

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



分布式爬虫

大纲

- 使用 Selenium + Phantoms 来抓取
- 微博接口分析
- 直接调用微博API来抓取

使用 Selenium + Phantoms

登录

最重要的是设置 User-Agent，否则无法跳转链接

```
from selenium.webdriver.common.desired_capabilities import DesiredCapabilities
```

```
user_agent = (  
    "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_8_4) " +  
    "AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.57 Safari/537.36"  
)
```

```
dcap = dict(DesiredCapabilities.PHANTOMJS)  
dcap["phantomjs.page.settings.userAgent"] = user_agent
```

```
driver = webdriver.PhantomJS(desired_capabilities=dcap)
```

输入用户名与密码



帐号登录 安全登录

邮箱/会员帐号/手机号

请输入密码

☒ 记住我 忘记密码

登录

还没有微博? [立即注册!](#)

其它登录: 淘 微博 人人 豆瓣 开心 腾讯 百度 搜狗

```
<input id="loginname"
type="text"
class="W_input " maxlength="128"
autocomplete="off"
action-data="text=邮箱/会员帐号/手机号"
action-type="text_copy"
name="username"
node-type="username" tabindex="1">
```

输入用户名与密码

为了与微博内容交互，需要使用 javascript

相关的 javascript 代码

```
document.getElementById('loginname').value='abc'
```

```
document.getElementsByName('password')[0].value='abc'
```

通过 Selenium 提供的 send_keys 来传递 value

```
driver.find_element_by_id('loginname').send_keys(username)
```

```
driver.find_element_by_name('password').send_keys(password)
```

微博 Web 图分析



关注、粉丝

关注列表、粉丝列表，作为漫游Weibo的外链

她的关注 452



贝塔斯曼龙宇 V 皇冠 女

关注 207 | 粉丝 27540 | 微博 312

地址 北京

贝塔斯曼中国总部CEO,贝塔斯曼亚洲投资基金董事总经理

通过 [微博搜索](#) 关注

+ 关注

更多 ▾



农家石嫣 V 女

关注 184 | 粉丝 56579 | 微博 12060

地址 北京 海淀区

国际社区支持农业联盟URGENCEI副主席，分享收获CSA创始人

通过 [微博搜索](#) 关注

+ 关注

更多 ▾



达沃斯DAVOS V 男

关注 386 | 粉丝 102650 | 微博 6242

地址 海外

达沃斯世界经济论坛

通过 [微博搜索](#) 关注

+ 关注

更多 ▾

获得微博外链

```
driver.find_element_by_xpath('//a[@class="t_link S_txt1"]')
```

0 是关注，1是粉丝，2是微博，我们只需要 0，关注的微博一般是有质量的，而粉丝的数量太多，并且有太多僵尸粉

打开关注列表页：

```
driver.find_element_by_xpath('//a[@class="t_link S_txt1"]').get_attribute('href')
```

获取所有关注的微博号的地址：

```
driver.find_elements_by_xpath('//*[contains(@class, "follow_item")]//a[@class="S_txt1"]')
```

获取微博用户信息

- 提取用户的基本信息
 - 链接：用正则表达式把用户的链接参数都去掉
`/u/1634431184?refer_flag=1005050006_`
 - 微博昵称及头像
 - 关注、粉丝及微博数量
- 过滤质量差的用户。对于微博数量少于阈值，或者关注数超过粉丝数 N 倍以上的，判定为僵尸粉或广告微博，直接跳过
 - 僵尸粉：微博数量极少
 - 纯广告、营销微博：关注数远远超过粉丝数量
- 提取下一页，可以继续查找更多的user

获得微博外链

```
driver.find_element_by_xpath('//a[@class="t_link S_txt1"]')
```

0 是关注，1是粉丝，2是微博，我们只需要 0，关注的微博一般是有质量的，而粉丝的数量太多，并且有太多僵尸粉

打开关注列表页：

```
driver.find_element_by_xpath('//a[@class="t_link S_txt1"]').get_attribute('href')
```

获取所有关注的微博号的地址：

```
driver.find_elements_by_xpath('//*[@contains(@class, "follow_item")]//a[@class="S_txt1"]')
```

