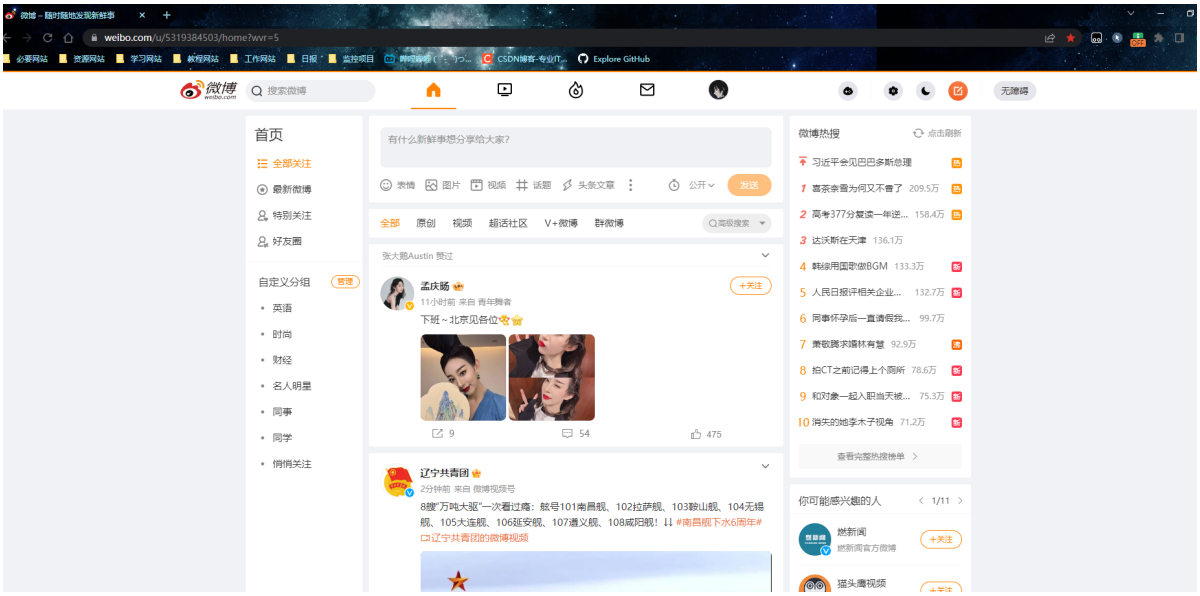
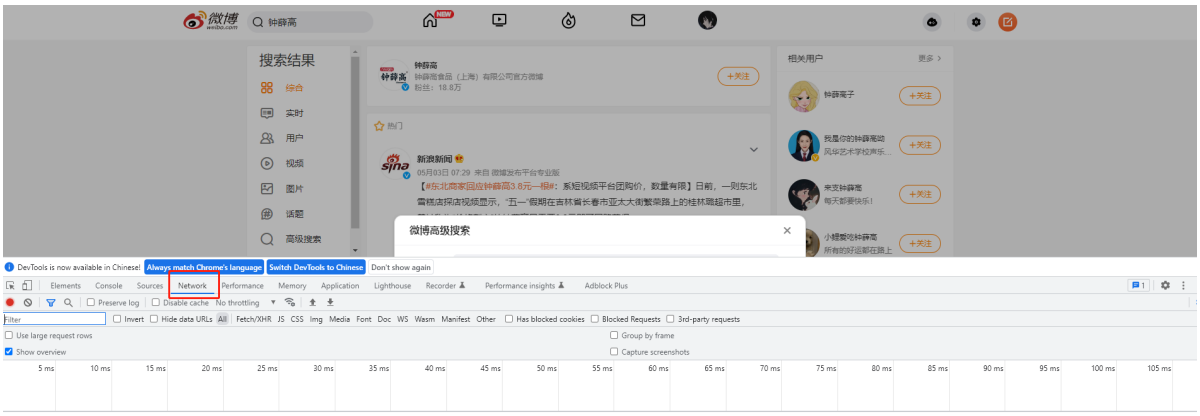
爬虫流程说明

