第二周：网络爬虫之提取

2.1BeautifulSoup 库入门

1.使用 BeautifulSoup 的方式

from bs4 import BeautifulSoup

soup = BeautifulSoup(' data ', 'html.parser')

第一个参数是一个 html 格式的信息。

2. BeautifulSoup 的基本元素

BS 库是解析、遍历、维护"标签树"的功能库。例如：

soup = BeautifulSoup("<html>data</html>","html.parser")

soup = BeautifulSoup(open("D://demo.html"),"html.parser")

表 1.1 Beautiful Soup 库解析器

解析器	使用方法	条件
bs4 的 HTML 解析器	BeautifulSoup(mk,'html.parser')	按照 bs4 库
lxml 的 HTML 解析器	BeautifulSoup(mk,'lxml')	pip install lxml
lxml 的 XML 解析器	BeautifulSoup(mk,'xml')	pip install lxml
html5lib 的解析器	BeautifulSoup(mk,'html5lib')	pip install html5lib

表 1.2 Beautiful Soup 类的基本元素

基本元素	说明
Tag	标签，最基本的信息组织单元，分别用<>和</>表明开头和结尾
Name	标签的名字。格式：<tag>.name
Attributes	标签的属性，字典形式组织。格式<tag>.attrs
NavigableString	标签内非属性字符串，<></>中字符串。格式：<tag>.string
Comment	标签内字符串的注释部分，一种特殊的 Comment 类型

当 html 网页中存在多个相同标签时，只能返回第一个。例如输入：

print(soup.a)

因为该页面有多个 a 链接，所以只返回第一个，即

Basic Python

当我们输入

print(soup.a.parent.name)

我们可以获取到 a 的父亲的名字，也即

'p'

当我们输入

print(soup.a.parent.parent.name)

得到 p 标签的父亲是

'body'

当我们来获取 a 标签的属性的时候，我们输入

print(soup.a.attrs)

得到的结果是

{'href': 'http://www.icourse163.org/course/BIT-268001', 'class': ['py1'], 'id': 'link1'}

这是用一个字典的方式得到的反馈。

3.基于 bs4 库的 HTML 内容遍历方法

标签树的遍历方式主要有：上行遍历、下行遍历和平行遍历。

（1）下行遍历

表 1.3 标签树的下行遍历

基本元素	说明
.contents	子节点的列表，将<tag>所有儿子节点存入列表。
.children	子节点的迭代类型，与.contents 类似，用于循环遍历儿子节点。
.descendants	子孙节点的迭代类型，包含所有子孙节点，用于循环遍历。

表 1.4 标签树的平行遍历

基本元素	说明
.next_sibling	返回按照 HTML 文本顺序的下一个平行节点标签
.previous_sibling	返回按照 HTML 文本顺序的上一个平行节点标签
.next_siblings	返回按照 HTML 文本顺序的后续所有平行节点标签（for 循环中）
.previous_siblings	返回按照 HTML 文本顺序的前续所有平行节点标签（for 循环中）

注意：所有平行遍历发生在同一个父节点下的各节点之间。

表 1.5 标签树的上行遍历

基本元素	说明
.parent	返回当前节点的父亲节点
.parents	返回当前节点的所有上面的节点

4.基于 bs4 的 HTML 格式化和编码

利用 prettify()。例如，我们输入

print(soup.prettify())

打印出来的网页代码就比较明了。

bs4 库将所有网页等内容转换成了 **UTF-8 码**。

2.2 信息组织与提取方法

1.信息标记的三种形式

信息标记的作用：

（1）可以形成信息组织结构，增加信息维度。

（2）可用于通信、存储或展示。

（3）标记的结构与信息一样具有重要价值。

国际通用信息标记三种形式有：XML（类似于 HTML）、JSON（有类型键值对）、YAML（无类型键值对）。

2.三种信息标记形式的比较

一段信息，用 XML 表示：

<person>

<firstname>Tian</firstname>

```
    <lastname>Song</lastname>
    <address>
        <streetAddr>中关村南大街 5 号</streetAddr>
        <city>北京市</city>
        <zipcode>100081</zipcode>
    </address>
    <prof>Computer System</prof><prof>Secuity</prof>
</person>
```

JSON 版本：

```
{
    "firstname":"Tian",
    "lastname":"Song",
    "address":{
                    "streetAddr":"中关村南大街 5 号",
                    "City":"北京市",
                    "zipcode":"100081"
                },
    "prof":["Computer System","Secuity"]
}
```

YAML 实例：

```
firstname:Tian
lastname:Song
address:
    streetAddr:中关村南大街 5 号
    City:北京市
    zipcode:100081
prof:
-Computer System
-Secuity
```

