Python 爬虫踩坑记录

计 97 胡沐彦

2020 年 9 月 10 日

爬取思路

- 运行环境：Ubuntu 20.04，爬取豆瓣电影信息与演员信息

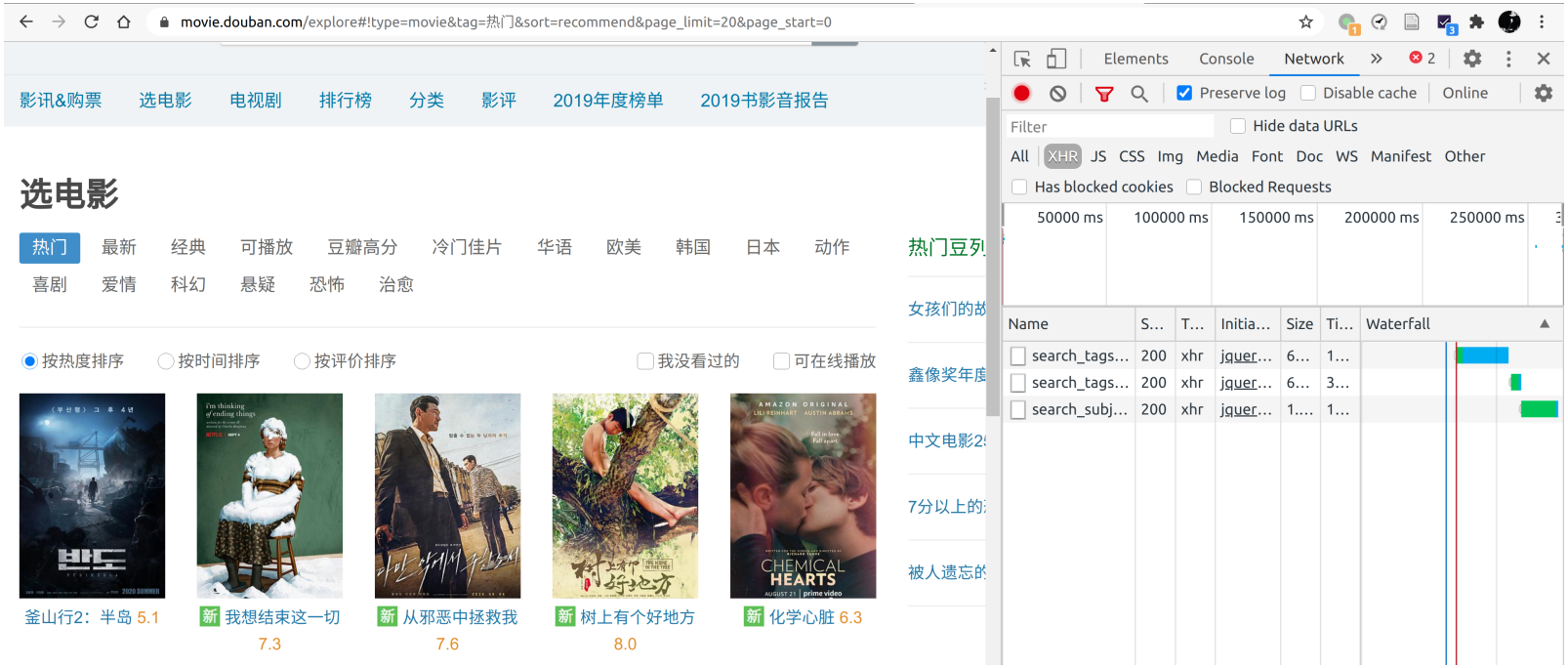
爬取思路

- 运行环境：Ubuntu 20.04，爬取豆瓣电影信息与演员信息
- 电影与演员的爬取分开进行

爬取思路

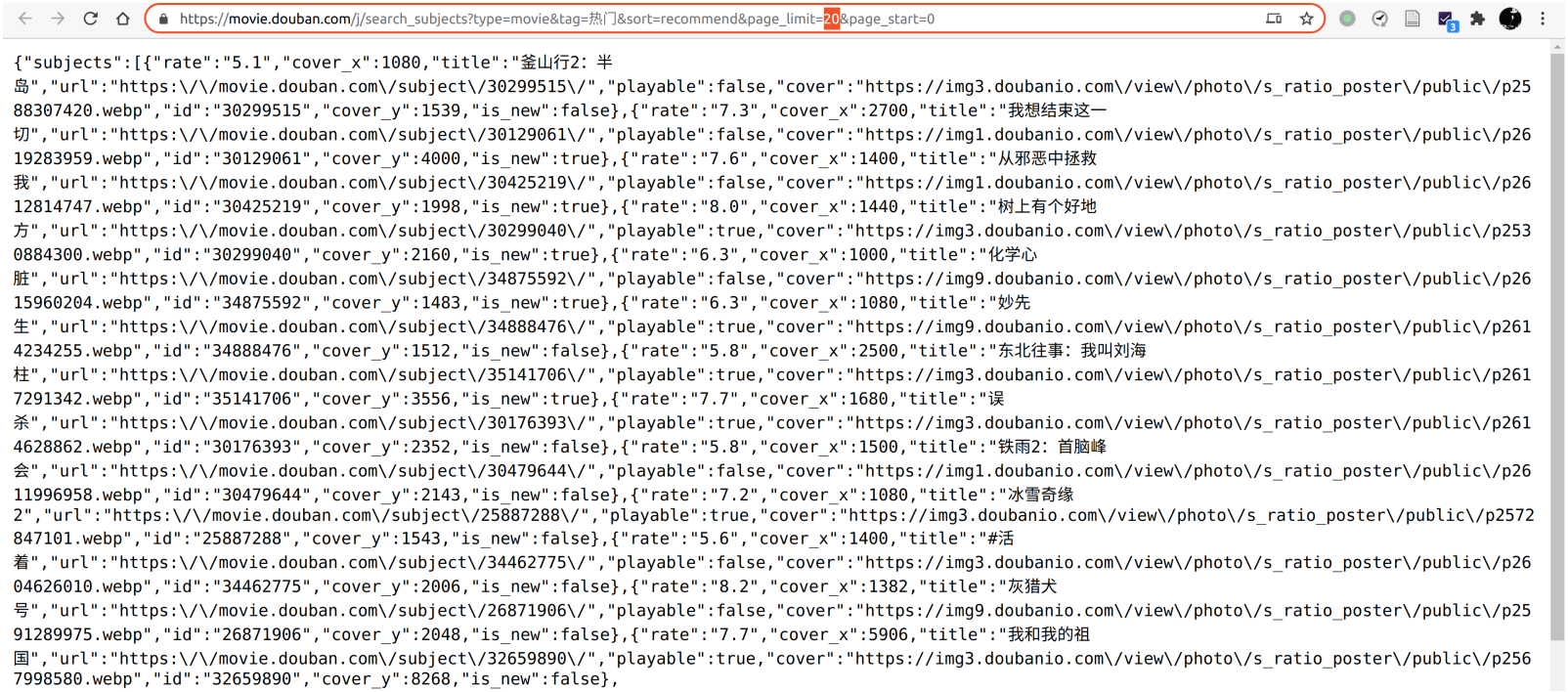
- 运行环境：Ubuntu 20.04，爬取豆瓣电影信息与演员信息
- 电影与演员的爬取分开进行
- 用 `urllib` 库进行网页页面的获取，然后利用正则表达式库 `re` 提取页面上的信息，最后用 `json` 库将信息以 json 格式保存

网页页面获取



无法直接从该页面源码中获取电影链接，用 chrome 浏览器自带的 devtools 中记录 network activity 功能抓取到需要的链接

网页页面获取



{"subjects":[{"rate":"5.1","cover_x":1080,"title":"釜山行2: 半
岛","url":"https:\/\/movie.douban.com\/subject\/30299515\/","playable":false,"cover":"https://img3.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p25
88307420.webp","id":"30299515","cover_y":1539,"is_new":false},{"rate":"7.3","cover_x":2700,"title":"我想结束这一
切","url":"https:\/\/movie.douban.com\/subject\/30129061\/","playable":false,"cover":"https://img1.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p26
19283959.webp","id":"30129061","cover_y":4000,"is_new":true},{"rate":"7.6","cover_x":1400,"title":"从邪恶中拯救
我","url":"https:\/\/movie.douban.com\/subject\/30425219\/","playable":false,"cover":"https://img1.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p26
12814747.webp","id":"30425219","cover_y":1998,"is_new":true},{"rate":"8.0","cover_x":1440,"title":"树上有个好地
方","url":"https:\/\/movie.douban.com\/subject\/30299040\/","playable":true,"cover":"https://img3.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p253
0884300.webp","id":"30299040","cover_y":2160,"is_new":true},{"rate":"6.3","cover_x":1000,"title":"化学心
脏","url":"https:\/\/movie.douban.com\/subject\/34875592\/","playable":false,"cover":"https://img9.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p26
15960204.webp","id":"34875592","cover_y":1483,"is_new":true},{"rate":"6.3","cover_x":1080,"title":"妙先
生","url":"https:\/\/movie.douban.com\/subject\/34888476\/","playable":true,"cover":"https://img9.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p261