目标网站微博:https://weibo.com/u/5319384503/home?wvr=5

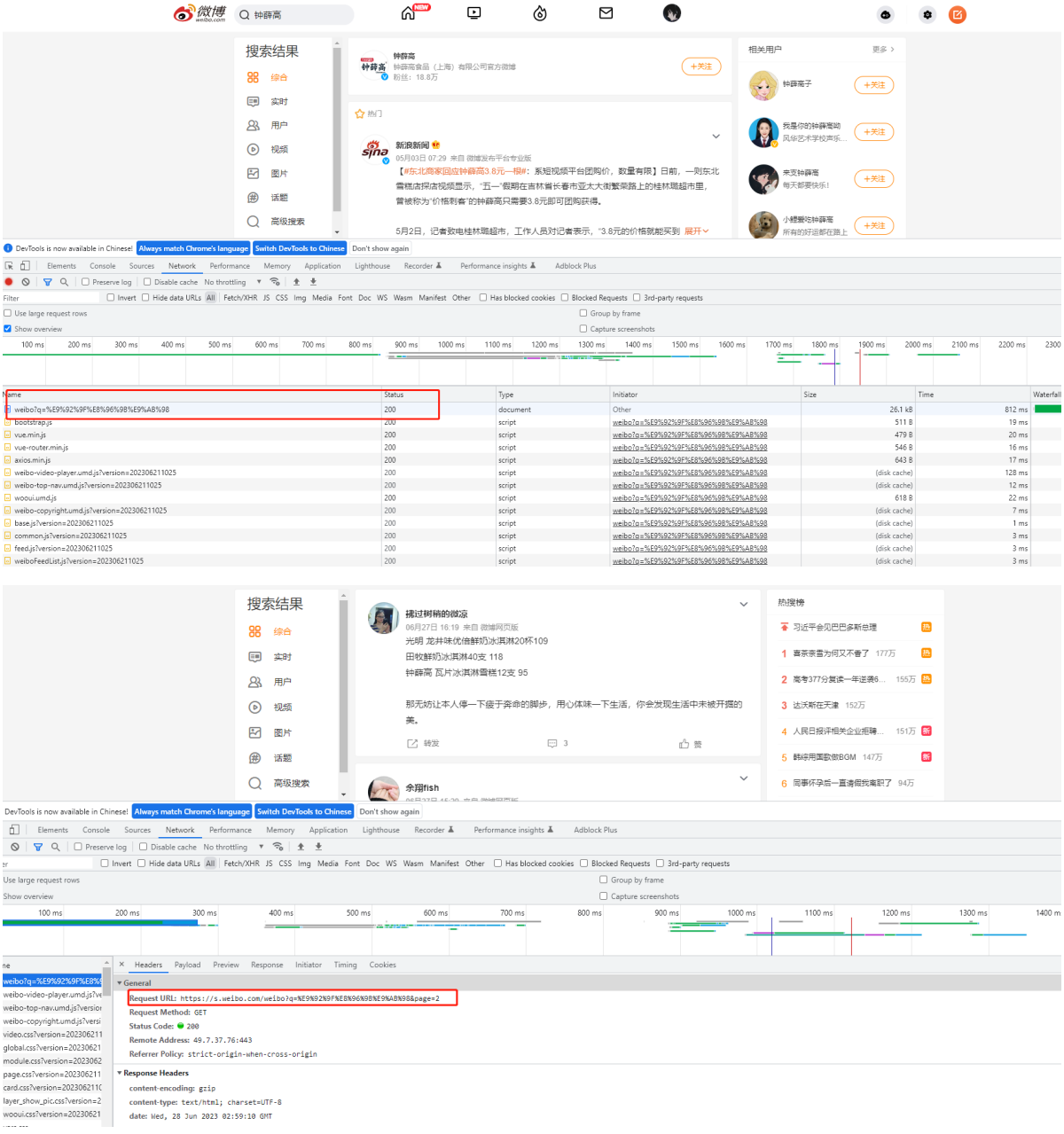


微博是需要先登录才能去获取数据的，登录之后，打开搜索列表

选择对应的时间信息，然后按F12，出现此页面



然后刷新一下页面，找到第一个



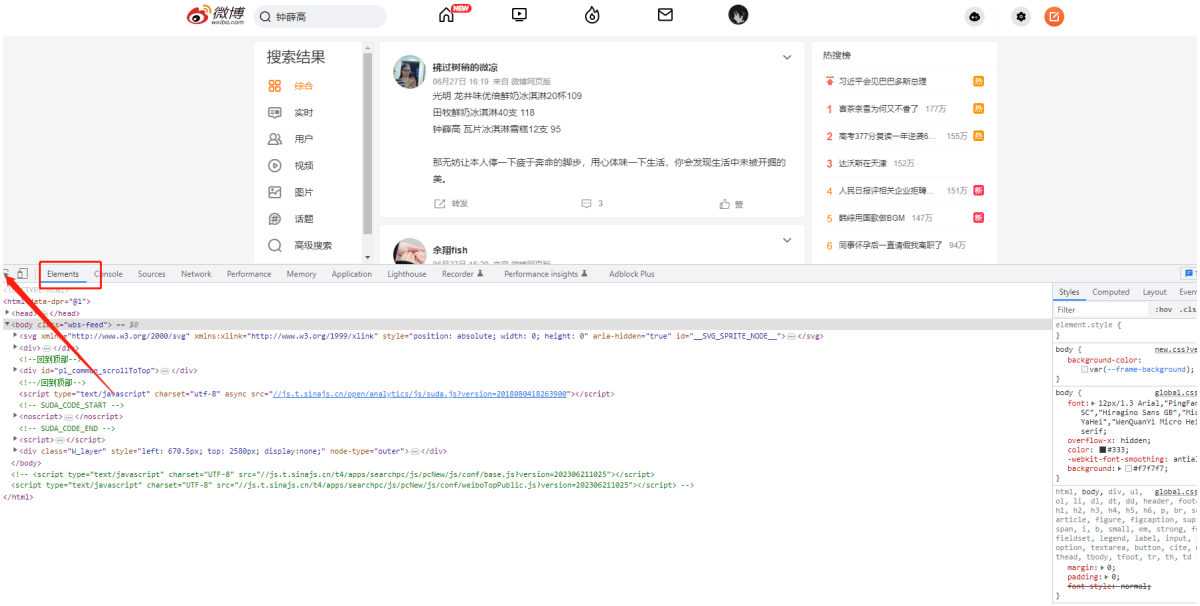
这样我们就获取对应的API

以及下面的COOKIE值

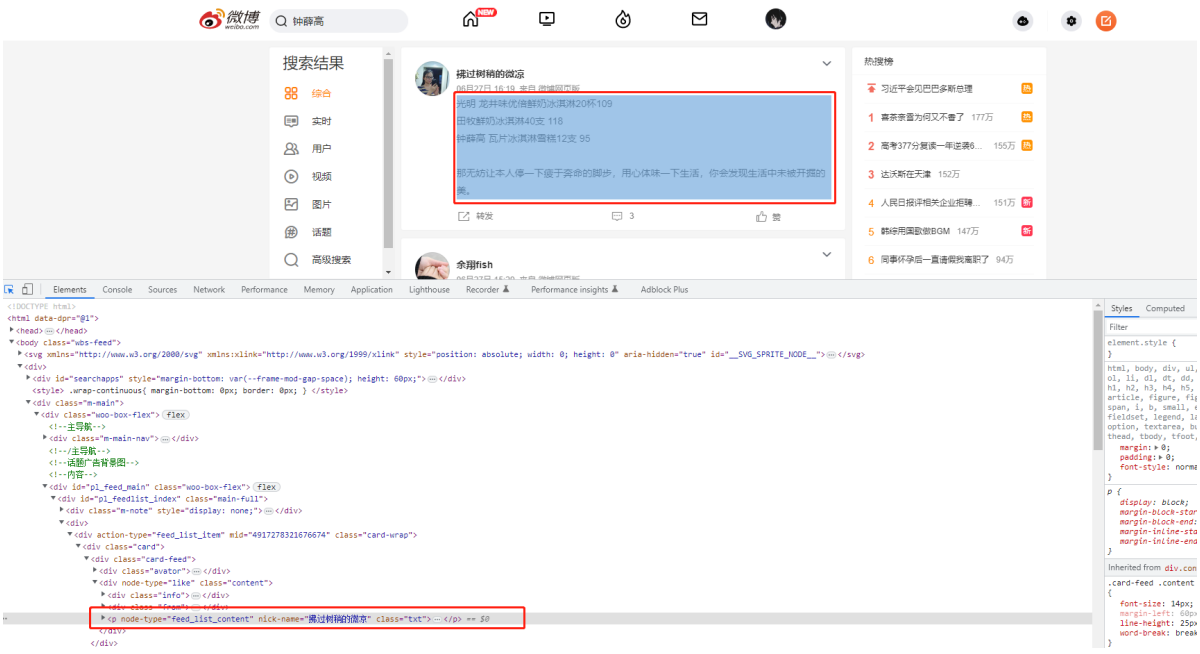
我们根据这里获取的API去构建，每一天对应的API,构建代码如下：

```
rng = pd.date_range(start='5/31/2022', end='7/31/2022')
list_time = []
for r in rng:
    r = str(r).split(" ")
    list_time.append(r[0])
for l in tqdm(range(len(list_time)-1)):
    for i in range(1,51,1):
        url = 'https://s.weibo.com/weibo?q=%E9%92%9F%E8%96%99&typeall=1&suball=1&timescope=custom:{}:{}&Refer=g&page={}'.format(list_time[l],list_time[l+1],i)
        main_url(url)
```