下面对三种类型进行比较：

（1）XML 是最早的通用信息标记语言，可扩展性好，但比较繁琐。在 Internet 上使用。

（2）JSON 信息有类型，适合程序处理（js），较 XML 简洁。移动应用云端和节点的信息通信，无注释。

（3）YAML 信息无类型，文本信息比例最高，可读性好。各类系统的配置文件，有注释，易懂。

3.信息提取的一般方法

方法一：W 按照解析信息的标记形式，再提取关键信息。需要标记解析器，优点是信息解析准确，缺点是提取过程繁琐，速度慢。（XML/JSON/YAML 均适用）

方法二：无视标记形式，直接搜索关键信息。优点是提取过程简洁，速度较快；缺点是提取结果准确性与信息内容相关。

融合方法：结合形式解析与搜索方法，提取关键信息。需要标记解析器及文本查找函数。

例如输入：

```
for link in soup.find_all('a'):
    print(link.get('href'))
```

结果为

http://www.icourse163.org/course/BIT-268001

http://www.icourse163.org/course/BIT-1001870001

4.基于 bs4 库的 HTML 内容查找方法

（1）方法：<>.find_all(name,attrs,recursive,string,**kwargs)：返回一个列表类型，存储查找的结果。

①.name：对标签名称的检索字符串。

如果参数列表我们使用 True，将返回所有的标签名称。

如果我们想显示 b 开头标签的内容，可以使用正则表达式库来实现，代码如下：

```
for tag in soup.find_all(re.compile('b')):
    print(tag.name)
```

这样返回的结果是：

```
body
b
```

②.attrs：对标签属性值的检索字符串，可标注属性检索。

③recursive：是否对子孙全部检索，默认为 True。

④string：<>…</>中字符串渔区的检索字符串。<tag>(..)等价于<tag>.find_all()，这是一种简写。

⑤**kwargs：7 种扩展方法。见表 2.1

表 2.1 扩展方法

方法	说明
<>.find()	搜索且只返回一个结果，字符串类型，同.find_all()参数
<>.find_parents()	在先辈节点中搜索，返回列表类型，同.find_all()参数
<>.find_parent()	在先辈节点中返回一个结果，字符串类型，同.find()参数
<>.find_next_siblings()	在后续平行节点中搜索，返回列表类型，同.find_all()参数
<>.find_next_sibling()	在后续平行节点中返回一个结果，字符串类型，同.find()参数
<>.find_previous_siblings()	在前序平行节点中搜索，返回列表类型，同.find_all()参数
<>.find_previous_sibling()	在前序平行节点中返回一个结果，字符串类型，同.find()参数

2.3 实例 1：中国大学排名爬虫

1.功能描述

输入：大学排名的 URL 链接

输出：大学排名信息的屏幕输出（排名、大学名称、总分）

技术路线：requests-bs4

定向爬虫：仅对输入 URL 进行爬取，不扩展爬取。

2.程序结构设计

步骤 1：从网络上获取大学排名网页内容。getHTMLText()

步骤 2：提取网页内容中信息到合适的数据结构。fillUnivList()

步骤 3：利用数据结构展示输出结果。PrintUnivList()

3.初始代码

```
import requests
```

```
from bs4 import BeautifulSoup
import bs4

def getHTMLText(url):
    try:
        r = requests.get(url ,timeout =30)
        r.raise_for_status()
        r.encoding = r.apparent_encoding
        return r.text
    except:
        return ""

def fillUnivList(ulist, html):
    soup = BeautifulSoup(html, "html.parser")
    for tr in soup.find('tbody').children:
        if isinstance(tr, bs4.element.Tag):
            tds = tr('td')
            ulist.append([tds[0].string,tds[1].string,tds[2].string])


def printUnivList(ulist, num):
    print("{:^10}\t{:^6}\t{:^10}".format("排名","学校名称","总分"))
    for i in range(num):
        u = ulist[i]
        print("{:^10}\t{:^6}\t{:^10}".format(u[0],u[1],u[2]))
    print("Suc"+ str(num))

def main():
    uinfo = []
    url = 'http://www.zuihaodaxue.com/zuihaodaxuepaiming2018.html'
    html = getHTMLText(url)
    fillUnivList(uinfo, html)
    printUnivList(uinfo, 55)#20 所大学

main()
```

结果：

排名	学校名称	总分
1	清华大学	北京
2	北京大学	北京
3	浙江大学	浙江
4	上海交通大学	上海
5	复旦大学	上海
6	中国科学技术大学	安徽
7	南京大学	江苏

