数据分析及实践

实验二

PB21000024

王一鸣

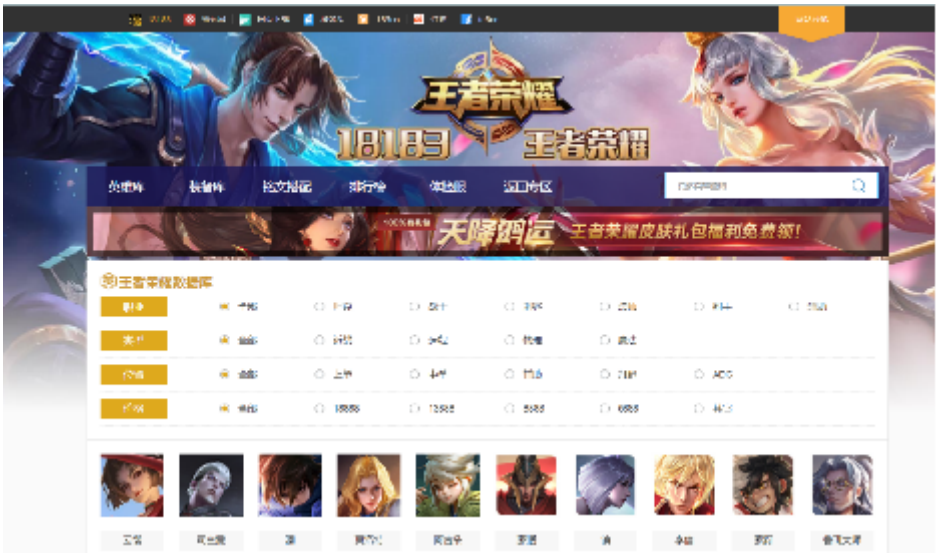
实验目的

练习获取数据的能力

实验内容

编写爬虫爬取特定网站内容并进行解析

王者英雄数据网站 **http://db.18183.com/wzry/** 英雄信息爬取



实验过程

使用python

- 首先获取对应网站的代码

```
url = 'http://db.18183.com/wzry/'
r=requests.get(url)
```

调用request库对对应网站发起请求，爬取网站代码

此时r存取了网站代码，现在我们要获取对我们有用的数据

- 解析网站代码

```
soup=BeautifulSoup(r.text,'lxml')
#从总页面得到各个分页面信息
heros=soup.find_all('li',class_='mod-iconitem')
```

```
        <!--  -->
<div class="section hero-result-box mod-bg clearfix">
    <ul class="mod-iconlist">
                        <li class="mod-iconitem" data-category="fighter" data-id="16364" data-price="13888" data-type="near,attack" data-pos
    style=""
    >
                <a href="/wzry/hero/16364.html">
                    <img class="mod-iconitem-img" src="//js.18183.com/uploads/pc/yxzjol/sjk/loading.gif" data-original="https://img.18183.com/up
                    <p class="mod-iconitem-tit">云缨</p>
                </a>
            </li>
                        <li class="mod-iconitem" data-category="fighter" data-id="16354" data-price="13888" data-type="near,magic" data-posi
    style=""
    >
                <a href="/wzry/hero/16354.html">
                    <img class="mod-iconitem-img" src="//js.18183.com/uploads/pc/yxzjol/sjk/loading.gif" data-original="https://img.18183.com/up
                    <p class="mod-iconitem-tit">司空震</p>
                </a>
            </li>
                        <li class="mod-iconitem" data-category="assassin" data-id="16344" data-price="13888" data-type="near,attack" data-po
    style=""
    >
                <a href="/wzry/hero/16344.html">
                    <img class="mod-iconitem-img" src="//js.18183.com/uploads/pc/yxzjol/sjk/loading.gif" data-original="https://img.18183.com/up
                    <p class="mod-iconitem-tit">澜</p>
                </a>
            </li>
                        <li class="mod-iconitem" data-category="fighter" data-id="16327" data-price="0" data-type="near,attack" data-position
    style=""
    >
                <a href="/wzry/hero/16327.html">
                    <img class="mod-iconitem-img" src="//js.18183.com/uploads/pc/yxzjol/sjk/loading.gif" data-original="https://img.18183.com/up
                    <p class="mod-iconitem-tit">夏洛特</p>
                </a>
            </li>
```

