PyShare D2

requests

In [ ]:
```
import requests
```

In [ ]:
```
re = requests.get("http://maoyan.com/board/4")
print(re.text)
```

- 没有添加请求头时,无法访问网页,尝试添加请求头headers

In [ ]:
```
headers = {
    "User-Agent":
    "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/68.0.3440.84 Safari/537.36"
}
```

- 添加请求头之后就可以正常访问网页了

In [ ]:
```
re = requests.get("http://maoyan.com/board/4", headers=headers)
```

In [ ]:
```
print(re.headers)
```

- 用于保持会话登陆信息的,本章节不涉及到这个内容,以后再讨论,

In [ ]:
```
print(re.cookies)
```

- 获取网络资源,比如下载图片

In [ ]:
```
import requests
url = "https://pic4.zhimg.com/v2-fb00cf1a31e087e2564f563f203ee098_xl.jpg"
re = requests.get(url, headers=headers)
with open("a.jpg", "wb") as jpg:
    jpg.write(re.content)
```

pyquery

In [ ]:
```
html = '''
<div id="container">
    <ul class="list">
         <li class="item-0" img="aaa">first item</li>
         <li class="item-1"><a href="link2.html">second item</a></li>
         <li class="item-0 active"><a href="link3.html"><span class="bold">third item</span></a></li>
         <li class="item-1 active2"><a href="link4.html">fourth item</a></li>
         <li class="item-0"><a href="link5.html">fifth item</a></li>
     </ul>
 </div>
'''
```

- div ul li a 这些都是标签
- id在CSS选择器中都是用#+名字表示,例如 #container
- 带class属性名的 都用 .+名字表示,例如 .item-0
- 有些标签带有两个class值,例如class="item-0 active,中间带有空格,所以是两个属性值,用.item-0.active表示
- 标签直接用标签名,例如li

In [ ]:

```
from pyquery import PyQuery as pq
doc = pq(html)
pagecode = doc('#container')
print(pagecode)
```

In [ ]:

```
html = '''
<div id="container">
    <ul class="list">
         <li class="item-0">first item</li>
         <li class="item-1"><a href="link2.html">second item</a></li>
         <li class="item-0 active"><a href="link3.html"><span class="bol
d">third item</span></a></li>
         <li class="item-1 active2"><a href="link4.html">fourth item</a>
</li>
         <li class="item-0"><a href="dlink5.html">fifth item</a></li>
     </ul>
 </div>
'''
from pyquery import PyQuery as pq
doc = pq(html)
pagecode = doc.find("#container")
print(pagecode)
```

In [ ]:

```
html = '''
<div id="container">
    <ul class="list">
         <li class="item-0">first item</li>
         <li class="item-1"><a href="link2.html">second item</a></li>
         <li class="item-0 active"><a href="link3.html"><span class="bol
d">third item</span></a></li>
         <li class="item-1 active2"><a href="link4.html">fourth item</a>
</li>
         <li class="item-0"><a href="dlink5.html">fifth item</a></li>
     </ul>
 </div>
'''
from pyquery import PyQuery as pq
doc = pq(html)
pagecode = doc("#container .list")
print(pagecode)
```

In [ ]:

```
html = '''
<div id="container">
    <ul class="list">
         <li class="item-0">first item</li>
         <li class="item-1"><a href="link2.html">second item</a></li>
         <li class="item-0 active"><a href="link3.html"><span class="bold">third item</span></a></li>
         <li class="item-1 active2"><a href="link4.html">fourth item</a></li>
         <li class="item-0"><a href="dlink5.html">fifth item</a></li>
     </ul>
 </div>
'''
from pyquery import PyQuery as pq
doc = pq(html)
pagecode = doc('#container .list .item-0')
print(pagecode)
```