8	华中科技大学	湖北	
9	中山大学	广东	
10	哈尔滨工业大学	黑龙江	
11	同济大学	上海	
12	武汉大学	湖北	
13	东南大学	江苏	
14	西安交通大学	陕西	
15	北京航空航天大学		北京
16	南开大学	天津	
17	四川大学	四川	
18	天津大学	天津	
19	华南理工大学	广东	
20	北京师范大学	北京	
21	北京理工大学	北京	
22	厦门大学	福建	
23	吉林大学	吉林	
24	山东大学	山东	
25	大连理工大学	辽宁	
26	中南大学	湖南	
27	苏州大学	江苏	
28	对外经济贸易大学		北京
29	西北工业大学	陕西	
30	中国人民大学	北京	
31	湖南大学	湖南	
32	华东师范大学	上海	
33	电子科技大学	四川	
34	华东理工大学	上海	
35	重庆大学	重庆	
35	南京航空航天大学		江苏
37	北京科技大学	北京	
37	南京理工大学	江苏	
39	上海财经大学	上海	
40	中国农业大学	北京	
41	上海大学	上海	
42	东北大学	辽宁	
43	华中师范大学	湖北	
43	南方科技大学	广东	
45	北京交通大学	北京	
46	首都医科大学	北京	
47	武汉理工大学	湖北	
48	北京化工大学	北京	
48	北京邮电大学	北京	
48	东华大学	上海	
51	北京外国语大学	北京	

52 天津医科大学 天津
52 中央财经大学 北京
54 西安电子科技大学 陕西
55 南京医科大学 江苏

我们发现，总分下面是省份，不符合要求。接下来对代码进行优化。

4.优化代码

在系统中，如果字符不足，默认采用西文填充，为了解决这个问题，我们可以采用中文字符的空格填充，格式为

char(12288)

优化完代码如下：

```
import requests
from bs4 import BeautifulSoup
import bs4

def getHTMLText(url):
    try:
        r = requests.get(url ,timeout =30)
        r.raise_for_status()
        r.encoding = r.apparent_encoding
        return r.text
    except:
        return ""

def fillUnivList(ulist, html):
    soup = BeautifulSoup(html, "html.parser")
    for tr in soup.find('tbody').children:
        if isinstance(tr, bs4.element.Tag):
            tds = tr('td')
            ulist.append([tds[0].string,tds[1].string,tds[3].string])


def printUnivList(ulist, num):
    textName =   'UniversityList.txt'
    title = '2018 年中国大学排行榜\n'
    f = open(textName, 'w+')
    tplt = "{0:^10}\t{1:{3}^10}\t{2:^10}"
    print(tplt.format("排名","学校名称","总分",chr(12288)))
    for i in range(num):
        u = ulist[i]
        print(tplt.format(u[0],u[1],u[2],chr(12288)))
    f.writelines('{}'.format(title))
    f.writelines(tplt.format("排名","学校名称","总分",chr(12288))+'\n')
    for i in range(num):
        u = ulist[i]
```

```
            f.writelines(tplt.format(u[0],u[1],u[2],chr(12288))+'\n')
    f.close()

def main():
    uinfo = []
    url = 'http://www.zuihaodaxue.com/zuihaodaxuepaiming2018.html'
    html = getHTMLText(url)
    fillUnivList(uinfo, html)
    printUnivList(uinfo, 55)#55 所大学
    print("2018 年中国大学排行榜已经生成，并已经导出为 UniversityList.txt 文件存储在您的文件夹中。")
    print("感谢您使用本程序！")

main()结果:
```

排名	学校名称	总分
1	清华大学	95.3
2	北京大学	78.6
3	浙江大学	73.9
4	上海交通大学	73.1
5	复旦大学	66.0
6	中国科学技术大学	61.9
7	南京大学	59.8
8	华中科技大学	59.1
9	中山大学	58.6
10	哈尔滨工业大学	57.4
11	同济大学	56.4
12	武汉大学	55.5
13	东南大学	55.3
14	西安交通大学	54.2
15	北京航空航天大学	54.0
16	南开大学	53.9
17	四川大学	53.3
18	天津大学	52.4
19	华南理工大学	51.8
20	北京师范大学	51.7
21	北京理工大学	51.1
22	厦门大学	50.9
23	吉林大学	50.2
24	山东大学	50.0
25	大连理工大学	49.7
26	中南大学	49.5
27	苏州大学	48.8
28	对外经济贸易大学	47.7
29	西北工业大学	47.6

30	中国人民大学	47.5
31	湖南大学	47.4
32	华东师范大学	46.5
33	电子科技大学	46.4
34	华东理工大学	45.5
35	重庆大学	45.2
35	南京航空航天大学	45.2
37	北京科技大学	44.5
37	南京理工大学	44.5
39	上海财经大学	44.3
40	中国农业大学	43.7
41	上海大学	43.6
42	东北大学	43.5
43	华中师范大学	43.3
43	南方科技大学	43.3
45	北京交通大学	43.0
46	首都医科大学	42.9
47	武汉理工大学	42.8
48	北京化工大学	42.4
48	北京邮电大学	42.4
48	东华大学	42.4
51	北京外国语大学	42.1
52	天津医科大学	42.0
52	中央财经大学	42.0
54	西安电子科技大学	41.9
55	南京医科大学	41.7

2018 年中国大学排行榜已经生成, 并已经导出为 UniversityList.txt 文件存储在您的文件夹中。感谢您使用本程序!

另外生成了相应的 txt 文档。