

Python爬虫实战-爬取豆瓣电影TOP250榜单

URL: <https://movie.douban.com/top250>

IDE: Pycharm Professional

前置准备

观察网页源代码，可以发现整个榜单的主体在class为article的div标签下


[illegible]

每一个电影单独放在一个li标签下

豆瓣电影 Top 250


li 675×176.05

☐ 我没看过

1  **肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台)** [可播放]


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

★★★★★ 9.7 2590072人评价
希望让人自由。 99

2  **霸王别姬 / 再见，我的妾 / Farewell My Concubine** [可播放]

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

★★★★★ 9.6 1923558人评价
风华绝代。 99

3  **阿甘正传 / Forrest Gump / 福雷斯脱·冈普** [可播放]

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

★★★★★ 9.5 1946214人评价
一部美国近现代史。 99

豆瓣用户每天都在对“看过”的电影进行的评价，豆瓣根据每部影片看过的人数得到的评价等综合数据，通过算法分析产生Top 250。

```

<link href="//img3.doubanio.com/dae/accounts/resources/3e96b44/movie/bundl_e_css" rel="stylesheet" type="text/css">
<div id="db-nav-movie" class="nav"></div>
<script id="suggResult" type="text/x-jquery-tmpl"></script>
<script src="//img3.doubanio.com/dae/accounts/3e96b44/movie/bund le_js" defer="defer"></script>
<div id="wrapper">
  <div id="content">
    <h1>豆瓣电影 Top 250</h1>
    <div class="grid-16-8 clearfix">
      <div class="article"> == $0
        <div class="opt mod"></div>
        <ol class="grid_view">
          <li></li>
          <li></li>
          <li></li>
          <li></li>
          <li></li>
          <li></li>
          <li></li>
          <li></li>
        </ol>
      </div>
    </div>
  </div>
</div>
html.ua-linux.ua-webkit body#t.neterror div#wrapper div#content div-grid-16-8.clearfix div ...
Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility
Filter :hov .cls +, <
Console
top Filter Default levels 2 Issues: 2
喜欢看豆瓣的代码，还是发现了什么bug? 不知和我们一起为豆瓣添砖加瓦吧! douban.js:2 http://jobs.douban.com/#position-zsmd
> crrbug/1173575, non-JS module files deprecated. {index}:6774
        
```


电影信息就放在下级class为info的标签下


影讯&购票 选电影 电视剧 排行榜 分类 影评 2021年度榜单 2021书影音报告 平仄电影网


豆瓣电影 Top 250

div.info 522.76 × 151.05 ☐ 我没看过的

豆瓣用户每天都在对“看过”的电影进行的评价，豆瓣根据每部影片看过的人数得的评价等综合数据，通过算法分析产生Top 250。

- 

肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]
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
★★★★★ 9.7 2590072人评价
希望让人自由。 99
- 

霸王别姬 / 再见，我的妾 / Farewell My Concubine [可播放]
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
★★★★★ 9.6 1923558人评价
风华绝代。 99
- 

阿甘正传 / Forrest Gump / 福雷斯特·冈普 [可播放]
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
★★★★★ 9.5 1946214人评价
一部美国近现代史。 99

豆瓣 让好电影： 由你

DevTools is now available in Chinese! Always match Chrome's language | Switch DevTools to Chinese | Don't show again

Elements Console Sources Network

HTML: <div id="wrapper"> <div id="content"> <h1>豆瓣电影 Top 250</h1> <div class="grid-16-8 clearfix"> <div class="article"> \$0 <div class="pic"> </div> <div class="info"> <ol class="grid_view"> <div class="item"> <div class="opt mod"> </div> <div class="hd"> </div> <div class="bd"> </div> </div> </div> </div> </div> </div>