In [ ]:

```
html = '''
<div id="container">
    <ul class="list">
         <li class="item-0">first item</li>
         <li class="item-1"><a href="link2.html">second item</a></li>
         <li class="item-0 active"><a href="link3.html"><span class="bold">third item</span></a></li>
         <li class="item-1 active2"><a href="link4.html">fourth item</a></li>
         <li class="item-0"><a href="dlink5.html">fifth item</a></li>
     </ul>
 </div>
'''
from pyquery import PyQuery as pq
doc = pq(html)
pagecode = doc('#container .list .item-0.active')
print(pagecode)
```

In [ ]:

```
html = '''
<div id="container">
    <ul class="list">
         <li class="item-0">first item</li>
         <li class="item-1"><a href="link2.html">second item</a></li>
         <li class="item-0 active"><a href="link3.html"><span class="bold">third item</span></a></li>
         <li class="item-1 active2"><a href="link4.html">fourth item</a></li>
         <li class="item-0"><a href="dlink5.html">fifth item</a></li>
     </ul>
 </div>
'''
from pyquery import PyQuery as pq
doc = pq(html)
pagecode = doc('#container .list a')
print(pagecode)
```

- 一些巧妙的特殊搜索法

- 匹配某个属性值以***开头的元素

In [ ]:

```
html = '''
<div id="container">
    <ul class="list">
         <li class="item-0">first item</li>
         <li class="item-1"><a href="link2.html">second item</a></li>
```

```
        <li class="item-0 active"><a href="link3.html"><span class="bol
d">third item</span></a></li>
        <li class="item-1 active2"><a href="link4.html">fourth item</a>
</li>
        <li class="item-0"><a href="dlink5.html">fifth item</a></li>
    </ul>
 </div>
'''
from pyquery import PyQuery as pq
doc = pq(html)
pagecode = doc("#container .list a[href^='d']")
print(pagecode)
```

- 匹配某个属性值以***结尾的元素

In [ ]:

```
html = '''
<div id="container">
    <ul class="list">
        <li class="item-0">first item</li>
        <li class="item-1"><a href="link2.html">second item</a></li>
        <li class="item-0 active"><a href="link3.html"><span class="bol
d">third item</span></a></li>
        <li class="item-1 active2"><a href="link4.html">fourth item</a>
</li>
        <li class="item-0"><a href="link5.html">fifth item</a></li>
    </ul>
 </div>
'''
from pyquery import PyQuery as pq
doc = pq(html)
pagecode = doc("#container .list a[href$='5.html']")
print(pagecode)
```

In [ ]:

```
html = '''
<div id="container">
    <ul class="list">
        <li class="item-0">first item</li>
        <li class="item-1"><a href="linoo2.html">second item</a></li>
        <li class="item-0 active"><a href="link3.html"><span class="bol
d">third item</span></a></li>
        <li class="item-1 active2"><a href="link4.html">fourth item</a>
</li>
        <li class="item-0"><a href="linook5.html">fifth item</a></li>
    </ul>
 </div>
'''
from pyquery import PyQuery as pq
doc = pq(html)
pagecode = doc("#container .list a[href*='oo']")
print(pagecode)
```

- 猫眼电影排行榜

In [ ]:

```
from pyquery import PyQuery as pq
import requests

headers = {
    "User-Agent":
```

```
    "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML
, like Gecko) Chrome/68.0.3440.84 Safari/537.36",
}


def urlw(indext):
    return f'http://maoyan.com/board/6?offset={str(indext)}'


for i in range(1, 5):
    html = requests.get(urlw(i), headers=headers)
    do = pq(html.text)
    aa = do("#app .board-item-content").items()
    for i in aa:
        print(i(".name a").attr("title"))
with open('a') as target:
    pass
```

- 豆瓣电影TOP250

In [ ]:

```
from pyquery import PyQuery as pq
import requests

headers = {
    "User-Agent":
    "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML
, like Gecko) Chrome/68.0.3440.84 Safari/537.36",
}

url = "https://movie.douban.com/top250"


def urlw(indext):
    return f'https://movie.douban.com/top250?start={str(indext)}'


for i in range(1, 100, 25):
    html = requests.get(urlw(i), headers=headers)
    do = pq(html.text)
    aa = do(".grid_view .item").items()
    for i in aa:
        print(i(".title").text())
```