微博信息抽取

微博名: `driver.find_element_by_tag_name('h1')`

所有的Feed: `driver.find_elements_by_class_name('WB_detail')`

```
feed = {}
```

```
feed['time'] = element.find_element_by_xpath('..div[@class="WB_from S_txt2"]').text
```

```
feed['content'] = element.find_element_by_class_name('WB_text').text
```

```
feed['image_names'] = []
```

```
for image in element.find_elements_by_xpath('..li[contains(@class,"WB_pic")]/img'):
```

```
    feed['image_names'].append(re.findall('/(^[^/]+)$', image.get_attribute('src')))
```

微博的图片，只需要保存图片名

<http://wx2.sinaimg.cn/thumb150/4b7a8989ly1fcws2sryvrj22p81sub2a.jpg>

<http://存储域名/分辨率/文件名>

微博图片信息

```
re.findall('(/[^[^/]+)$', image.get_attribute('src'))
```

微博的图片，只需要保存图片名

<http://wx2.sinaimg.cn/thumb150/4b7a8989ly1fcws2sryvrj22p81sub2a.jpg>

<http://存储域名/分辨率/文件名>

名称	宽度	定义
thumb150	150 像素	缩略图
mw690	690 像素	中图
mw1024	1024像素	大图

滚频与翻页

每次滚动后，检查是否已经出现了

- 微博的下一页的 class:

点击重新载入

```
page next S_txt1 S_line1
```

```
driver.find_element_by_xpath('//a[@class="page next S_txt1 S_line1"]').click()
```

- 翻页命令

```
driver.execute_script('window.scrollTo(0, document.body.scrollHeight)')
```

