

Selenium技术按标签爬取豆瓣电影

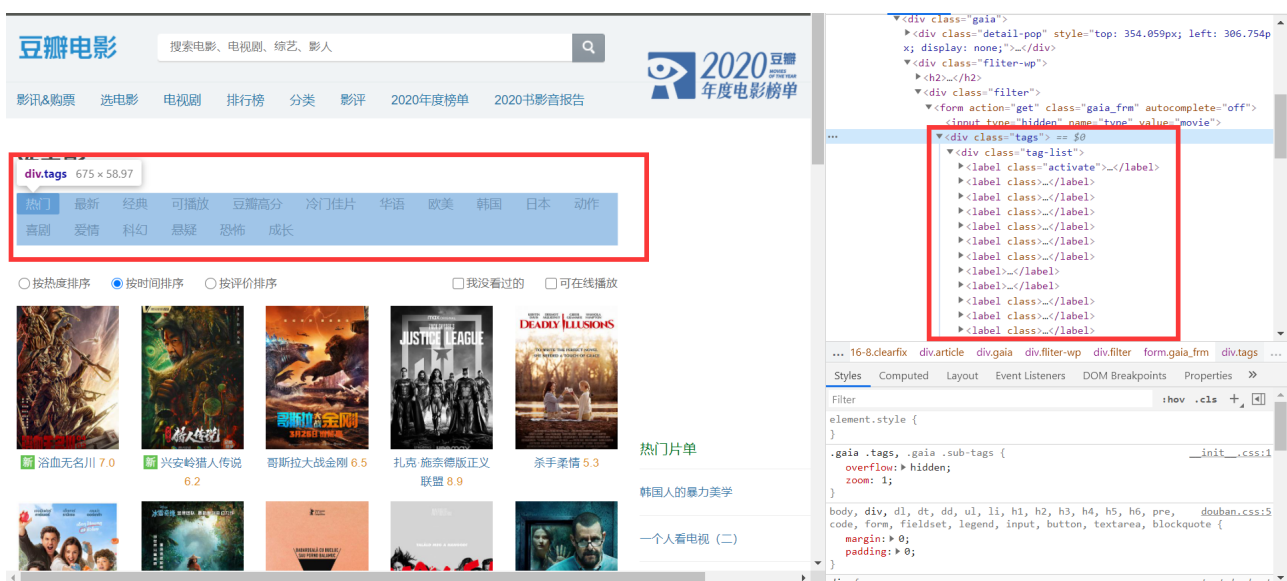
1.爬取准备

这次爬取的是豆瓣的“选电影”页面

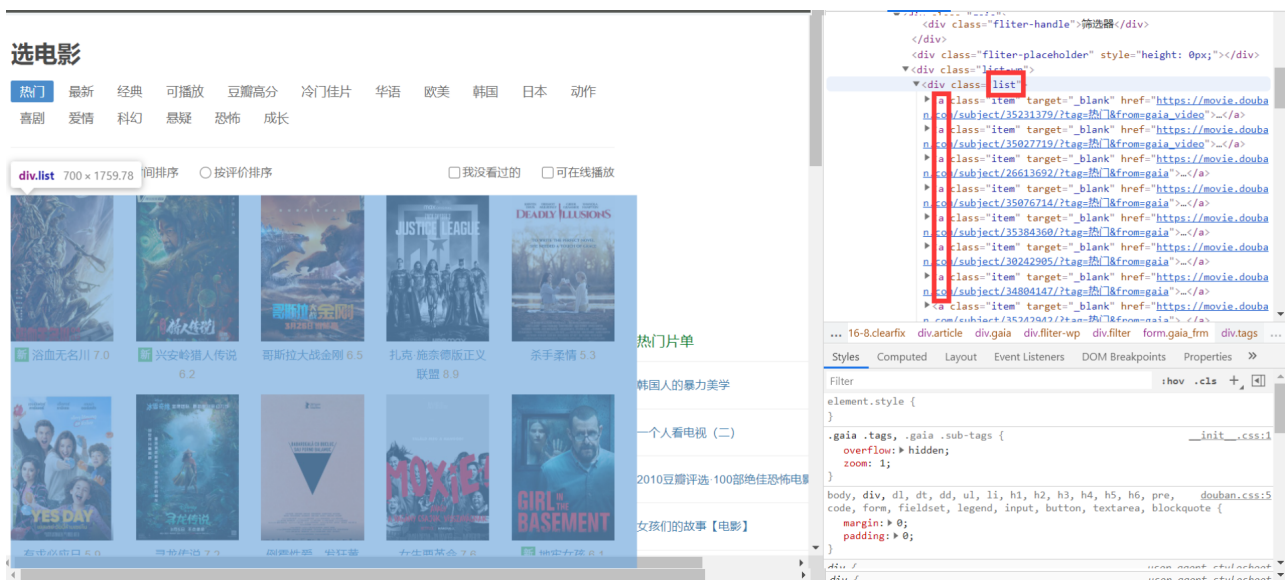
”页面。



简单的分析可以发现，导航栏的所有标签在一个div中，而每个标签存放在一个label中。我们只需要模拟点击去访问这些label标签就可以实现下方电影内容的变化。

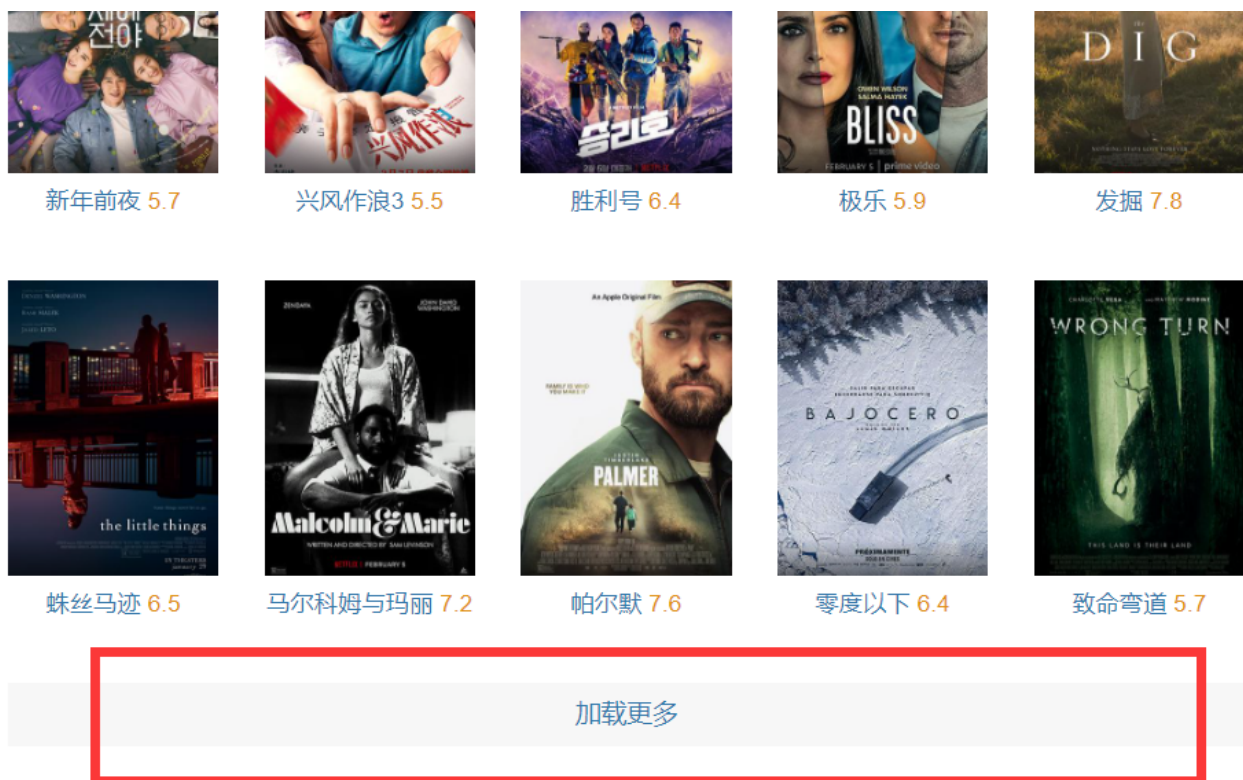


电影信息存放较为简单，一个a标签存放了其转跳的连接，而电影名和评分直接在下方的两个字符标签中。



思路整理：外层的循环遍历所有的**tags**，遍历的同时捕获页面下方的电影信息，从而达到爬取所有标签下电影的效果。最后将数据保存在**csv**文件中。

改进之处：可以在日后加入点击下方“加载更多”的按钮以实现更多数据的爬取。



2.爬取代码

```
from selenium import webdriver
import selenium
from bs4 import BeautifulSoup
from time import sleep
```

```

import pandas as pd

