

selenium概述

我们在抓取一些普通网页的时候requests基本上是可以满足的. 但是, 如果遇到一些特殊的网站. 它的数据是经过加密的. 但是呢, 浏览器却能够正常显示出来. 那我们通过requests抓取到的内容可能就不是我们想要的结果了. 例如,

票房

实时票房

单日票房

单周票房

周末票房

单月票房

年度票房

全球票房

历史票房

影院票房

排片

中美

情报

年度票房

年度首周票房

2021年


	影片名称	类型	总票房(万)	平均票价	场均人次	国家及地区	上映日期
1	 送你一朵小红花	剧情	104,689	37	12	中国	2020-12-31
2	温暖的抱抱	喜剧	59,660	37	10	中国	2020-12-31
3	拆弹专家2	动作	53,984	39	10	中国	2020-12-24
4	心灵奇旅	动画	23,597	38	10	美国	2020-12-25
5	缉魂	科幻	9,353	36	5	中国	2021-01-15
6	许愿神龙	动画	8,689	35	6	中国/美国	2021-01-15
7	晴雅集	动作	8,264	38	19	中国	2020-12-25
8	大红包	喜剧	5,490	34	6	中国	2021-01-22
9	海底小纵队：火焰之环	动画	4,612	33	6	中国	2021-01-08

电影票房数据. 在浏览器上看的时候是正常的. 那么按照之前的逻辑. 我们只需要看看数据是通过哪个请求拿到的就可以进行模拟请求了. 但是!

年度票房

年度首周票房

2020年

	影片名称	类型
1	 八佰	战争
2	我和我的家乡	喜剧
3	姜子牙	动画
4	金刚川	剧情/战争
5	夺冠	剧情
6	拆弹专家2	动作
7	除暴	动作
8	宠爱	剧情
9	我在时间尽头等你	爱情
10	误杀	剧情
11	信条	科幻
12	紧急救援	剧情
13	叶问4：完结篇	动作

Filter

Hide data URLs AllXHRJS CSS Img Media Font Doc WS Manifest OtherHas blocked cookies

500 ms

1000 ms

1500 ms

2000 ms

2500 ms

3000 ms

3500 ms

4000 ms

Name

×

Headers

Preview

Response

Initiator

Timing

GetData.ashx

/API

Content-Type: text/plain; charset=utf-8

Date: Mon, 25 Jan 2021 09:36:53 GMT

Expires: -1

Pragma: no-cache

Server: nginx/1.14.1

X-AspNet-Version: 4.0.30319

X-Powered-By: ASP.NET

Request Headers

view source

Accept: text/plain, */*; q=0.01

Accept-Encoding: gzip, deflate, br

Accept-Language: zh-CN,zh;q=0.9,en;q=0.8

Cache-Control: no-cache

Connection: keep-alive

Content-Length: 46

Content-Type: application/x-www-form-urlencoded; charset=UTF-8

DNT: 1

Host: www.endata.com.cn

Origin: https://www.endata.com.cn

Pragma: no-cache

sec-ch-ua: "Chromium";v="88", "Google Chrome";v="88", ";Not A Brand";v="99"

sec-ch-ua-mobile: 70

Sec-Fetch-Dest: empty

Sec-Fetch-Mode: cors

Sec-Fetch-Site: same-origin

User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_4) AppleWebKit/537.36 (KHTML, like Gecko) Safari/537.36

X-Requested-With: XMLHttpRequest

Form Data

view source

view URL encoded

year: 2020

Method Name: BoxOffice_GetYearInfoData

1 / 28 requests


13.4

数据找到了. 接着看"预览"吧

年度票房

年度首周票房

2020年

	影片名称	类型
1	 八佰	战争
2	我和我的家乡	喜剧
3	姜子牙	动画
4	金刚川	剧情/战争
5	夺冠	剧情
6	拆弹专家2	动作
7	除暴	动作
8	宠爱	剧情
9	我在时间尽头等你	爱情
10	误杀	剧情
11	信条	科幻
12	紧急救援	剧情
13	叶问4：完结篇	动作

