python爬取豆瓣高分电影

1.获取网页源代码

In [35]:
```
import requests
from bs4 import BeautifulSoup
import xlwt
headers = {
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) "
                  "Chrome/83.0.4103.116 Safari/537.36 "
}
def request_douban(url):
    try:
        response = requests.get(url=url, headers=headers)
        #print(response)
        if response.status_code == 200:
            return response.text
    except requests.RequestException:
        return None
```

2.解析各项信息并写入execl文件

In [36]:
```
book = xlwt.Workbook(encoding='utf-8', style_compression=0)
sheet = book.add_sheet('豆瓣电影Top250', cell_overwrite_ok=True)
sheet.write(0, 0, '名称')
sheet.write(0, 1, '图片')
sheet.write(0, 2, '排名')
sheet.write(0, 3, '评分')
sheet.write(0, 4, '作者')
n = 1


def save_to_excel(soup):
    list = soup.find(class_='grid_view').find_all('li')

    for item in list:
        item_name = item.find(class_='title').string    #由于class在python里是一个关键字，所以在这里class后面需要加一个下划线
        item_img = item.find('a').find('img').get('src')
        item_index = item.find(class_='').string
        item_score = item.find(class_='rating_num').string
        item_author = item.find('p').text
        if (item.find(class_='inq') != None):
            item_intr = item.find(class_='inq').string

        # print('爬取电影：' + item_index + ' | ' + item_name +' | ' + item_img +' | ' + item_score +' | ' + item_author +' | ' + item_intr )
        print('爬取电影：' + item_index + ' | ' + item_name + ' | ' + item_score + '|'+item_author)

        global n

        sheet.write(n, 0, item_name)
        sheet.write(n, 1, item_img)
        sheet.write(n, 2, item_index)
        sheet.write(n, 3, item_score)
        sheet.write(n, 4, item_author)
        n = n + 1
def main(page):
    url = 'https://movie.douban.com/top250?start=' + str(page * 25) + '&filter='
    html = request_douban(url)
    #print(html)
    soup = BeautifulSoup(html, 'lxml')
    save_to_excel(soup)
```

In [37]:
```
url = 'https://movie.douban.com/top250?start=0&filter='
html = request_douban(url)
print(html)
```

```
<a href="https://movie.douban.com/subject/1292052/" class="">
    <span class="title">肖申克的救赎</span>
            <span class="title"> / The Shawshank Redemption</span>
        <span class="other"> / 月黑高飞(港)  /  刺激1995(台)</span>
</a>


        <span class="playable">[可播放]</span>
</div>
<div class="bd">
    <p class="">
        导演: 弗兰克·德拉邦特 Frank Darabont   主演: 蒂姆·罗宾斯 Tim Robbins /...<br>
        1994 / 美国 / 犯罪 剧情
    </p>


    <div class="star">
            <span class="rating5-t"></span>
            <span class="rating_num" property="v:average">9.7</span>
            <span property="v:best" content="10.0"></span>
```

In [38]:
```
if __name__ == '__main__':
    for i in range(0, 10):
        main(i)

book.save(u'豆瓣最受欢迎的250部电影.xls')
```

```
爬取电影：1 | 肖申克的救赎 | 9.7|
                        导演: 弗兰克·德拉邦特 Frank Darabont      主演: 蒂姆·罗宾斯 Tim Robbins /...
                        1994 / 美国 / 犯罪 剧情

爬取电影：2 | 霸王别姬 | 9.6|
                        导演: 陈凯歌 Kaige Chen      主演: 张国荣 Leslie Cheung / 张丰毅 Fengyi Zha...
                        1993 / 中国大陆 中国香港 / 剧情 爱情 同性

爬取电影：3 | 阿甘正传 | 9.5|
                        导演: 罗伯特·泽米吉斯 Robert Zemeckis      主演: 汤姆·汉克斯 Tom Hanks / ...
                        1994 / 美国 / 剧情 爱情

爬取电影：4 | 这个杀手不太冷 | 9.4|
                        导演: 吕克·贝松 Luc Besson      主演: 让·雷诺 Jean Reno / 娜塔莉·波特曼 ...
                        1994 / 法国 美国 / 剧情 动作 犯罪

爬取电影：5 | 泰坦尼克号 | 9.4|
                        导演: 詹姆斯·卡梅隆 James Cameron      主演: 莱昂纳多·迪卡普里奥 Leonardo...
                        1997 / 美国 墨西哥 澳大利亚 加拿大 / 剧情 爱情 灾难
```

3.正则表达式的使用

In [39]:
```
#src="https://img2.doubanio.com/view/photo/s_ratio_poster/public/p480747492.jpg"
import re
patten = re.compile(r'src="(.*)" class')
imgs = re.findall(patten, html)

print(imgs)
```

```
['https://img2.doubanio.com/view/photo/s_ratio_poster/public/p480747492.jpg', 'https://img3.doubanio.com/view/photo/s_ratio_poster/public/p2561716440.jp
2.doubanio.com/view/photo/s_ratio_poster/public/p2372307693.jpg', 'https://img3.doubanio.com/view/photo/s_ratio_poster/public/p511118051.jpg', 'https://
m/view/photo/s_ratio_poster/public/p457760035.jpg', 'https://img2.doubanio.com/view/photo/s_ratio_poster/public/p2578474613.jpg', 'https://img1.doubanio.
_ratio_poster/public/p2557573348.jpg', 'https://img2.doubanio.com/view/photo/s_ratio_poster/public/p492406163.jpg', 'https://img2.doubanio.com/view/phot
public/p2616355133.jpg', 'https://img1.doubanio.com/view/photo/s_ratio_poster/public/p524964039.jpg', 'https://img1.doubanio.com/view/photo/s_ratio_post
8097.jpg', 'https://img2.doubanio.com/view/photo/s_ratio_poster/public/p479682972.jpg', 'https://img9.doubanio.com/view/photo/s_ratio_poster/public/p257
tps://img3.doubanio.com/view/photo/s_ratio_poster/public/p579729551.jpg', 'https://img3.doubanio.com/view/photo/s_ratio_poster/public/p1461851991.jpg',
banio.com/view/photo/s_ratio_poster/public/p1910824951.jpg', 'https://img2.doubanio.com/view/photo/s_ratio_poster/public/p2564556863.jpg', 'https://img1.
w/photo/s_ratio_poster/public/p2614500649.jpg', 'https://img9.doubanio.com/view/photo/s_ratio_poster/public/p2455050536.jpg', 'https://img9.doubanio.com/
io_poster/public/p1363250216.jpg', 'https://img9.doubanio.com/view/photo/s_ratio_poster/public/p616779645.jpg', 'https://img9.doubanio.com/view/photo/s_
ic/p2614359276.jpg', 'https://img9.doubanio.com/view/photo/s_ratio_poster/public/p2540924496.jpg', 'https://img1.doubanio.com/view/photo/s_ratio_poster/
8.jpg', 'https://img1.doubanio.com/view/photo/s_ratio_poster/public/p1505392928.jpg']
```

In [ ]:
```

```