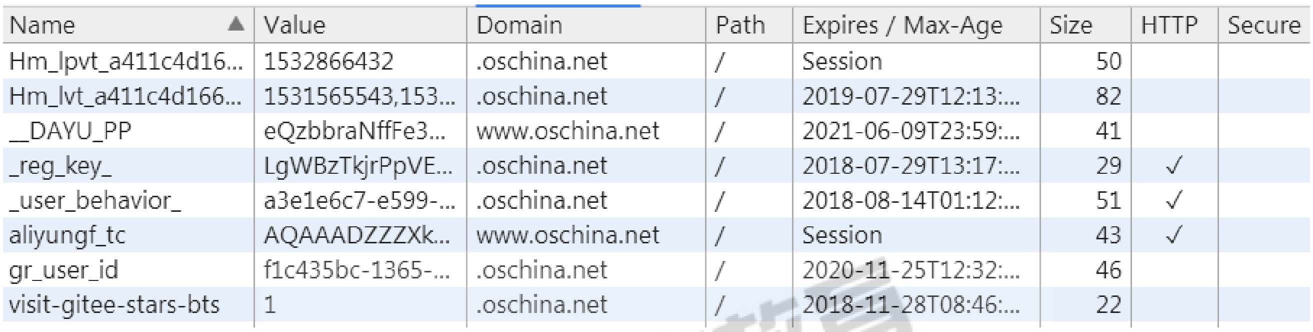
模拟登陆oschina（新浪）

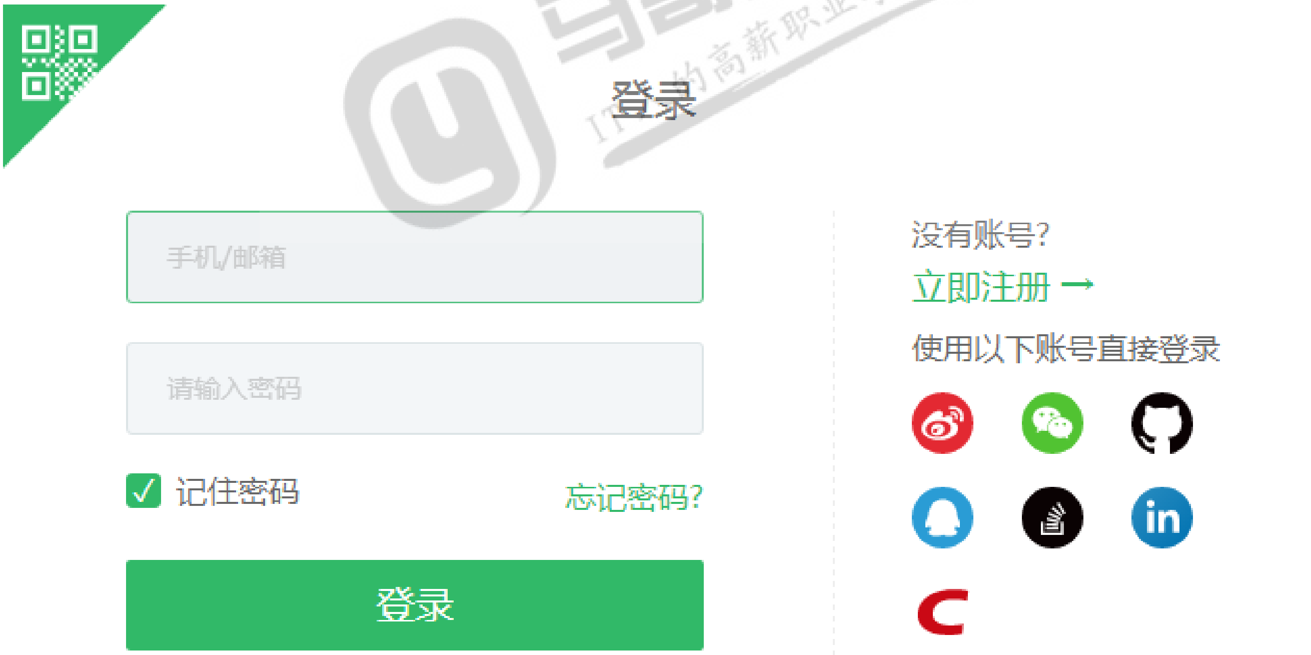
一般登录后，用户就可以一段时间内可以使用该用户身份操作，不需要频繁登录了。这背后往往使用了Cookie技术。
登录后，用户获得一个cookie值，这个值在浏览器当前会话中保存，只要不过期甚至可以保存很久。
用户每次向服务器提交请求时，将这些Cookie提交到服务器，服务器经过分析Cookie中的信息，以确认用户身份，确认是信任的用户身份，就可以继续使用网站功能。

Cookie
网景公司发明。cookie一般是一个键值对name=value，但还可以包括expire过期时间、path路径、domain域、secure安全等信息。



