爬取思路

网页页面获取

页面信息提

实际效果

# Python 爬虫踩坑记录

计 97 胡沐彦

2020年9月10日

网页页面获取

页面信息提耳

实际效果

### 爬取思路

• 运行环境: Ubuntu 20.04, 爬取豆瓣电影信息与演员信息

网页页面获取

实际效果

### 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

- 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

爬取思路

网页页面获取

页面信息提取

**实际**效果

### 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

胡汰彦

**伸取用** 段

网页页面获取

页面信息提明

空际新田

### 网页页面获取

切,"u"f:"https:///movie.douban.com/subject/39129061/","playable:false,"cover:"https://ingl.doubanio.com/view/photo/s\_ratio\_poster//public/p26 19283959.webp","d="39129061","cover\_y=:0009,"sep="on-trues|,frate":".6",rcover\_x=1000,"title";"MissBei推推 東","u"f:"https://movie.douban.com/subject/38425219/","playable:false,"cover\_x=1000;"https://ingl.doubanio.com/view/photo/s\_ratio\_poster/public/p26

7g, or : https:///movie-bubban.com/subject/jabs22223/, p.aspace intege\_/cover : https://imglubbanic.com/yiew/photo/s\_ratio\_poster/pab 12814747.webp", "idi-:"30425219", "cover y":1998, "is\_new":tnue), {"rate":"8.0", "cover : "ittps://imglubbanic.com/yiew/photo/s\_ratio\_poster/pablic/p253 方","url":"https:///movie.douban.com/subject/30299040/", "playable":true, "cover":"https://imglubbanic.com/yiew/photo/s\_ratio\_poster/public/p253

ンテ、"urt:"inttps:///movie.douban.com/ysubjectv/3/299440V", "p(3/301e::true, cover::inttps://mgs.doubanio.com/view/pnoto/s\_ratio\_poster/public/p255
0884300.webp", "id":"30299040", "cover y":2160, "is\_new":true}, "("ste":"6.3", "cover x":1060, "title":"化学心

#"."url":"https:///movie.douban.com/view/choto//s ratio\_poster/public//26
#"."url":"https:///movie.douban.com/view/choto//s ratio\_poster/public//26

m: nt: https://imyva.couooan.com/ysubject/34804765/2", j.cyasute: latee, over: https://imys.couoanio.com/yiew/photo/s-ratio\_poster/public/pzo 15960204.webp", "id:":"34875592", "cover\_y':1483, "is\_nem":true), {"rate":"63,", "cover\_x":1808, "title": "95\* #"."url": "https:///myvic.douban.com/ysubject/34808476//"."olavable:"true "cover": "https://imyd.doubanio.com/yiew/photo//s ratio\_poster/public/p261

主"、urt: inttps://ymovie.couoban.com/yubjectv/34004/0/", playadie:true\_cover: intps://jingy.couobanio.com/yubw/photo//s\_ratio\_poster/public/poi 4234255.webp\*, "id": '34880476", "cover y":1512, "is\_new":false), {"rate". "5.8," cover\_x":2500, "title": "5.0% "1.8," 我则第 ## "url": "https:///myke.douban.com/yubjectv/35141706//", "olayable": true, "cover": "https://ing.doubanio.com/yubw/ohoto//s ratio\_poster/youblic/yo261

7291342.webp", "id": "35141766", "cover\_y":3556, "is\_new":true}, {"rate":"7.7", "cover\_x":1680, "title":"限 条", "url":"https:///novie.douban.com/subject//38176393\/","playable":true, "cover":"https://ing3.doubanio.com/view/photo\/s ratio poster/public/p261

ポー、パーに (PS-/ V) (May 24-00-00-11-00-1

1199958.webp\*,"id\*:30479644","cover\_y":2143,"is\_new":false),("rate":"7.2","cover\_x":1080,"title":"本面積 2","uf":"https://v/movie.douban.com/>subjectv/25887286V","playoble\*:true,"cover":https://imp3.doubanio.com/view/photo//s\_ratio\_poster//public//p2572 847101.webp\*,"pd\*:725897288","cover\_y\*:1543\_";sew=fisles\_fi-rate":5.6","cover\_x\*:1404\_"title":"海道

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":2806,"is\_new":false},{"rate":"8.2","cover\_x":1382,"title":"来購大

号","url":"https:\/\movie.douban.com\/subject\/26871996\/"."playable":false,"cover":"https://img9.doubanio.com\/view\/photo\/s\_ratio\_poster\/public\/p25
91289975.webp","id":"28871996","cover\_y":2848,"is\_new":false},{"rate":"7.7","cover\_x":5906,"title":"我和我的祖

圖","url":"https:\/\movie.douban.com\/subject\/32659890\/"."playable":true,"cover":"https://ing3.doubanio.com\/view\/photo\/s\_ratio\_poster\/public\/p2567998580.webp","id":"326598980","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

空际苏田

# 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

爬取思路

网页页面获取

贝囬信息提明

空际初里

### 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  $2 \sim 3s$  的爬取速度下不会被封

页面信息提取

实际效果

### 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

 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

爬取思路

网页页面获取

页面信息提取

实际效果

### 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 网页页面获取

页面信息提取

空际初里

### 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

爬取思路

网页页面获取

页面信息提取

# 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

爬取思路

网页页面获取

页面信息提取

买际效果

### 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)

爬取思路

网页页面获取

页面信息提取

实际效果

### 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

爬取思路

网页页面获取

页面信息提取

实际效果

### 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

爬取思路

网页页面获取

页面信息提取

实际效果

### 用一天时间完成全部爬取工作

实际运行效果

网页页面获取

页面信息提

实际效果

# 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? 网页页面获取

页面信息提

实际效果

# 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

爬取思路

网页页面获取

页面信息提取

实际效果

# 谢谢大家!