通过观察网页的代码结构我们发现英雄子页面链接存放在标识符为 *<li>*所在处，并且具有 *class=“mod-iconitem”* 属性，于是可以调用bs4进行查找，得到一个包含各英雄子页面的列表

- 爬取子网站的信息

  通过观察我们发现子网站地址即为主站地址下的 */hero/id.html*　id即为英雄id，于是可以通过拼接得到子网站链接

```
url_=url+'hero/'+hero['data-id']+'.html'#访问分页面
```

- 解析子网站

  同样通过观察子网站代码结构，发现标识符并进行查找

  同样还有基础属性，不予赘述

```
                    </div>
                </div>
                <!-- 基础属性 -->
                <div class="otherinfo-item">
                    <div class="otherinfo-datapanel">
                        <ul>
                            <li><p>最大生命：3237</p></li>
                            <li><p>最大法力：430</p></li>
                            <li><p>物理攻击：175</p></li>
                            <li><p>法术攻击：0</p></li>
                            <li><p>物理防御：101</p></li>
                            <li><p>物理减伤率：14.4%%</p></li>
                            <li><p>法术防御：  50</p></li>
                            <li><p>法术减伤率：  7.6%%</p></li>
                            <li><p>移速：  380</p></li>
                            <li><p>物理护甲穿透：0</p></li>
                            <li><p>法术护甲穿透：0</p></li>
                            <li><p>攻速加成：0</p></li>
                            <li><p>暴击几率：0</p></li>
                            <li><p>暴击效果：200%%</p></li>
                            <li><p>物理吸血：0</p></li>
                            <li><p>法术吸血：0</p></li>
                            <li><p>冷却缩减：0%</p></li>
                            <li><p>攻击范围：近程</p></li>
                            <li><p>韧性：0</p></li>
                            <li><p>生命回复：49</p></li>
                            <li><p>法力回复：15</p></li>
                        </ul>
                    </div>
                </div>
            </div>
        </div>
    </div>
```

提取 *class=“otherinfo-datapanel”* ,以换行符为界就能分开每个属性，然后进行字符串处理制成字典

```
soup=BeautifulSoup(r.text,'lxml')
out=soup.find('div', class_='otherinfo-datapanel')#提取英雄信息列表
pattern=re.compile('\n+')
```

```
except Exception:
    continue
else:
    temp={}
    temp['id']=hero['data-id']
    temp['name']=hero.p.text
    temp['生存能力']=base_attri[0]
    temp['攻击伤害']=base_attri[1]
    temp['技能效果']=base_attri[2]
    temp['上手难度']=base_attri[3]
    for i in range(len(out)):
        if(out[i]==''):
            continue
        out[i]=out[i].split(':')
        out[i][1]=re.sub(re.compile('\s+'),'',out[i][1])
        out[i][1]=re.sub(re.compile('%+'),'%',out[i][1])
        temp[out[i][0]]=out[i][1]
    sample.append(temp)#生成字典
    #print(temp)
```

- 写入文件

```
#文件写入
file=open('result.json','w',encoding='utf8')
for item in sample:

    file.write(json.dumps(item,ensure_ascii=False))
    file.write('\n')
file.close()
```


"生命回复": "42",
"法力回复": "17"
}
{
"id": "9512",
"name": "扁鹊",
"生存能力": "4",
"攻击伤害": "7",
"技能效果": "3",
"上手难度": "4",
"最大生命": "3205",
"最大法力": "490",
"物理攻击": "168",
"法术攻击": "0",
"物理防御": "87",
"物理减伤率": "12.6%",
"法术防御": "50",
"法术减伤率": "7.6%",
"移速": "350",
"物理护甲穿透": "0",
"法术护甲穿透": "0",
"攻速加成": "0",

以上即为本次实验过程