Name ▲	Value	Domain	Path	Expires / Max-Age	Size	HTTP	Secure
Hm_lpvt_a411c4d16...	1532866432	.oschina.net	/	Session	50		
Hm_lvt_a411c4d166...	1531565543,153...	.oschina.net	/	2019-07-29T12:13:...	82		
__DAYU_PP	eQzbbraNffFe3...	www.oschina.net	/	2021-06-09T23:59:...	41		
_reg_key_	LgWBzTkjrPpVE...	.oschina.net	/	2018-07-29T13:17:...	29	✓	
_user_behavior_	a3e1e6c7-e599-...	.oschina.net	/	2018-08-14T01:12:...	51	✓	
aliyungf_tc	AQAAADZZZXk...	www.oschina.net	/	Session	43	✓	
gr_user_id	f1c435bc-1365-...	.oschina.net	/	2020-11-25T12:32:...	46		
visit-gitee-stars-bts	1	.oschina.net	/	2018-11-28T08:46:...	22		

清空oschina.net的所有cookies，重新登录，勾选"记住密码"。



```
使用wei.xu@magedu.com/magedu.com18登录oschina后，HTTP请求头如下

Accept:text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Accept-Encoding:gzip, deflate
```

```
Accept-Language:zh-CN
Cache-Control:no-cache
Connection:keep-alive
Cookie:aliyungf_tc=AQAAABGJsEw8KgYA3Fj2cgAxjv0siYI4; _user_behavior_=4d327315-a3e1-4db5-9653-
79fc9640b1dc;
oscid=ZV2oveUqo28xv80qumQtfRqukWzpKq2brNqjn0Y0a5kFTeUQUUbcPj2dwLIiVt%2FusptVh3rtTPug%2B42Y%2FtTx
Q27Y6jEJ5FAls%2BjV0jPXPwaripdfSXBaumHSPFRAUipThiNBOrmH7B%2BTEOAcGyx4CLF%2Fzg8l36%2FA0ZZEWRnEvpU%
3D
DNT:1
Host:www.oschina.net
Pragma:no-cache
Referer:https://www.oschina.net/home/login
Upgrade-Insecure-Requests:1
User-Agent:Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Maxthon/5.0
Chrome/55.0.2883.75 Safari/537.36
X-DevTools-Emulate-Network-Conditions-Client-Id:a638e770-f2be-4986-8ddb-2b40d75f5abd
```

对比登录前后的cookie值，发现登录后有oscid。

那就把这个HTTP 请求头放在代码中。

```python
import requests

url = 'https://my.oschina.net/'

headers = {
    'Host': "www.oschina.net",
    'Connection': "keep-alive",
    'Pragma': "no-cache",
    'Cache-Control': "no-cache",
    'X-DevTools-Emulate-Network-Conditions-Client-Id': "a638e770-f2be-4986-8ddb-2b40d75f5abd",
    'Upgrade-Insecure-Requests': "1",
    'User-Agent': "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko)
Maxthon/5.0 Chrome/55.0.2883.75 Safari/537.36",
    'Accept': "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8",
    'DNT': "1",
    'Referer': "https://www.oschina.net/home/login",
    'Accept-Encoding': "gzip, deflate",
    'Accept-Language': "zh-CN",
    'Cookie': "aliyungf_tc=AQAAABGJsEw8KgYA3Fj2cgAxjv0siYI4; _user_behavior_=4d327315-a3e1-4db5-
9653-79fc9640b1dc;
Hm_lvt_a411c4d1664dd70048ee98afe7b28f0b=1532283265,1532536915,1532866432,1532867986;
Hm_lpvt_a411c4d1664dd70048ee98afe7b28f0b=1532868325; _reg_key_=50QBpSefZ587WIOAWnyP;
oscid=ZV2oveUqo28xv80qumQtfRqukWzpKq2brNqjn0Y0a5kFTeUQUUbcPj2dwLIiVt%2FusptVh3rtTPug%2B42Y%2FtTx
Q27Y6jEJ5FAls%2BjV0jPXPwbY%2Bw3K%2FN2OBE5Odc%2FCbQprhiNBOrmH7B%2BTEOAcGyx4CLF%2Fzg8l36%2FA0ZZEWR
nEvpU%3D",
    }
# headers = {
#     'User-Agent': "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko)
Maxthon/5.0 Chrome/55.0.2883.75 Safari/537.36",
#     }

# 使用session
```

```
response = requests.request("GET", url, headers=headers)

with response:
    with open('o:/profile.html', 'w', encoding='utf-8') as f:
        text = response.text
        f.write(text)
        print(text) # 搜索 user-info
        print(response.status_code, '~~~~~~~~~~~~~~~~~~~~')
```