滚屏与翻页

每次滚动后，检查是否已经出现了“下一页”的按钮，如果是则可以停止翻页，否则检查是否出现了“网络超时”的链接，是的话，点击这个链接来重新加载

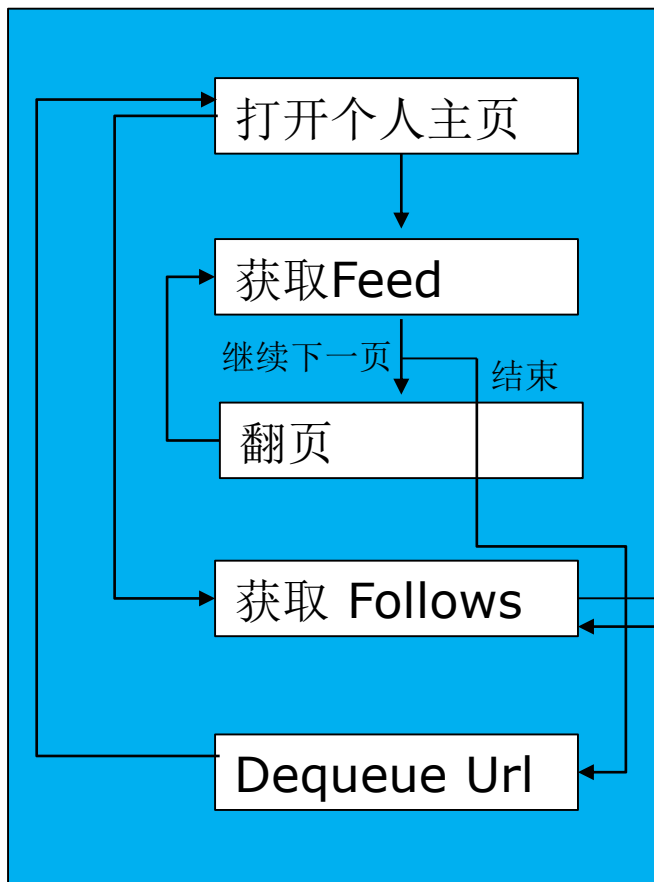


滚屏

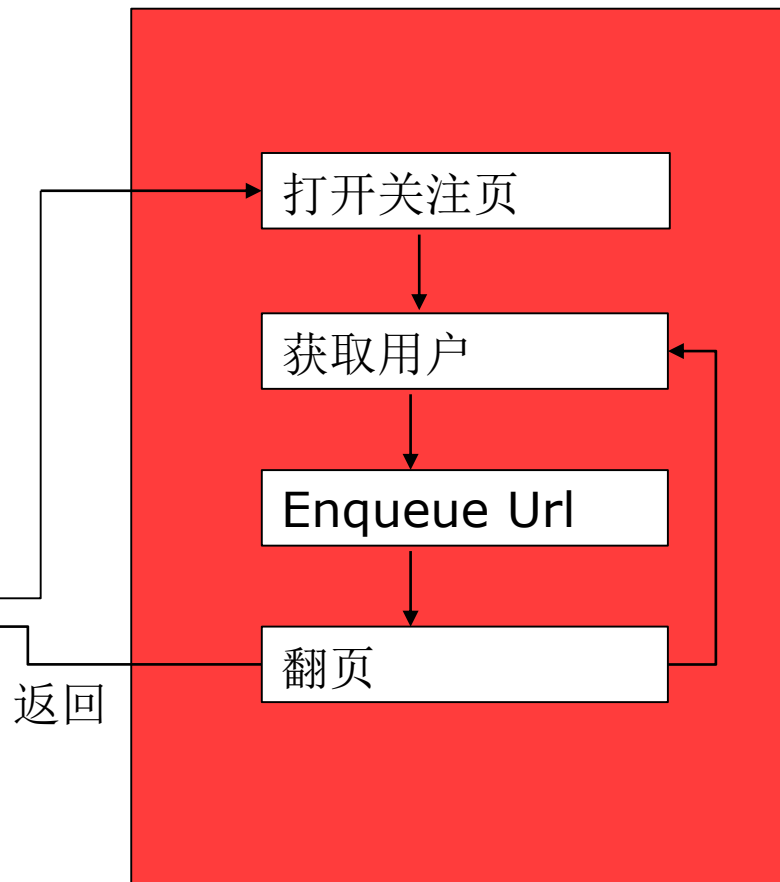
```
for i in range(0,10):  
    driver.execute_script('window.scrollTo(0, document.body.scrollHeight)')  
    html = driver.page_source  
    tr = etree.HTML(html)  
    next_page_url = tr.xpath('//a[contains(@class,"page next")]')  
    if len(next_page_url) > 0:  
        return next_page_url[0].get_attribute('href')  
    if len(re.findall('点击重新载入', html)) > 0:  
        driver.find_element_by_link_text('点击重新载入').click()
```

微博抓取框架

Web 1 : Crawler



Web 2: User Info



微博接口分析

微博域名

http://m.weibo.cn

这是微博的首页网页版，能看到结构非常简单，我们能直接拿到 **Feed** 流，我们尝试从移动端来分析微博的数据API 接口



个人首页

<https://m.weibo.cn/u/1266321801?uid=1266321801&luicode=10000011&lfid=100103type%3D1%26q%3D%E5%A7%9A%E6%99%A8&featurecode=20000320>

上面的是姚晨的微博个人主页，包含的参数列表：

uid、luicode、lfid、featurecode

我们并不清楚每个参数的含义，但是很明显，

1266321801 是她的微博ID

可以直接尝试

<https://m.weibo.cn/u/1266321801>



个人feed流

Name

- ☐ content.min.css
- ☐ getIndex?type=uid&value=1266321801&containerid=1076031266321801
- ☐ getIndex?type=uid&value=1266321801&containerid=1076031266321801

ajax 请求

x Headers Preview Response Cookies Timing

▼ General

Request URL: https://m.weibo.cn/api/container/getIndex?type=uid&value=1266321801&containerid=1005051266321801

Request Method: GET

Status Code: 200 OK

Remote Address: 180.149.139.248:443

Referrer Policy: no-referrer-when-downgrade

▼ Response Headers view source

Connection: keep-alive

Content-Encoding: gzip

Content-Type: application/json; charset=utf-8

Date: Sun, 11 Jun 2017 12:16:57 GMT

PROC_NODE: web-v8-005.mweibo.yz.sinanode.com

Server: Tengine/2.2.0

Set-Cookie: WEIBOCN_FROM=deleted; expires=Thu, 01-Jan-1970 00:00:01 GMT; Max-Age=0; path=/; domain=.weibo.cn

https://m.weibo.cn/api/container/getIndex?type=uid&value=1266321801&containerid=1076031266321801