Filter

Hide data URLs AllXHRJS CSS Img Media Font Doc WS Manifest OtherHas blocked cookiesBlocked Requests

500 ms

1000 ms

1500 ms

2000 ms

2500 ms

3000 ms

3500 ms

4000 ms

4500 ms

Name

×

Headers

Preview

Response

Initiator

Timing

GetData.ashx

/API

073119D5DF1A2F0A8451E5B5C28A1AD98B6A9DE1CB7F75D5CC96E2E7B1492E7FE337EA8F2929FA94D81109F

E51348597F041D112070B254249349A40A338EEA69C11C32CE0A510210B2D5054D7972AF18D678AB98900737

20BE89DDE934AB541CC46994575693179B5BF2F38E06B3AAC85F9D4AFD5108B91048C63CC31C8EAD83C285B

ABEAFB7A18760B17C6806BAEDDB6D78CEE5169F50C40C2C45864E5E2FD3254418B5DDF34D3A4605F5229636

2BCF90979596A7348334FEC1FF47F23512F738FFAFEE485941075D005F2561C64E244360F582F4708B5A96851

761A76399063DD0CCB67F265EF41FD5CCF28E80325FE2AA6AC0D0C9190A97F15E6424FD5EAAAC25B5092FBD

E46217742C5DA9F50964A6195FDF1145ED9A4EFB4F0A1873AD348C835F8174DBD6479087027DC792A70D9A6

8CDD276FDA0C4632CE51B6E56DA236E8354C1904393BD4079568A1D9DFED0ED02FCF312185E3E8AD82F0B8

2FBF3DAD3877E80D9916E48A280DCF3EF22B242B3249D258879A944DABBE2A05E29398820FF1ED201A448

C315733AA66D474877321EE9907A0E7042D30A77E23DB0BF12C35154374B33CECF126157592ED647A00622

7109FF4FD663819318712C4F88830B34E47CF42185A5D588FF2E43FAD1142317E69449DC98AF59581CD945A

CFC48BAF87634EAA7ED654D254489A50CE4EF72717958E55884C8B713DDF45813DF32CFB7D586AB254E8876

5A6F7C9B8C29F2AA68702AB014F59F1E8BA9E93ED5CD72F207982EDA7469B128BFAD00C9F1753086FB1E

80061A90FE388512FBF7389CDDA77CD7BE00DC87C0805BCA7ADE5BC80270678EBE244F117DF8A707DC4F335

59342FEE5C9E1AFF18AD8A988E5A8895D3C9A411BC4EC100F8558F563066BA4952C9A435C15FE21646C8D58B

43191F147E0B035B7D761DD824838A4FD0609C8D666C9BD7682637F7F66989CC86969C4D8F7DB0552D8631D