def get_cur_page():
    # 爬取当前页面的所有内容
    buffer = []
    items_list = browser.find_elements_by_css_selector(".list a")
    for ele in items_list:
        tmp = ele.text.split(" ")
        print(tmp[0], tmp[-1], ele.get_attribute('href'))
        try:
            buffer.append([tmp[0], tmp[-1],
ele.get_attribute('href')])
        except exception as ex:
            print("此电影没有评分")
    return pd.DataFrame(buffer, columns=['title', 'rate', 'url'])

if __name__ == '__main__':
    df = pd.DataFrame(columns=['title', 'rate', 'url'])
    # 1.爬取tag-list
    browser =
selenium.webdriver.Edge(executable_path="S:/web_browser_driver/msed
gedriver.exe")
    target_url =
'https://movie.douban.com/explore#!type=movie&tag=%E7%83%AD%E9%97%A
8&sort=time&page_limit=20&page_start=0'
    browser.implicitly_wait(5)
    browser.get(target_url)
    page_list = browser.find_elements_by_css_selector(".tag-list
label")

    # 2.爬取每个页面的内容
    for page in page_list:
        page.click()

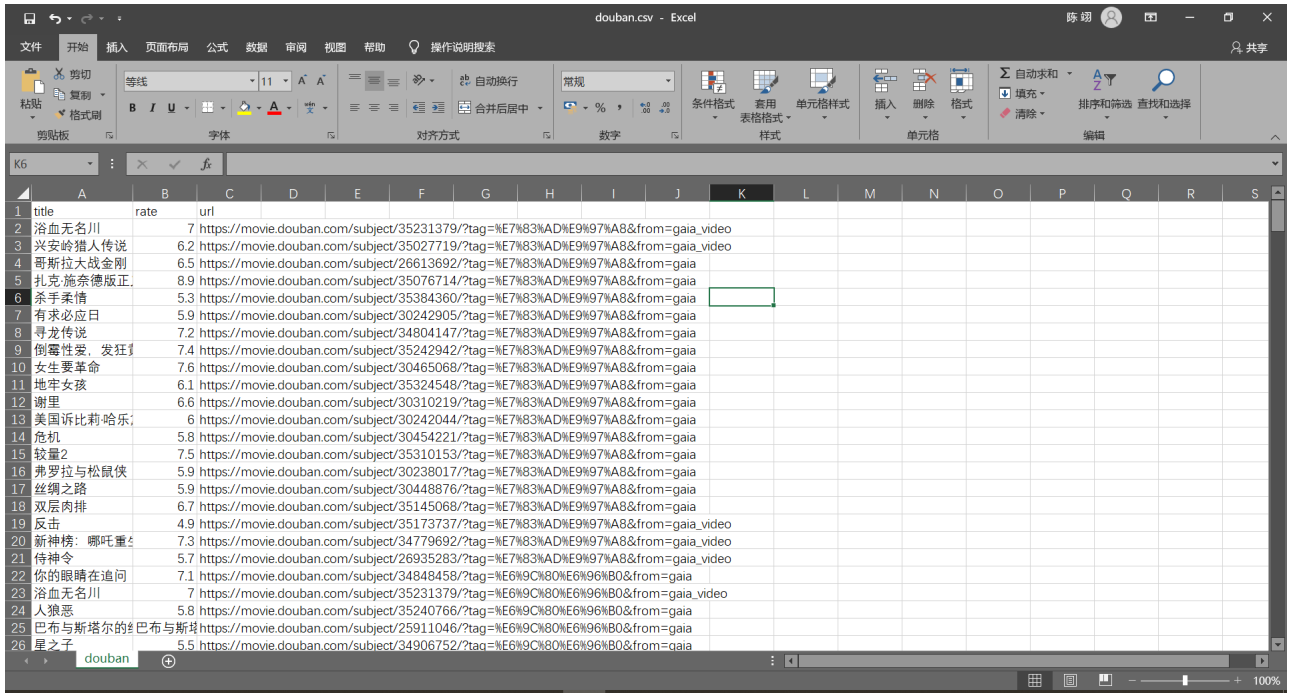
        df = df.append(get_cur_page(), ignore_index=True)
        sleep(2)
        print(df)

    # 关闭selenium并保存
    path = "E:/douban.csv"
    browser.close()