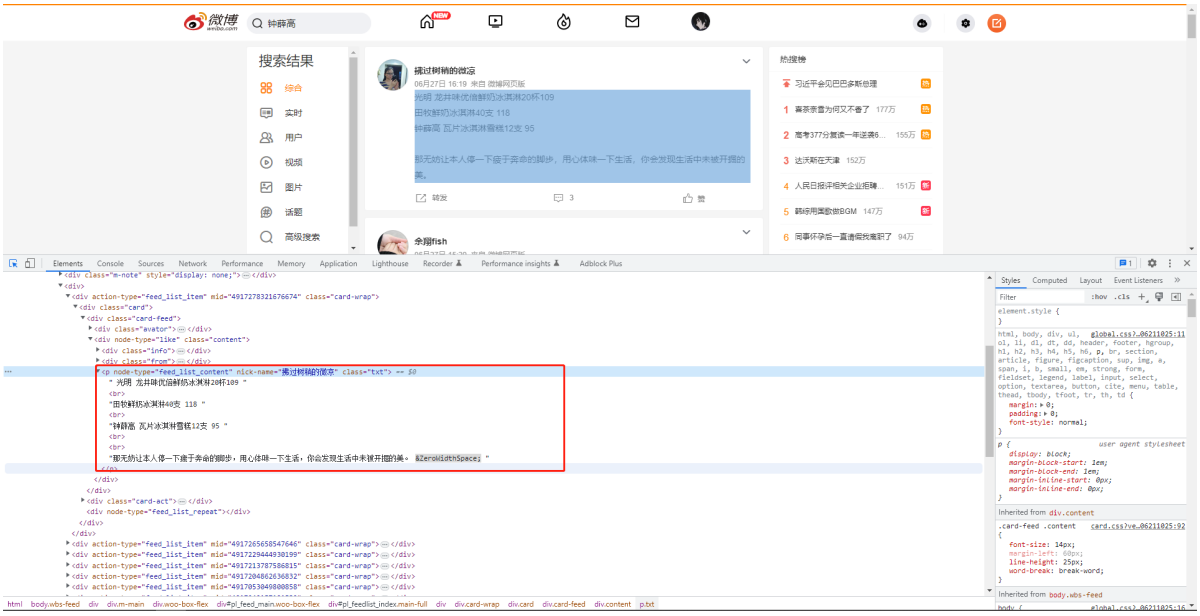
通过这样我们就获取每一个，50页的内容，因为微博，最多只能呈现50页而已，所以我们就也构建每一天的50页，来获取全部时间的内容，在获取整个页面信息后，我们还需定位来获取每个对应的信息