82A1CC75857CC50226B1E9690B6572376088DCFB283BC0C1A50335E401491354F9D033A148B06481B045E789

055DD195A43D75C6B4CFB26C96186866DB3308FEC3B742A322EED7ECA5AD5C13C368AFB7F3A9112746BD1

FA915C406B6D958699B68F7A63F13437852F50FA37C8B389325134859F7041D112CE789A753D090334A338EE

A69C11C32C5C1D039886515A00DF4D1F8128FC94FCD22A7274FFDFBA98ABC6D08F3AAC89B6A57EC513FE6F6

0D5E8A17A6A75E38E39E235D7A4C1D0F7034356A589C17219E39A645FE36CB44CDF3411C183294F0CA88FE34

D80814412FAB05A714465A7C49795DF98D5680E1D41B65E4664F5E09FA83152145AE2DAF55B777E28FF008F

AD20DED5AD2C04F2A1FD5666880EEB87A23593A895E09483D679816D7E76EF406DF523B87D27E533AFA1FA

7894AC3EF3D2095EE0A83DED190A345AE1F11E67568CFD9090144C0F53CAF977393A51B246E5E323224EC19F6

7DAE123881EB476447433A1E8B8C6E9422F0F13B84B1836539482EC68B3E4C4134BA67C235250ED8678554076

37835789646D2A5ED7ED59F526E8023B3AD2C0944CBFBFA3906F6AAE8518BF4EA8A25CC1E0F9D76ED261D

1AD70C4EE73FD586F4505E98C69045CE7649CBF7B0FA881A532455D53A7ACF725992B45F80B0483128A98

16634345770F37CFA01076A0730C379A9EC7FD7C9C48601E80219882736831CAC3F67D2FEF9879AF416388813

99EA6B8780D4D1B527C0F4D16138245E3B1AB6DB28200B681F2973133ABDD2DF35A928F8DB7DDC20FF7A633D

我们发现这个数据是经过加密算法的. 这就头疼了. 直接通过 requests 拿到这些内容必须要解密才能看到真实数据. 但是该网站采用的加密方式又不是那么容易破解. 此时, 各位想想如果我能通过我的程序直接调用浏览器. 让浏览器去解密这些内容. 我们直接拿结果岂不妙哉. 哎~这就引出了我们本章要重点讲解的 selenium 了. 它可以完美解决上述问题

简单介绍一下 selenium, 它本身是一个自动化测试的工具. 可以启动一个全新的浏览器. 并从浏览器中提取到你想要的内容. 随着各种网站的反爬机制的出现. selenium 越来越受到各位爬sir的喜爱. selenium 最大的缺点其实就一个, 慢! 你想啊. 他要启动一个第三方的软件(浏览器), 并且还要等待浏览器把数据渲染完毕. 这个过程必然是很耗时的. 所以它慢.

接下来, 我们来聊聊 selenium 如何安装和使用.

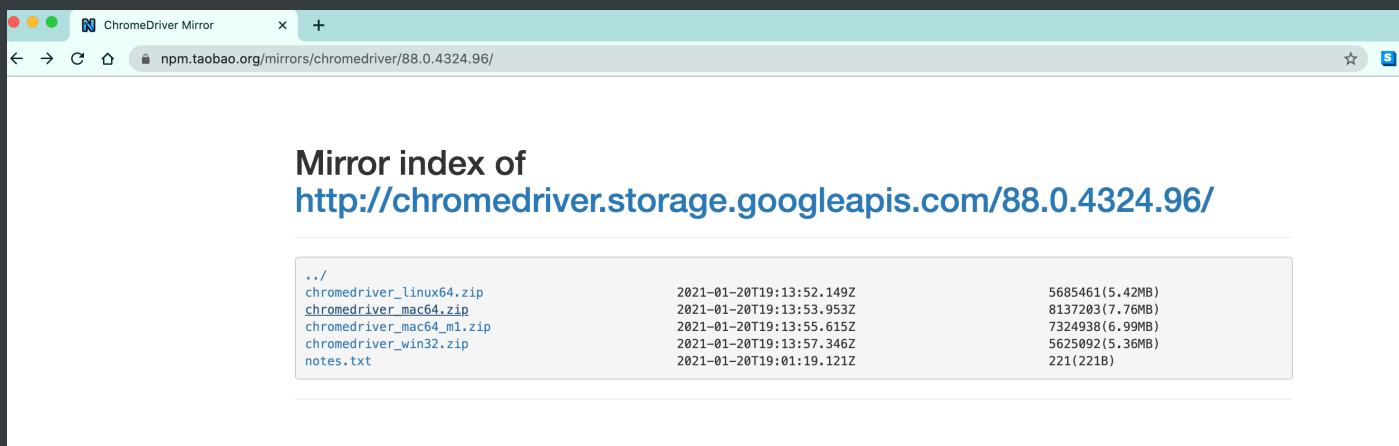
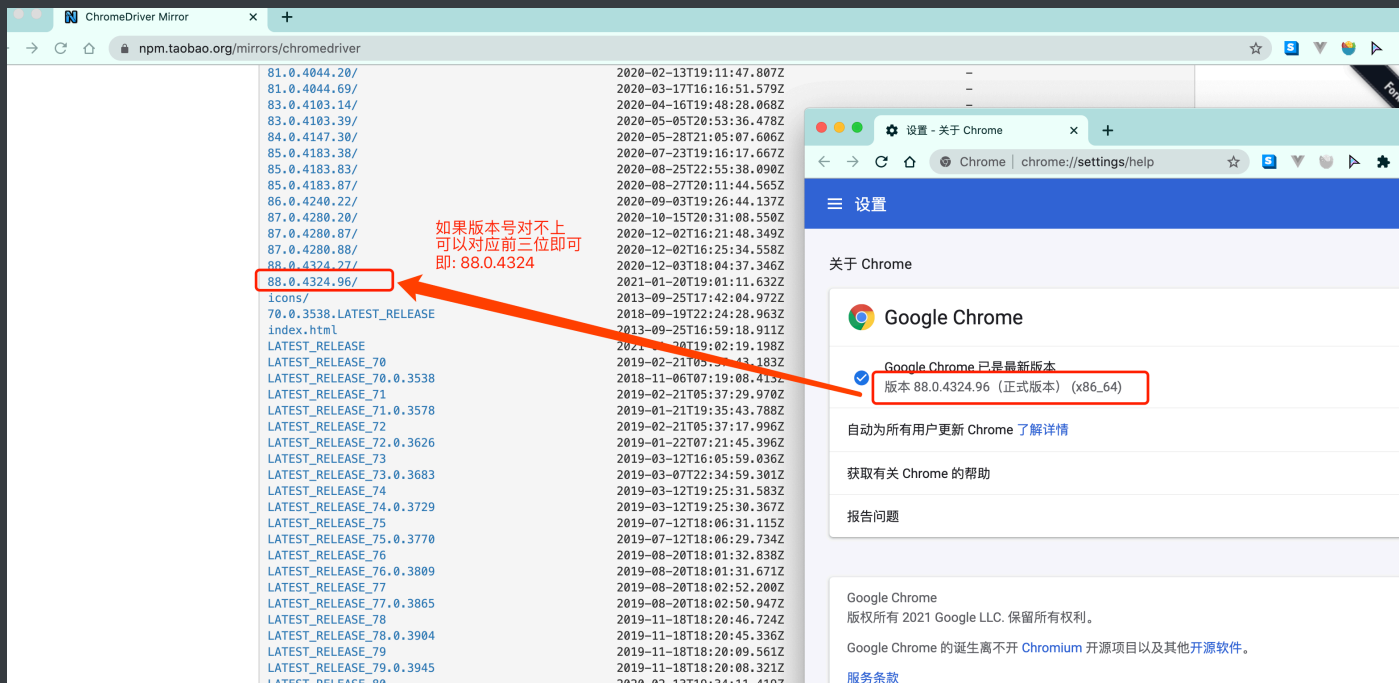
就像其他第三方库一样, selenium 直接用 pip 就可以安装了

```
1 pip install selenium
```