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls +

Console

喜欢着豆瓣的代码，还是发现了什么bug? 不如和我们一起为豆瓣添砖加瓦吧! douban.js:2 http://jobs.douban.com/#position-zs9d

crbug/1173575, non-JS module files deprecated. (index):6774

清楚结构后，就可以开始写代码了

获取数据

如果不更改请求头中的user-agent，会返回响应码418，因此要修改请求头，如下所示

```
import requests

url = "https://movie.douban.com/top250"
headers = {"user-agent": "Mozilla/5.0 (X11; Linux x86_64)"}
res = requests.get(url=url, headers=headers)
print(res.status_code)
```

现在响应码等于200了，表明成功

```
/usr/bin/python3.7 /home/hwx/PycharmProjects/work/main.py
200
```

```
Process finished with exit code 0
```


接下来直接按照标签对数据进行提取

```
res = requests.get(url=url, headers=headers)
soup = BeautifulSoup(res.text, "html.parser")
targets = soup.find("ol", class_="grid_view").find_all("div", class_="info")
```

这个info标签组成的列表，就包括了第一页所有电影的信息

豆瓣电影 Top 250

div.hd 522.76 × 21.85 ☐ 我没看过的




肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]

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

★★★★★ 9.7 2591442人评价

希望让人自由。 99




霸王别姬 / 再见，我的妾 / Farewell My Concubine [可播放]

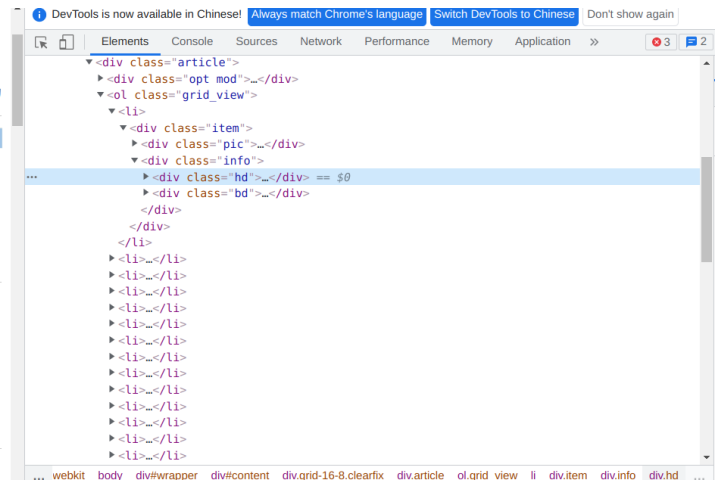
导演: 陈凯歌 Kaige Chen 主演: 张国荣 Leslie Cheung / 张丰毅 Fengyi Zha...
1993 / 中国大陆 中国香港 / 剧情 爱情 同性

★★★★★ 9.6 1924610人评价

风华绝代。 99



阿甘正传 / Forrest Gump / 福雷斯特·冈普 [可播放]



豆瓣电影 Top 250

div.bd 522.76 × 109.2 ☐ 我没看过的



肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]

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

★★★★★ 9.7 2591442人评价

希望让人自由。 99

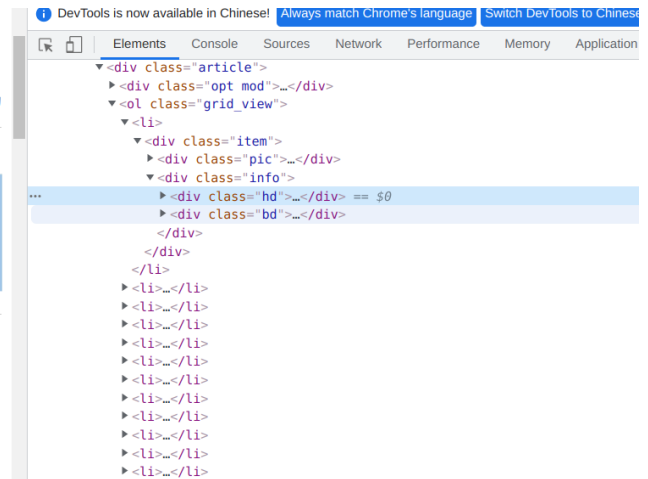


霸王别姬 / 再见，我的妾 / Farewell My Concubine [可播放]