4234255.webp","id":"34888476","cover_y":1512,"is_new":false},{"rate":"5.8","cover_x":2500,"title":"东北往事: 我叫刘海
柱","url":"https:\/\/movie.douban.com\/subject\/35141706\/","playable":true,"cover":"https://img3.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p261
7291342.webp","id":"35141706","cover_y":3556,"is_new":true},{"rate":"7.7","cover_x":1680,"title":"误
杀","url":"https:\/\/movie.douban.com\/subject\/30176393\/","playable":true,"cover":"https://img3.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p261
4628862.webp","id":"30176393","cover_y":2352,"is_new":false},{"rate":"5.8","cover_x":1500,"title":"铁雨2: 首脑峰
会","url":"https:\/\/movie.douban.com\/subject\/30479644\/","playable":false,"cover":"https://img1.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p26
11996958.webp","id":"30479644","cover_y":2143,"is_new":false},{"rate":"7.2","cover_x":1080,"title":"冰雪奇缘
2","url":"https:\/\/movie.douban.com\/subject\/25887288\/","playable":true,"cover":"https://img3.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p2572
847101.webp","id":"25887288","cover_y":1543,"is_new":false},{"rate":"5.6","cover_x":1400,"title":"#活
着","url":"https:\/\/movie.douban.com\/subject\/34462775\/","playable":false,"cover":"https://img3.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p26
04626010.webp","id":"34462775","cover_y":2006,"is_new":false},{"rate":"8.2","cover_x":1382,"title":"灰猎犬
号","url":"https:\/\/movie.douban.com\/subject\/26871906\/","playable":false,"cover":"https://img9.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p25
91289975.webp","id":"26871906","cover_y":2048,"is_new":false},{"rate":"7.7","cover_x":5906,"title":"我和我的祖
国","url":"https:\/\/movie.douban.com\/subject\/32659890\/","playable":true,"cover":"https://img3.doubanio.com\/view\/photo\/s_ratio_poster\/public\/p256
7998580.webp","id":"32659890","cover_y":8268,"is_new":false},