但是呢, 它与其他库不同的地方是他要启动你电脑上的浏览器, 这就需要有一个驱动程序来辅助.

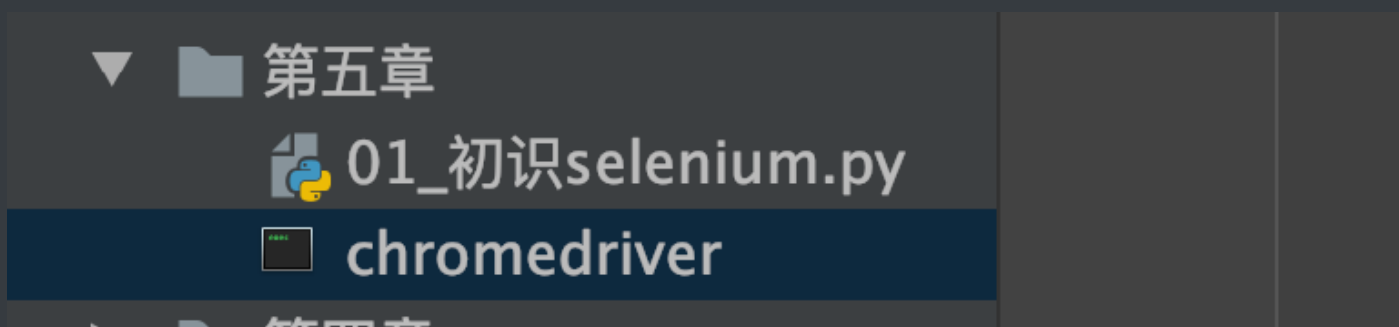
chrome 驱动地址: <https://npm.taobao.org/mirrors/chromedriver>

这里推荐用 chrome 浏览器. 其他浏览器的驱动请自行百度.