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

★★★★★ 9.6 1924610人评价

风华绝代。 99



info标签下包括两部分，一部分在hd标签中，作为电影信息的头部，主要包括电影名称和链接。bd标签中是电影信息的主体，包含导演、演员、评分和引语

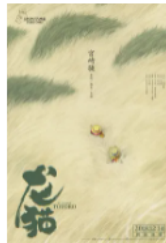
```
<div class="hd">
  <a href="https://movie.douban.com/subject/1292052/" class=
**
    <span class="title">肖申克的救赎</span> == $0
    <span class="title">&nbsp;/&nbsp;/&nbsp;The Shawshank Redemption</span>
    <span class="other">&nbsp;/&nbsp;/&nbsp;月黑高飞(港) / 刺激1995(台)</span>
  </a>
  <span class="playable">[可播放]</span>
</div>
<div class="bd">...</div>
```

hd中title标记的标签中，第一个就是电影在大陆的中文名称，bd的p标签中所包含了演员信息、类别和年份。

豆瓣电影 Top 250

☐ 我没看过的(250)

26



龙猫 / ととなりのトトロ / 邻居托托罗 / 邻家的豆豆龙 [\[可播放\]](#)

导演: 宫崎骏 Hayao Miyazaki 主演: 日高法子 Noriko Hidaka / 坂本千夏 Ch...
1988 / 日本 / 动画 奇幻 冒险

★★★★★ 9.2 1145720人评价

“人人心中都有个龙猫，童年就永远不会消失。”

[想看](#) [看过](#)

27



末代皇帝 / The Last Emperor / 末代皇帝溥仪(港) [\[可播放\]](#)

导演: 贝纳尔多·贝托鲁奇 Bernardo Bertolucci 主演: 尊龙 John Lone / 陈...
1987 / 英国 意大利 中国大陆 法国 / 剧情 传记 历史

★★★★★ 9.3 766467人评价

“不要跟我比惨，我比你更惨”再适合这部电影不过了。”

[想看](#) [看过](#)

start每递增25，就能访问到下一页

```

for i in range(10):
    url = f"https://movie.douban.com/top250?start={i*25}"
    res = requests.get(url=url, headers=headers)
    if res.status_code != 200:
        print("访问异常")
        break
    soup = BeautifulSoup(res.text, "html.parser")
    targets = soup.find("ol", class_="grid_view").find_all("div", class_="info")
    for target in targets:
        head = target.find("div", class_="hd")
        title = head.find("span", class_="title").text # 标题
        link = head.find("a")["href"] # 链接
        body = target.find("div", class_="bd")
        temp = body.find("p").text.strip().split("\n") # 切割
        person = "".join(temp[0].strip().split("\xa0")).strip() # 人员信息
        form = "".join(temp[1].strip().split("\xa0")).strip() # 电影分类
        try:
            quote = body.find_all("p")[1].text.strip()
        except:
            quote = "无评语"
        points = body.find("div", class_="star").text.strip().split()[0]
        result.append((title, form, person, points, quote, link))
    print(url)
    time.sleep(1)

```

加一层循环，并且每次循环睡眠一秒种，降低访问频率，防止被封IP

最后是文件保存

```

f = open("./result.txt", "w+")
for items in result:
    print(items)
    for item in items:
        f.write(item+"\t\t")
    f.write("\n")
f.close()

