Xpath解析

- xpath是什么

```
1. xpath原本是在可扩展标记语言中进行数据查询---语言
        可扩展标记语言：XML
2. 后期发展成为支持解析各种标记语言，的优选语言
        HTML
3. xpath不是标准库中的模块，需要第三方安装
        pip install lxml
4. xpath语法被嵌入到了lxml模块中
        lxml是一个包
        lxml下又一个模块：etree
        etree模块中的对象支持xpath语法


lxml：是一个第三方的解析库
xpath：是一种语言，一种语法，一种规范
```

- etree模块

```
etree模块中的函数
1. HTML(text):
        可以返回一个etree._Element对象（简称Element对象）
        text:是一个html/xml文件，可以使str，bytes
2. toString（ElementOBJ）：
        可以将Element对象中的标签进行修正，并以字符串形式返回
        结果是一个二进制字符串
3. Element（）：
        用于创建Element对象，通过配合其他函数可以构建xml对象，不是爬虫范畴
```

- lxml.etree._Element对象（简称Element对象）

```
1. Element对象的底层使用了多维思想
        Element 是一个序列：
                支持切片操作
                支持下标操作
        是一个自定义的序列：
                __len__()
                __getitem__()
                __setitem__()
                __delitem__()
        并且将list封装到了Element对象之内
2. Element对象的常用属性：
        1. tag：获得标签名（str）
        2. attrib：获得属性名（dict）
        3. text：获得当前节点的内容（str）
        4. tail：获得Element闭合之后的尾迹（str）
```

```
3. Element对象的常用方法：
        1. xpath（）
        使用xpath表达式选取元素的节点，返回值的是一个Element列表
        2. getchildren（）
            获得当前节点的子节点
            返回值的是一个Element列表
        3. getparent（）
            获得当前节点的父节点
        4. getprevious()
            获得当前节点前的节点
        4. getnext（）
            获得当前节点后的节点
```

xpath的使用

```
1. xpath（self,_path,*,namespaces,extensions,smart_strings,**_varibles）
        使用xpath表达式选取节点，返回一个Element列表

2. 节点：（node）
        七种节点
        元素，属性，文本，命名空间，处理指令，注释，文档
```

- xpath支持的路径表达式和内置方法

表达式	描述
node name---节点名	选取当前节点的所有同名子节点
/	从当前节点开始递归的选取子节点
//	选取所有符合要求的节点，不考虑位置
.	当前节点
..	选取父节点
@	选取属性
string（path）	获取path指定的节点和子节点的内容（是函数） 用法：xpath('string(path)') 如果有多个节点符合要求，只获取第一个节点的内容
text()	获取指定节点的内容（不包括子节点），如果有多个节点符合要求，则全部取出 用法：xpath（'path/text()'）---text()不是函数
[]	谓语 ：用于过滤（筛选）
*	通配符：匹配任意节点
node（）	返回指定节点的所有子节点和格式（包括换行符）

xpath中的谓语

1. 用于查找某个特定节点或者包含某个特定内容的节点，使用谓语进行筛选

表达式	说明
name node[index]	谓语中写入要获取的元素的下标，以获取指定下标的Element对象（下标从1开始）
last()	获取最后一个Element对象
position（ ）	指代节点的下标
starts-with(@属性，值)	筛选属性以指定属性名和指定属性值为开头的Element对象
contains（@属性，值）	筛选包含某属性，且属性值包含指定值得Element对象
not（ 函数）	结果

- xpath支持的运算符

```
| 选取多个节点，不应用于谓语
+，-，*，div 基本运算
== != > < >= <=布尔运算
and or not逻辑运算
```

- 总结

```
1. 使用xpath语法：
        from lxml import etree
        初始化html为Element对象
        E.xpath(规则)解析
2. 分清节点关系
3. xpath，是一种语言，熟练应用路径表达式，函数，谓语，运算符
4. Element对象的下标从1开始
```

- 作业：

1. 菜鸟教程100例 （标题，题目，分析）--（多个模板）
2. 猫眼电影-榜单-top100（翻页，电影名，主演，上映时间，评分）
3. 51job（*城市：北，杭，广，深， *职位关键字：爬虫，python，数据分析 ）
    - 职位名称
    - 公司名称
    - 工资（直接存字符串）
    - 招聘的基本要求（地区，经验，学历，人数等）
    - 职位信息
    - 任职资格
    - 练习方式
    - 公司信息
    - 公司规模

说明：所有信息要入库，数据库名称：以网站名称命名