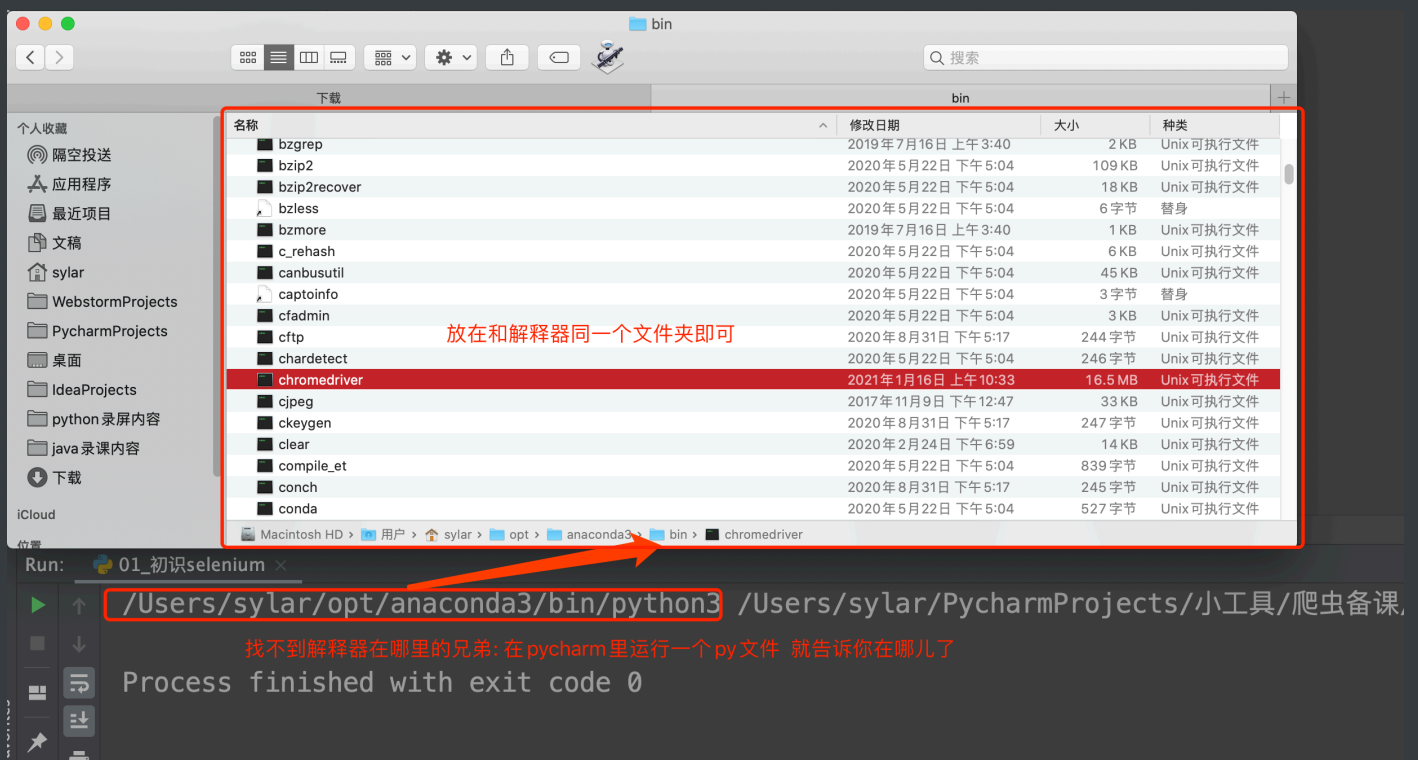


根据你电脑的不同自行选择吧. win64选win32即可.

然后关键的来了. 把你下载的浏览器驱动放在程序所在的文件夹. 或者放到python解释器所在的文件夹. 两种二选其一.