```

```
df.to_csv(path, index=False, encoding="utf_8_sig")
print("保存成功", path)
```

3.爬取结果



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	title	rate	url																
2	浴血无名川	7	https://movie.douban.com/subject/35231379/?tag=%E7%83%AD%E9%97%A8&from=gaia_video																
3	兴安岭猎人传说	6.2	https://movie.douban.com/subject/35027719/?tag=%E7%83%AD%E9%97%A8&from=gaia_video																
4	哥斯拉大战金刚	6.5	https://movie.douban.com/subject/26613692/?tag=%E7%83%AD%E9%97%A8&from=gaia																
5	扎克施奈德版正	8.9	https://movie.douban.com/subject/35076714/?tag=%E7%83%AD%E9%97%A8&from=gaia																
6	杀手柔情	5.3	https://movie.douban.com/subject/35384360/?tag=%E7%83%AD%E9%97%A8&from=gaia																
7	有求必应日	5.9	https://movie.douban.com/subject/30242905/?tag=%E7%83%AD%E9%97%A8&from=gaia																
8	寻龙传说	7.2	https://movie.douban.com/subject/34804147/?tag=%E7%83%AD%E9%97%A8&from=gaia																
9	倒置性爱, 发狂	7.4	https://movie.douban.com/subject/35242942/?tag=%E7%83%AD%E9%97%A8&from=gaia																
10	女生要革命	7.6	https://movie.douban.com/subject/30465068/?tag=%E7%83%AD%E9%97%A8&from=gaia																
11	地牢女孩	6.1	https://movie.douban.com/subject/35324548/?tag=%E7%83%AD%E9%97%A8&from=gaia																
12	谢里	6.6	https://movie.douban.com/subject/30310219/?tag=%E7%83%AD%E9%97%A8&from=gaia																
13	美国诉比莉·哈乐	6	https://movie.douban.com/subject/30242044/?tag=%E7%83%AD%E9%97%A8&from=gaia																
14	危机	5.8	https://movie.douban.com/subject/30454221/?tag=%E7%83%AD%E9%97%A8&from=gaia																
15	较量2	7.5	https://movie.douban.com/subject/35310153/?tag=%E7%83%AD%E9%97%A8&from=gaia																
16	弗罗拉与松鼠侠	5.9	https://movie.douban.com/subject/30238017/?tag=%E7%83%AD%E9%97%A8&from=gaia																
17	丝绸之路	5.9	https://movie.douban.com/subject/30448876/?tag=%E7%83%AD%E9%97%A8&from=gaia																
18	双层肉排	6.7	https://movie.douban.com/subject/35145068/?tag=%E7%83%AD%E9%97%A8&from=gaia																
19	反击	4.9	https://movie.douban.com/subject/35173737/?tag=%E7%83%AD%E9%97%A8&from=gaia_video																
20	新神榜: 哪吒重生	7.3	https://movie.douban.com/subject/34779692/?tag=%E7%83%AD%E9%97%A8&from=gaia_video																
21	侍神令	5.7	https://movie.douban.com/subject/26935283/?tag=%E7%83%AD%E9%97%A8&from=gaia_video																
22	你的眼睛在追问	7.1	https://movie.douban.com/subject/34848458/?tag=%E6%9C%80%E6%96%B0&from=gaia																
23	浴血无名川	7	https://movie.douban.com/subject/35231379/?tag=%E6%9C%80%E6%96%B0&from=gaia_video																
24	人狼恶	5.8	https://movie.douban.com/subject/35240766/?tag=%E6%9C%80%E6%96%B0&from=gaia																
25	巴布与斯塔尔的维	巴布与斯塔尔的维	https://movie.douban.com/subject/25911046/?tag=%E6%9C%80%E6%96%B0&from=gaia																
26	星之子	5.5	https://movie.douban.com/subject/34906752/?tag=%E6%9C%80%E6%96%B0&from=gaia																

演示gif（加载时间或许很长）