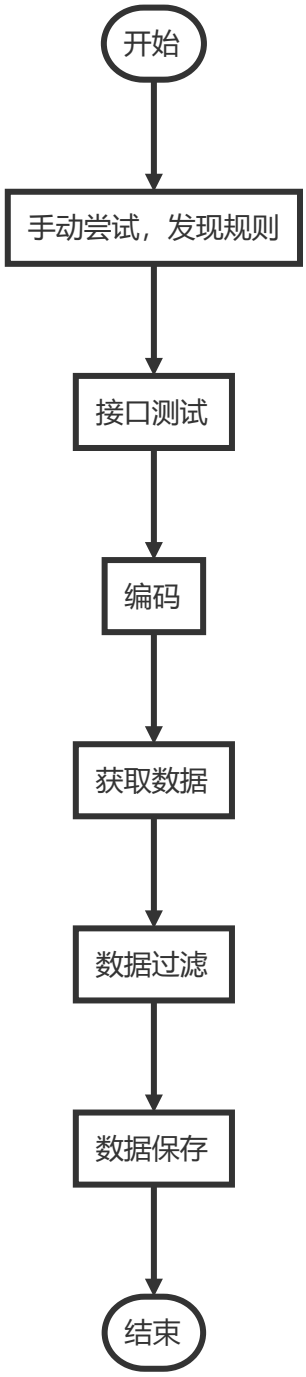
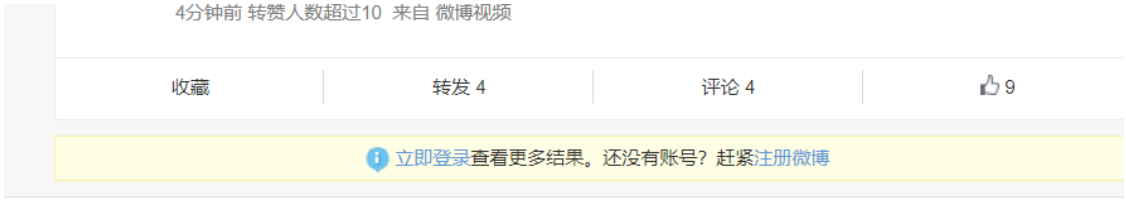
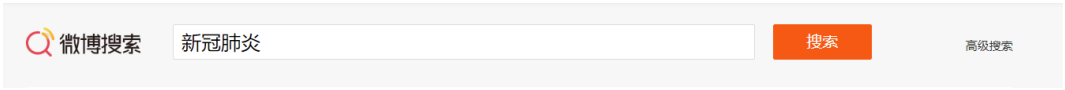
数据获取