通过移动鼠标箭头，我们可以定位到对应div值



然后来获取全部信息内容：



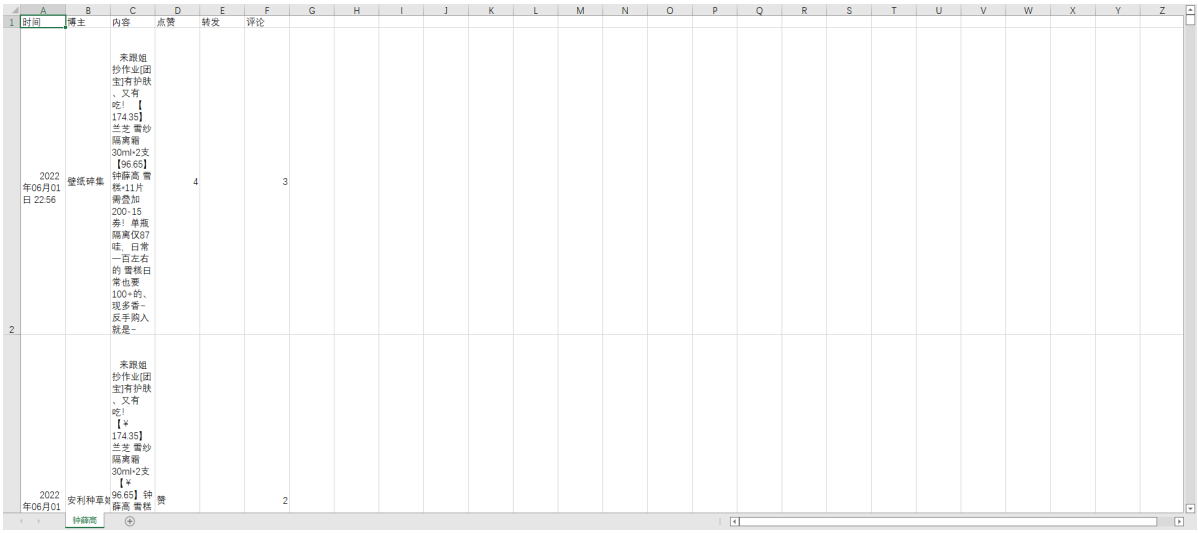
这里采用的是XPATH语法定位规则来获取内容：

具体实现代码如下：

```
soup = etree.HTML(content)
node = soup.xpath('//div[@class="card-feed"]')
act = soup.xpath('//div[@class="card-act"]/ul')
for n,a in zip(node,act):
        try:
            name = n.xpath('./div[@node-type="like"]/div/div[2]/a/@nick-name')[0]

        except:
            name = np.NaN
        # try:
        #     title = n.xpath('./div[@class="avator"]/a/span/@title')[0]
        #
        # except:
        #     title = np.NaN
        try:
            timedate = n.xpath('./div[@node-type="like"]/div[2]/a/text()')[0]
        except:
            timedate = np.NaN
        try:
            comtent = n.xpath('./div[@node-type="like"]/p[@node-
type="feed_list_content"]/text()')
            comtent1 = ' '.join(comtent)
        except:
            comtent1 = np.NaN
        try:
            dianzan = a.xpath('./li[3]/a/button/span[2]/text()')[0]

        except:
            dianzan =  np.NaN
        try:
            zhuanfa = a.xpath('./li[1]/a/text()')[0]

        except:
```

```
            zhuanfa =  np.NaN
        try:
            pinglun = a.xpath('./li[2]/a/text()')[0]

        except:
            pinglun =  np.NaN
        df = pd.DataFrame()
        df['时间'] = [timedate]
        df['博主'] = [name]
        df['内容'] = [comtent1]
        df['点赞'] = [dianzan]
        df['转发'] = [zhuanfa]
        df['评论'] = [pinglun]
        df.to_csv('钟薛高.csv', index=None, header=None, mode='a+', encoding='utf-8-sig')
    time.sleep(0.2)
```

最后的输出结果如下：



	A	B	C	D	E	F
1	时间	博主	内容	点赞	转发	评论
2	2022年06月01日 22:56	壁纸碎集	来跟姐抄作业[团宝]有护肤、又有吃！【174.35】兰芝 雪纱隔离霜30ml+2支【96.65】钟薛高 雪糕*11片 需叠加200-15券！单瓶隔离仅87哇。日常一百左右的 雪糕日常也要100+的、现多省~反手购入就是~	4		3
	2022年06月01	安利种草姐	来跟姐抄作业[团宝]有护肤、又有吃！【¥174.35】兰芝 雪纱隔离霜30ml+2支【¥96.65】钟薛高 雪糕	赞		2

	A	B	C	D	E	F
20537	今天 20:27	-高学历疯	私信收到的大吓人了	5542		420
20538	今天 20:15	萧敬腾	大家好 我是萧敬腾 今天想跟大家分享我的喜悦 分享始我爱的人及爱我的人 我知道這很突然 但我想讓大家知道 我很幸福 我真的很幸福 我有音樂 我有籃球 我有好友 我有愛我的家人 我有支持我的你們 我擁有了好多好多 現在 我擁有了愛情 一個真正	144622		30558
20539	今天 20:12	他的同学	这篇文章教你如何不痛苦地早起。	11247		81

一共是2万多条数据内容