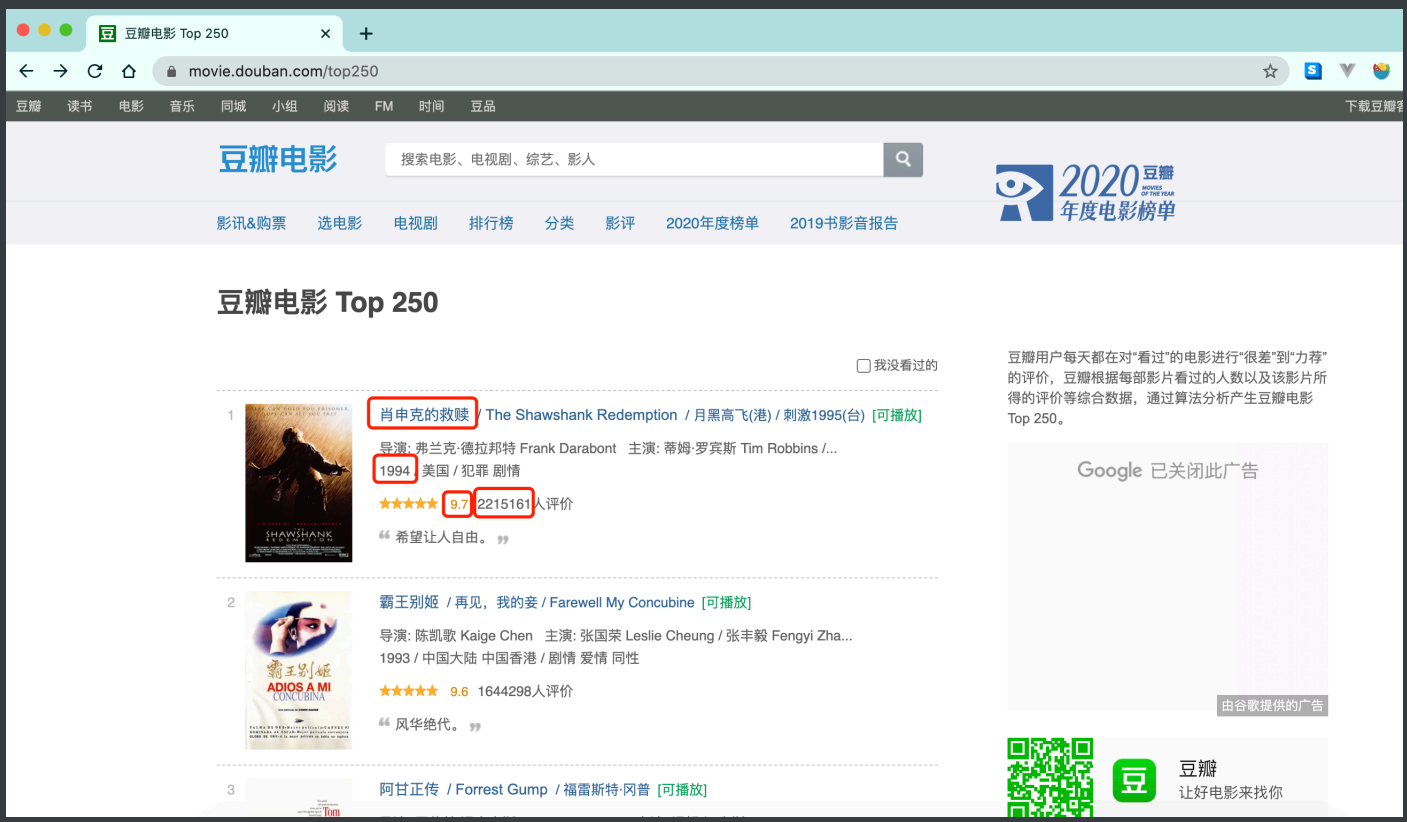
手刃豆瓣TOP250电影信息

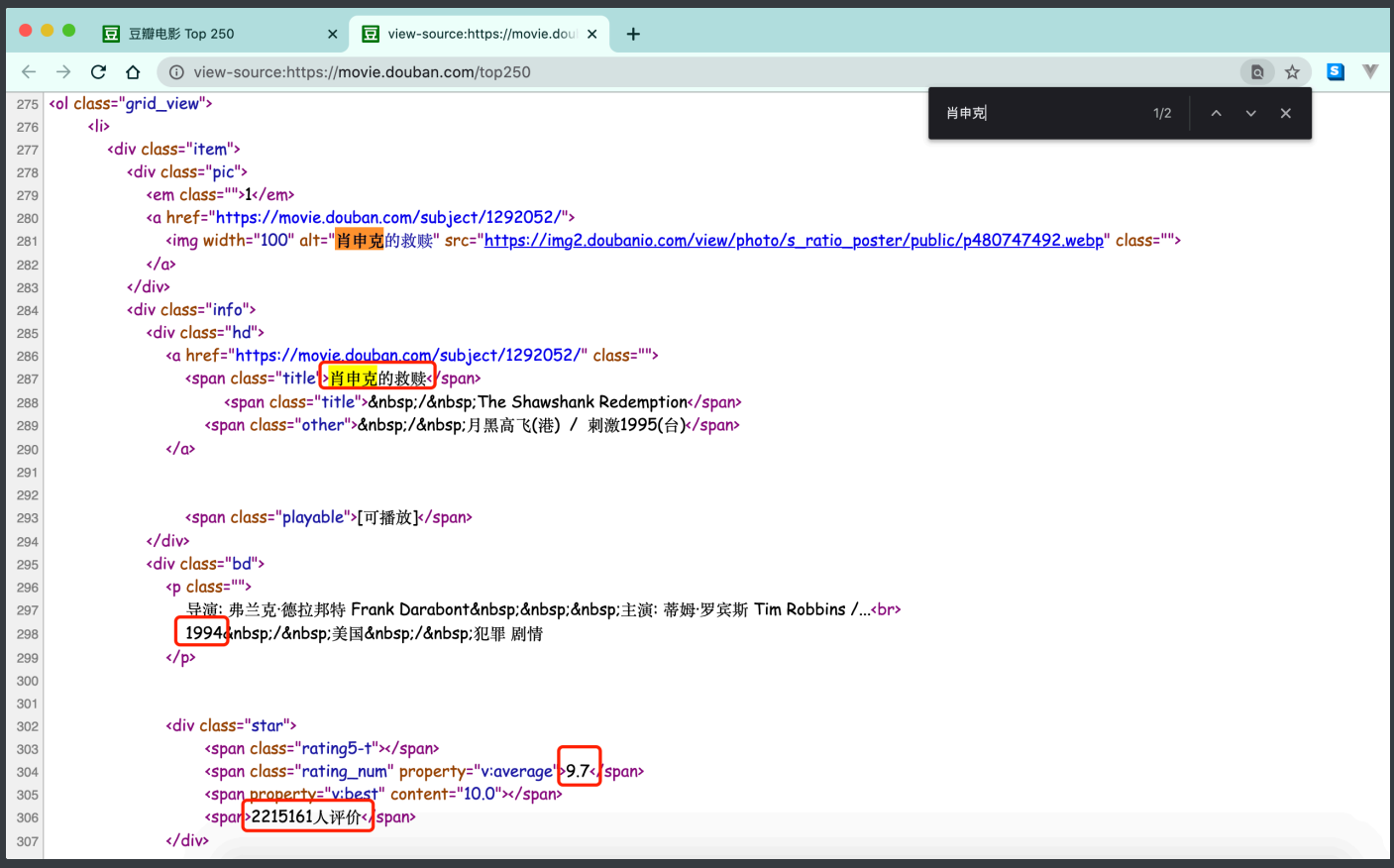
终于可以放开手脚干一番事业了. 今天我们的目标是豆瓣电影TOP250排行榜. 没别的意思, 练手而已

先看需求:



目标: 抓取"电影名称","上映年份","评分","评分人数"四项内容.

怎么做呢? 首先, 先看一下页面源代码. 数据是否是直接怼在源代码上的?



很明显, 我们想要的数据全部都在页面源代码中体现了. 所以, 我们不需要考虑js动态加载数据的情况了. 那么接下来就是编写爬虫代码的第一步了. 拿到页面源代码:

```python
import requests

headers = {
    "user-agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/87.0.4280.88 Safari/537.36"
}

url = "https://movie.douban.com/top250?start=0&filter="
resp = requests.get(url, headers=headers)
print(resp.text)

```

然后呢. 从页面源代码中提取我们需要的内容. 这时候我们就可以去写正则了.

```
obj = re.compile(r'<li>.*?<div class="item">.*?<div class="pic">.*?<em class="">(?P<num>\d+)</em>'
                 r'.*?<span class="title">(?P<name>.*?)</span>'
                 r'.*?<p class="">.*?<br>\n(?P<year>.*?) '
                 r'.*?property="v:average">(?P<average>.*?)</span>'
                 r'.*?<span>(?P<people>\d+)人评价</span>', re.S)

```

开始匹配, 将最终完整的数据按照自己喜欢(需要)的方式写入文件.

```
it = obj.finditer(resp.text)
with open("movie.csv", mode="w", encoding="utf-8") as f:
    csvwriter = csv.writer(f)  # 创建csv文件写入工具, 也可以直接f.write()
    for item in it:
        dic = item.groupdict()
        dic['year'] = dic['year'].strip()
        csvwriter.writerow(dic.values())  # 写入数据
```

代码还有优化空间, 各位可以思考一下如何进一步对代码优化(时间复杂度, 空间复杂度), 各位可以自行想办法将豆瓣TOP250条数据全部抓取到. 我这里就偷工减料了. 嘿嘿~