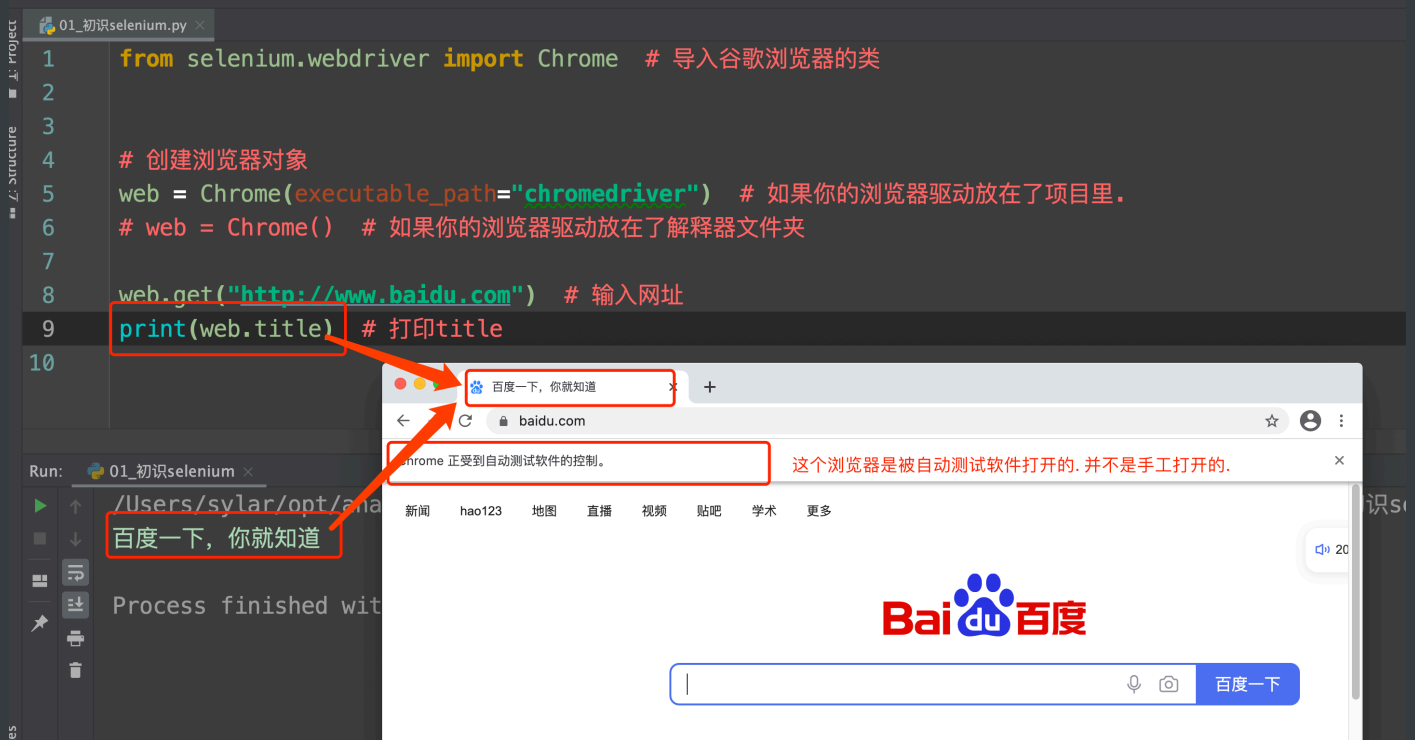
或者



OK~ 前期准备工作完毕. 上代码看看, selenium是个什么鬼

```
1 from selenium.webdriver import Chrome # 导入谷歌浏览器的类
2
3
4 # 创建浏览器对象
5 web = Chrome(executable_path="chromedriver") # 如果你的浏览器驱动放在了项目里.
6 # web = Chrome() # 如果你的浏览器驱动放在了解释器文件夹
7
8 web.get("http://www.baidu.com") # 输入网址
9 print(web.title) # 打印title
```

运行一下你会发现神奇的事情发生了. 浏览器自动打开了. 并且输入了网址. 也能拿到网页上的title标题.



cool~