- type: 通过uid方式查询
- value: user id
- containerid: 容器的ID号 = 107603 + uid

个人 Feed 流翻页

向下滚动，获取更多feed流，观察新的请求

Name	× Headers Preview Response Cookies Timing
<input type="checkbox"/> content.min.css	
<input type="checkbox"/> getIndex?type=uid&value=1266321801&container...	
<input type="checkbox"/> getIndex?type=uid&value=1266321801&container...	
<input type="checkbox"/> config	
<input type="checkbox"/> getIndex?type=uid&value=1266321801&container...	<div><div>▼ General</div><div>Request URL: https://m.weibo.cn/api/container/getIndex?type=uid&value=1266321801&containerid=1076031266321801&page=2</div><div>Request Method: GET</div><div>Status Code: 200 OK</div><div>Remote Address: 180.149.153.242:443</div><div>Referrer Policy: no-referrer-when-downgrade</div></div>
<input type="checkbox"/> config	<div><div>▼ Response Headers view source</div><div>Connection: keep-alive</div><div>Content-Encoding: gzip</div><div>Content-Type: application/json; charset=utf-8</div><div>Date: Sun, 11 Jun 2017 12:24:05 GMT</div><div>PROC_NODE: web-v8-132.mweibo.yz.sinanode.com</div><div>Server: Tengine/2.2.0</div></div>

https://m.weibo.cn/api/container/getIndex?type=uid&value=1266321801&containerid=1076031266321801&page=2

https://m.weibo.cn/api/container/getIndex?type=uid&value=1266321801&containerid=1076031266321801

- type: 通过uid方式查询
- value: user id
- containerid: 容器的ID号 = 107603 + uid
- page: 当前请求的页码

点击打开个人关注页



主页 微博



@杨子姗: 办婚礼了, 我和吴中天。😄



推荐 她的关注 她的粉丝

好友关注动态



遂昌快活林

浙江省遂昌金矿有限公司经济师...

获取第一手投资理财信息



尖叫耐撕男女

微博知名视频博主

我们都爱讲笑话



沸点观察

知名财经博主 微博股评师 微博签..

获取第一手投资理财信息



姚志远要学雷锋

微博资深汽车达人 微博汽车视频...

一起交流买车养车心得



无敌可爱哦

#world6#超级话题小主持人

最in游戏玩家



西威Civet

果壳网艺术领域达人

来普及科学知识



有三类关注:

- 推荐
- 她的关注
- 她的粉丝

点击打开个人关注页

Name	×	Headers	Preview	Response	Cookies	Timing
<input type="checkbox"/> getIndex?containerid=231051_-_followersrecomm_-_126632...		General Request URL: https://m.weibo.cn/api/container/getIndex?containerid=231051_-_followers_-_1266321801&luicode=10000011&lfid=1005051266321801&featurecode=20000320&type=uid&value=1266321801 Request Method: GET Status Code: 200 OK Remote Address: 180.149.153.216:443 Referrer Policy: no-referrer-when-downgrade				
<input type="checkbox"/> getIndex?containerid=231051_-_followersrecomm_-_126632...						
<input type="checkbox"/> getIndex?containerid=231051_-_followers_-_1266321801&luicode=10000011&lfid=1005051266321801&featurecode=20000320&type=uid&value=1266321801						
<input type="checkbox"/> config						
<input type="checkbox"/> getIndex?containerid=231051_-_fans_-_1266321801&luicode=10000011&lfid=1005051266321801&featurecode=20000320&type=uid&value=1266321801		Response Headers view source Connection: keep-alive Content-Encoding: gzip Content-Type: application/json; charset=utf-8				

3个接口

关注列表:

https://m.weibo.cn/api/container/getIndex?containerid=231051_-_followers_-_1266321801&luicode=10000011&lfid=1005051266321801&featurecode=20000320&type=uid&value=1266321801

推荐列表:

https://m.weibo.cn/api/container/getIndex?containerid=231051_-_followersrecomm_-_1266321801&luicode=10000011&lfid=1005051266321801&featurecode=20000320&type=uid&value=1266321801&since_id=1497184436%7C20_1

粉丝列表:

https://m.weibo.cn/api/container/getIndex?containerid=231051_-_fans_-_1266321801&luicode=10000011&lfid=1005051266321801&featurecode=20000320&type=uid&value=1266321801

疑问

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