已登录访问首页，右上角会有用户信息，如下

```
<div class="box-fr user-bar">
    <div class="box user-info">
        <span class="name">magedu_wayne</span>，您好&nbsp;
            <div id="site_header_tooltip"></div>
            <div class="user-menu split menu-drop user-menu">
                <a id="MySpace" href="https://my.oschina.net/u/3881396" data-tooltips-model="my_space">
                    我的空间
                    <i class="icon-svg icon-arr-down-white"></i>
                </a>
                <div class="menu-drop-down">
                    <ul class="drop-list myspace">
                        <li class="msg"><a href="https://www.oschina.net/home/go?page=admin%2Finbox">我的私信</a></li>
                        <li class="discuss"><a href="https://www.oschina.net/home/go?page=%3ftab=activity">我的讨论记录</a></li>
                        <li class="code"><a href='https://www.oschina.net/code/list_by_user?id=3881396'>我分享的代码</a></li>
                        <li class="blog"><a href="https://my.oschina.net/u/3881396">我的博客</a></li>
                        <li class="friends"><a href="https://www.oschina.net/home/go?page=following">我关注的人</a></li>
                        <li class="favorites"><a href="https://www.oschina.net/home/go?page=favorites">我的收藏夹</a></li>
                        <li class="profile"><a href="https://www.oschina.net/home/go?page=admin%2Fprofile">个人资料修改</a></li>
                    </ul>
                </div>
            </div>
            <a class="user-menu split sm-hide" href="https://www.oschina.net/home/go?page=admin%2Fnew-project">添加软件</a>
            <a class="user-menu split sm-hide" href="https://www.oschina.net/home/go?page=admin%2Fnew-release">投递新闻</a>
            <a href="/action/user/logout?session=b154d9256999699302&goto_page=https%3A%2F%2Fwww.oschina.net%2F">退出</a>
    </div>
</div>
```

未登录访问首页，右上角显示登录、注册链接，如下

```
<div class="box-fr user-bar">
    <div class="box user-info">
        [ <a href='https://www.oschina.net/home/login?
goto_page=https%3A%2F%2Fwww.oschina.net%2F'>登录</a> | <a
 href='https://www.oschina.net/home/reg?goto_page=https%3A%2F%2Fwww.oschina.net%2F'>注册</a> ]
    </div>
</div>
```

新浪微博等都一样，只要允许记住用户登录，就可以通过上述方法登录后爬取内容。

多线程爬取博客园

博客园的新闻分页地址https://news.cnblogs.com/n/page/10/，多线程成批爬取新闻的**标题和链接**。

`https://news.cnblogs.com/n/page/2/`，这个url中变化的是最后的数字一直在变，它是页码

```
from concurrent.futures import ThreadPoolExecutor
import requests
import logging
from queue import Queue
import threading
from bs4 import BeautifulSoup
import time

FORMAT = "%(asctime)s %(threadName)s %(thread)d %(message)s"
logging.basicConfig(format=FORMAT, level=logging.INFO)

# 'https://news.cnblogs.com/n/page/10/'
BASE_URL = 'https://news.cnblogs.com'
NEWS_PAGE = '/n/page/'
headers = {
    'User-Agent': "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko)
Maxthon/5.0 Chrome/55.0.2883.75 Safari/537.36",
}

# 使用池，以后可以使用第三方消息队列完成
urls = Queue() # url的队列
htmls = Queue() # 响应数据队列
outputs = Queue() # 结果输出队列

event = threading.Event()

# 创建博客园的新闻urls，每页30条新闻
def create_urls(start, end, step=1):
    for i in range(start, end + 1, step):
        urls.put('{}{}{}/'.format(BASE_URL, NEWS_PAGE, i))
    print('url创建完毕')


# 爬取页面线程函数
```

```python
def crawler():
    while not event.is_set():
        try:
            url = urls.get(True, 1)
            with requests.get(url, headers=headers) as response:
                html = response.text
                htmls.put(html)
                print(url)
        except:
            pass


# 解析线程函数
def parser():
    while not event.is_set():
        try:
            html = htmls.get(True, 1)
            soup = BeautifulSoup(html, 'lxml')
            titles = soup.select('h2.news_entry a')
            for title in titles:
                # <a href="/n/602987/" target="_blank">特斯拉推最新生活方式产品：1500美元冲浪板</a>
                val = title.text, BASE_URL + title.get('href')
                outputs.put(val)
                print(val)
        except:
            pass


# 持久化线程函数
def save(path):
    with open(path, 'a+') as f:
        while not event.is_set():
            try:
                text, url = outputs.get(True, 1)
                print(text, url, '~~~~~~~~~')
                f.write('{} {}\n'.format(text, url))
                f.flush()
            except:
                pass


# 线程池
executor = ThreadPoolExecutor(10)

executor.submit(create_urls, 1, 10)
executor.submit(parser)
executor.submit(save, 'o:/news.txt')
for i in range(7):
    executor.submit(crawler)


while True:
    inp = input('>>>')
```

```
    if inp.strip() == 'quit':
        event.set()
        print('closing ......')
        time.sleep(4)
        break
```