```

220	新龙门客栈	1992/中国香港	中国大陆/动作	爱情 武侠 古装	导演: 李惠民 Raymond Lee主演: 张曼玉 Maggie Cheung / 林青霞 Bri...	8.7	287	^	v	
221	疯狂的麦克斯4：狂暴之路	2015/澳大利亚	美国/动作	科幻 冒险	导演: 乔治·米勒 George Miller主演: 汤姆·哈迪 Tom Hardy / 查理兹·塞...	8.8				
222	无耻混蛋	2009/德国	美国/剧情	犯罪	导演: Quentin Tarantino主演: 布拉德·皮特 Brad Pitt / 梅拉尼·罗兰 M...	8.6	昆汀同学越来越变态			
223	魂断蓝桥	1940/美国	美国/剧情	爱情 战争	导演: 茂文·勒鲁瓦 Mervyn LeRoy主演: 费雯·丽 Vivien Leigh / 罗伯特·...	8.8	中国式内在的美国情			
224	血钻	2006/美国	德国 英国/剧情	惊悚 冒险	导演: 爱德华·兹威克 Edward Zwick主演: 莱昂纳多·迪卡普里奥 Leonardo ...	8.7	每个美丽事物			
225	步履不停	2008/日本	日本/剧情	家庭	导演: 是枝裕和 Hirokazu Koreeda主演: 阿部宽 Hiroshi Abe / 夏川结衣 Yu...	8.8	日本的家庭电影已经			
226	黑客帝国2：重装上阵	2003/美国	美国/动作	科幻	导演: 拉娜·沃卓斯基 Lana Wachowski / 莉莉·沃卓斯基 Lilly Wachowski...	8.7	一个精彩的世界			
227	千钧一发	1997/美国	美国/剧情	科幻 惊悚	导演: 安德鲁·尼科尔 Andrew Niccol主演: 伊桑·霍克 Ethan Hawke / 乌玛...	8.8	一部能引人思			
228	彗星来的那一夜	2013/美国	英国/科幻	悬疑 惊悚	导演: 詹姆斯·沃德·布柯特 James Ward Byrkit主演: 艾米丽·芭尔多尼 Em...	8.5	小成本大			
229	战争之王	2005/法国	德国 美国/剧情	犯罪	导演: 安德鲁·尼科尔 Andrew Niccol主演: 尼古拉斯·凯奇 Nicolas Cage / ...	8.7	做一颗让			
230	崖上的波妞	2008/日本	日本/动画	奇幻 冒险	导演: 宫崎骏 Hayao Miyazaki主演: 奈良柚莉爱 Yuria Nara / 土井洋辉 Hir...	8.6	无评语			
231	燃情岁月	1994/美国	美国/剧情	爱情 战争 西部	导演: 爱德华·兹威克 Edward Zwick主演: 布拉德·皮特 Brad Pitt / 安东...	8.8	传奇，不是每			
232	谍影重重2	2004/美国	德国/动作	悬疑 惊悚	导演: 保罗·格林格拉斯 Paul Greengrass主演: 马特·达蒙 Matt Damon / ...	8.7	谁说王家卫镜			
233	爱乐之城	2016/美国	美国/剧情	爱情 歌舞	导演: 达米恩·查泽雷 Damien Chazelle主演: 瑞恩·高斯林 Ryan Gosling / ...	8.4	无评语			
234	阿飞正传	1990/中国香港	香港/犯罪	剧情 爱情	导演: 王家卫 Kar Wai Wong主演: 张国荣 Leslie Cheung / 张曼玉 Maggie C...	8.5	王家卫最			
235	海洋	2009/法国	瑞士 西班牙 美国	阿联酋 摩纳哥/纪录片	导演: 雅克·贝汉 Jacques Perrin / 雅克·克鲁奥德 Jacques Cluzaud主演:...	9.0				
236	谍影重重	2002/美国	德国 捷克/动作	悬疑 惊悚	导演: 道格·里曼 Doug Liman主演: 马特·达蒙 Matt Damon / 弗兰卡·波坦...	8.6	哗啦啦啦			
237	穿越时空的少女	2006/日本	日本/剧情	爱情 科幻 动画	导演: 细田守 Mamoru Hosoda主演: 仲里依纱 Riisa Naka / 石田卓也 Takuya...	8.6	爱			
238	再次出发之纽约遇见你	2013/美国	美国/喜剧	爱情 音乐	导演: 约翰·卡尼 John Carney主演: 凯拉·奈特莉 Keira Knightley / 马克...	8.6				
239	心灵奇旅	2020/美国	美国/动画	奇幻 音乐	导演: 彼特·道格特 Pete Docter / 凯普·鲍尔斯 Kemp Powers主演: 杰米·...	8.7	无评语		ht	
240	香水	2006/德国	法国 西班牙 美国	比利时/剧情	犯罪 奇幻	导演: 汤姆·提克威 Tom Tykwer主演: 本·卫肖 Ben Whishaw / 艾伦·瑞克...	8.5			
241	地球上的星星	2007/印度	印度/剧情	儿童 家庭	导演: 阿米尔·汗 Aamir Khan主演: 达席尔·萨法瑞 Darsheel Safary / 阿...	8.9	天使保护事件			
242	火星救援	2015/英国	美国 匈牙利 约旦	美国/剧情	科幻 冒险	导演: 雷德利·斯科特 Ridley Scott主演: 马特·达蒙 Matt Damon / 杰西卡...	8.5	无		
243	朗读者	2008/美国	德国/剧情	爱情	导演: 斯蒂芬·戴德利 Stephen Daldry主演: 凯特·温丝莱特 Kate Winslet ...	8.6	当爱情跨越年			
244	我爱你	2011/韩国	韩国/剧情	爱情	导演: 秋昌民 Chang-min Choo主演: 宋在河 Jae-ho Song / 李顺载 Soon-jae...	9.1	你要相信，这世上真			
245	完美陌生人	2016/意大利	意大利/剧情	喜剧	导演: 保罗·格诺维瑟 Paolo Genovese主演: 马可·贾利尼 Marco Giallini ...	8.5	来啊，互相伤			
246	末路狂花	1991/美国	美国 法国/犯罪	剧情 惊悚	导演: 雷德利·斯科特 Ridley Scott主演: 吉娜·戴维斯 Geena Davis / 苏...	8.8	没有了道			
247	千年女优	2001/日本	日本/动画	剧情 爱情	导演: 今敏 Satoshi Kon主演: 庄司美代子 Miyoko Shôji / 小山茉美 Mam...	8.8	爱情是一场没有尽头			
248	驴得水	2016/中国大陆	中国/剧情	喜剧	导演: 周申 Shen Zhou / 刘露 Lu Liu主演: 任素汐 Suxi Ren / 大力 Da Li ...	8.3	过去的如果就			
249	聚焦	2015/美国	美国/剧情	传记	导演: 托马斯·麦卡锡 Thomas McCarthy主演: 马克·鲁弗洛 Mark Ruffalo / ...	8.8	新闻人的理性求真。			
250	小萝莉的猴神大叔	2015/印度	印度/剧情	喜剧 动作	导演: 卡比尔·汗 Kabir Khan主演: 萨尔曼·汗 Salman Khan / 哈莎莉·马...	8.4	宝莱坞的			
251										

