type属性的元素）
  - //a[text()="下一页"]选择文本为“下一页”的a标签
  - //a[1]选择第一个a标签（从1开始而不是从0开始，倒数是last(),last()-1,last()-3而不是-1,-2,-3），//a[position<3]选择前两个a标签，使用、、
  - //book[price<35]条件选择
- *和node()表示任意节点
  - //a/@*获取a标签下所有属性的值
- //a[last()]|//a[3]同时选择|

> Chrome插件 XPath Helper,需要注意爬虫爬到数据和element中的内容不一定一致。

lxml库

- 入门

```
import lxml.etree

html = etree.HTML(html_content)
html.xpath("//input[@type="submit"]")
```

> lxml可以自动修正html代码，但不一定正确，使用etree.tostring()观察

- 使用xpath()后u会得到一个列表，可能是**Element对象列表，可以再次进行xpath()**

```
ret1 = html.xpath("//li)
for i in ret1:
    title = i.xpath("./a/text()")[0] if len(i.xpath("./a/text()"))>0 else None #
列表可能为空，使用三元运算符
    url = i.xpath("./a/@href")[0] if len(i.xpath("./a/text()"))>0 else None
```

> 提取页面上数据的思路:先分组，遍历每一组，不会造成数据的错乱