第3课 我的第一个小爬虫

1、BeautifulSoup 是什么

解析和提取网页中的数据：

（1）解析数据：把服务器返回来的 HTML 源代码翻译为我们能理解的方式；

（2）提取数据：把我们需要的数据从众多数据中挑选出来。

2、BeautifulSoup 怎么用

2-1、BeautifulSoup 安装

win：pip install BeautifulSoup4；

Mac：pip3 install BeautifulSoup4。

2-2、BeautifulSoup 解析数据


```
bs对象 = BeautifulSoup（要解析的文本,'解析器'）
```

括号中，要输入两个参数：

①、第 0 个参数是要被解析的文本（必须是字符串）

②、第 1 个参数用来标识解析器，我们要用的是一个Python内置库：html.parser。（不是唯一的解析器）

```
import requests
from bs4 import BeautifulSoup
#引入BS库
res =
requests.get('https://xiaoke.kaikeba.com/example/canteen/index.html')
html = res.text
soup = BeautifulSoup(html,'html.parser') #把网页解析为BeautifulSoup对象
```

2-3、BeautifulSoup 提取数据

2-3-1、find() 与 find_all()

find() 与 find_all() 是 BeautifulSoup 对象的两个方法，它们可以匹配 html 的标签和属性，把 BeautifulSoup 对象里符合要求的数据都提取出来：

①、find()只提取首个满足要求的数据；

```
import requests
from bs4 import BeautifulSoup
url = 'https://localprod.pandateacher.com/python-manuscript/crawler-html/spder-men0.0.html'
res = requests.get (url)
print(res.status_code)
soup = BeautifulSoup(res.text,'html.parser')
item = soup.find('div') #使用find()方法提取首个<div>元素，并放到变量item里。
print(type(item)) #打印item的数据类型
print(item) #打印item
#结果：
200
<class 'bs4.element.Tag'>
<div>大家好，我是一个块</div>
```

②、find_all()提取出的是所有满足要求的数据。

```
import requests
from bs4 import BeautifulSoup
url = 'https://localprod.pandateacher.com/python-manuscript/crawler-html/spder-men0.0.html'
res = requests.get (url)
print(res.status_code)
soup = BeautifulSoup(res.text,'html.parser')
item = soup.find_all('div') #使用find()方法提取首个<div>元素，并放到变量item里。
```

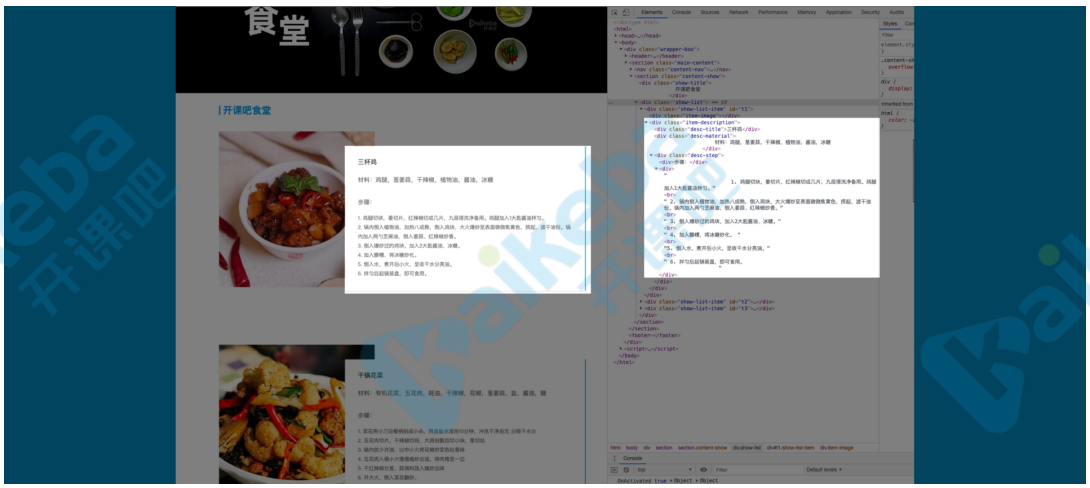
```
print(type(item)) #打印item的数据类型
print(item) #打印item
#结果：
200
<class 'bs4.element.ResultSet'>
[<div>大家好，我是一个块</div>, <div>我也是一个块</div>, <div>我还是一个块
</div>]
```

注意：

find() 或 find_all() 括号中的参数：标签和属性可以任选其一，也可以两个一起使用，这取决于我们要在网页中提取的内容。

（1）中括号里的class_，这里有一个下划线，是为了和python语法中的类 class区分，避免程序冲突。当然，除了用class属性去匹配，还可以使用其它属性，比如style属性等；

（2）只用其中一个参数就可以准确定位的话，就只用一个参数检索。如果需要标签和属性同时满足的情况下才能准确定位到我们想找的内容，那就两个参数一起使用。




```
import requests
from bs4 import BeautifulSoup
res =
requests.get('https://xiaoke.kaikeba.com/example/canteen/index.html')
html = res.text
soup = BeautifulSoup(html,'html.parser')
items = soup.find_all(class_='show-list-item')
print("想找的菜的信息都在这里了：")
for item in items:
    print(item) # 打印item
```

2-3-2、Tag 对象

Tag对象的三种常用属性与方法

属性/方法	作用
Tag.find()和Tag.find_all()	提取Tag中的Tag
Tag.text	提取Tag中的文字
Tag['属性名']	输入参数：属性名，可以提取Tag中这个属性的值

```
import requests
from bs4 import BeautifulSoup
res =
requests.get('https://xiaoke.kaikeba.com/example/canteen/index.html')
html = res.text
soup = BeautifulSoup(html,'html.parser')
items = soup.find_all(class_='show-list-item')
for item in items:
    title = item.find(class_='desc-title') # 在列表中的每个元素里，匹配属性class_='title'提取出数据
    material = item.find(class_='desc-material') #在列表中的每个元素里，匹配属性class_='desc-material'提取出数据
    step = item.find(class_='desc-step') #在列表中的每个元素里，匹配属性class_='desc-step'提取出数据
    print(title.text,'\n',material.text,'\n',step.text)

```

3、对象的变化过程

对象操作：Response对象——字符串——BS对象：

①、一条是BS对象——Tag对象；

②、另一条是BS对象——列表——Tag对象。