通过观察，调整链接中的 `page_limit` 获取所有电影列表

网页页面获取

```
req = Request(url)
req.add_header("User-Agent", Header)
sourceCode = urlopen(req).read().decode("utf-8")
```

- Header 为一个消息头字符串：Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/76.0.3809.100 Safari/537.36

网页页面获取

```
req = Request(url)
req.add_header("User-Agent", Header)
sourceCode = urlopen(req).read().decode("utf-8")
```

- Header 为一个消息头字符串：Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/76.0.3809.100 Safari/537.36
- 加入后可降低被反爬虫的概率
  - 实测在随机 sleep $2 \sim 3$s 的爬取速度下不会被封

基本信息提取

根据观察 HTML 代码设计尽可能准确的正则表达式

```
433        <span class="pl">片长:</span> <span property="v:runtime" content="115">115分钟</span><br/>
```

- 如电影长度为 `int(re.search(r'<span property="v:runtime" content="([0-9]+)">', html).group(1))`

用 `os.system` 运行 `wget` 命令下载图片，以电影/演员的豆瓣 ID 作为保存文件名

豆瓣网页的一些小 trick

- 豆瓣电影有专门的影评网页（电影链接后加`/reviews`），爬取影评更方便

豆瓣网页的一些小 trick

- 豆瓣电影有专门的影评网页（电影链接后加/reviews），爬取影评更方便
- 有些内容可能需要点击“显示全部”查看完整内容，在 HTML 源码中一般显示在 `<span class="all hidden">` 之后

Python 的一些小 trick

- 网页源代码中含有 HTML 转义字符，可以用 `html.unescape` 转换

Python 的一些小 trick

- 网页源代码中含有 HTML 转义字符，可以用 `html.unescape` 转换
- 获得满足正则表达式 `pattern` 的前 k 个匹配可以用 `for x in isslice(re.finditer(pattern, string), k)`

Python 的一些小 trick

- 网页源代码中含有 HTML 转义字符，可以用 `html.unescape` 转换
- 获得满足正则表达式 `pattern` 的前 k 个匹配可以用 `for x in isslice(re.finditer(pattern, string), k)`
- 输出为 json 文件用 `json.dumps`，为使中文正常显示需加 `ensure_ascii = False` 参数

Python 的一些小 trick

- 网页源代码中含有 HTML 转义字符，可以用 `html.unescape` 转换
- 获得满足正则表达式 `pattern` 的前 k 个匹配可以用 `for x in isslice(re.finditer(pattern, string), k)`
- 输出为 json 文件用 `json.dumps`，为使中文正常显示需加 `ensure_ascii = False` 参数
- 有的电影/演员因为信息不全可能出现匹配失败，可以用 Python 的 `try-except` 机制防止程序终止

实际运行效果

用一天时间完成全部爬取工作

实际运行效果

用一天时间完成全部爬取工作

很少被封的原因？Ubuntu？wget？

实际运行效果

用一天时间完成全部爬取工作

很少被封的原因？Ubuntu？wget？

一种成本较低的防 ban 方案：被 ban 之后路由器重播获得不同的 IP

谢谢大家！