保存下来的结果

完整代码


```

import requests
import time
from bs4 import BeautifulSoup

headers = {"user-agent": "Mozilla/5.0 (X11; Linux x86_64)"}
result = []
for i in range(10):
    url = f"https://movie.douban.com/top250?start={i*25}"
    res = requests.get(url=url, headers=headers)
    if res.status_code != 200:
        print("访问异常")
        break
    soup = BeautifulSoup(res.text, "html.parser")
    targets = soup.find("ol", class_="grid_view").find_all("div", class_="info")
    for target in targets:
        head = target.find("div", class_="hd")
        title = head.find("span", class_="title").text # 标题
        link = head.find("a")["href"] # 链接
        body = target.find("div", class_="bd")
        temp = body.find("p").text.strip().split("\n") # 切割
        person = "".join(temp[0].strip().split("\xa0")).strip() # 人员信息
        form = "".join(temp[1].strip().split("\xa0")).strip() # 电影分类
        try:
            quote = body.find_all("p")[1].text.strip()
        except:
            quote = "无评语"
        points = body.find("div", class_="star").text.strip().split()[0]
        result.append((title, form, person, points, quote, link))
    print(url)
    time.sleep(1)

f = open("./result.txt", "w+")
for items in result:
    print(items)
    for item in items:
        f.write(item+"\t\t")
    f.write("\n")
f.close()

```