手动尝试阶段

- url格式：`https://s.weibo.com/weibo?q=新冠肺炎&Refer=index`
- 未登录：只能获取少量数据

- 

- 登录后
  - 每次检索，能够获取最多50页信息，每页20条
  - 搜索界面有“高级搜索”入口
  - 

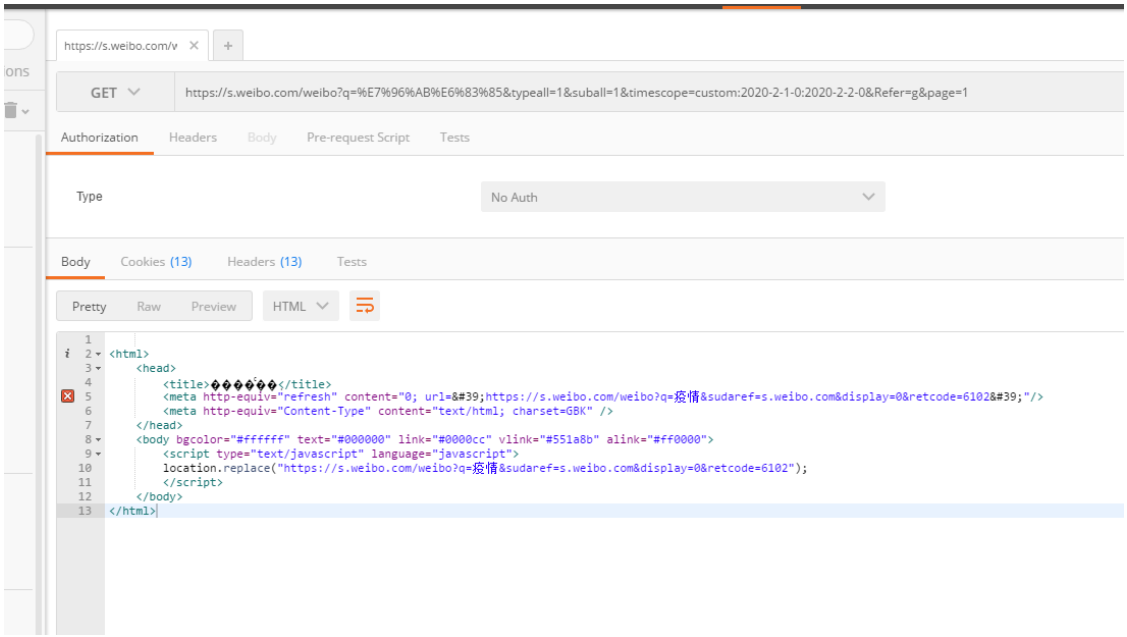
  - 

  - 高级搜索url格式

    `https://s.weibo.com/weibo?q=新冠肺炎&typeall=1&suball=1&timescope=custom:2020-04-02-1:2020-04-12-11&Refer=g`
  - 翻页:
  - 第二页：`https://s.weibo.com/weibo?q=新冠肺炎&typeall=1&suball=1&timescope=custom:2020-04-02-1:2020-04-12-11&Refer=g&page=2`
  - 第一页：`https://s.weibo.com/weibo?q=新冠肺炎&typeall=1&suball=1&timescope=custom:2020-04-02-1:2020-04-12-11&Refer=g&page=1`
    - 等同于：`https://s.weibo.com/weibo?q=新冠肺炎&typeall=1&suball=1&timescope=custom:2020-04-02-1:2020-04-12-11&Refer=g`

接口测试

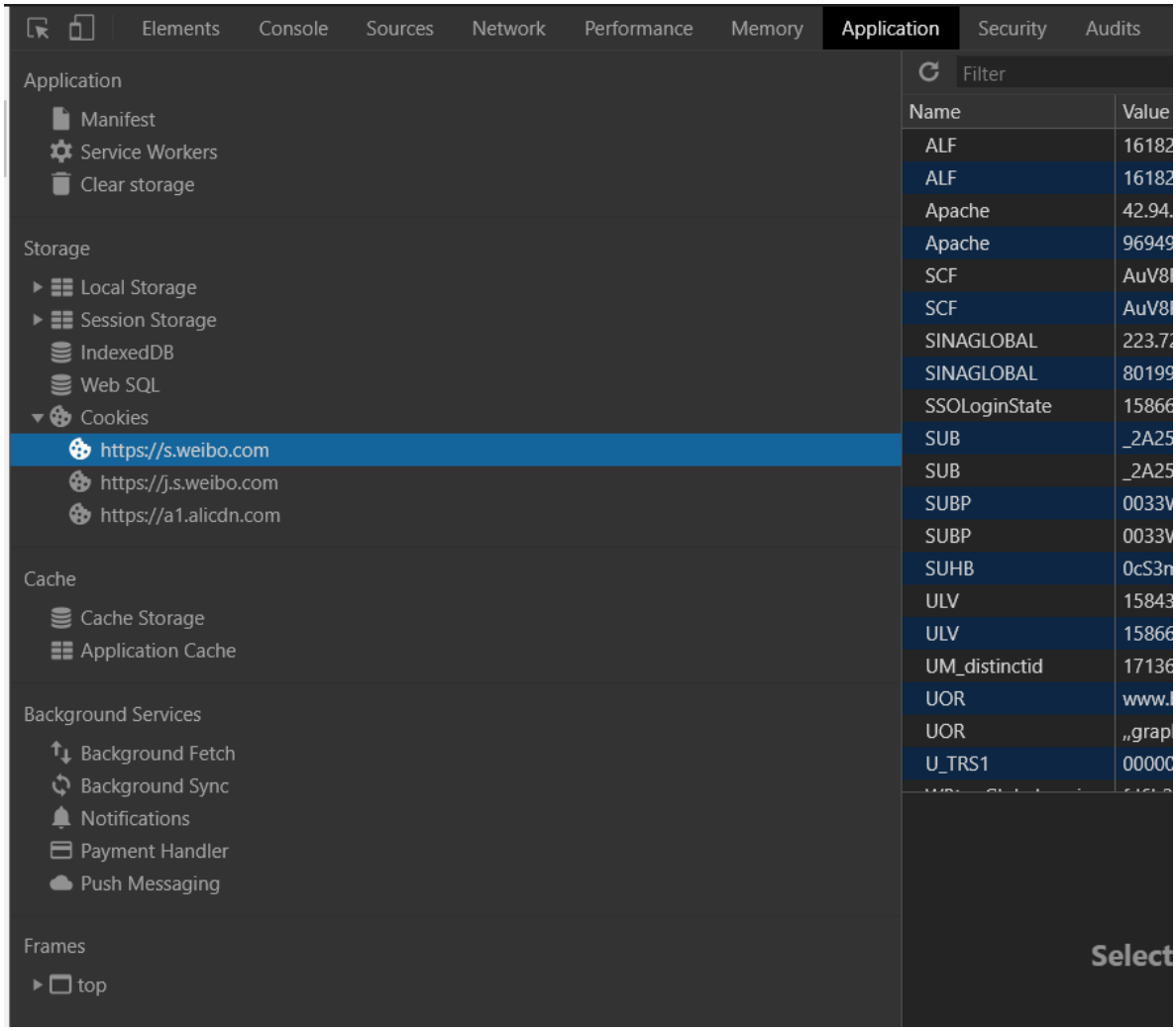
- 使用工具 postMan进行接口测试

- 

问题：直接使用url获取不到和浏览器相同的信息

解决方法: 添加cookie

cookie获取方式：浏览器先登录新浪微博，使用`F12`打开开发者模式，一次点击`Application`,`Cookies`



Name	Value	Domain	Path	Expires / Max-Age	Size	HttpOnly	Secure	SameSite
ALF	1618205919	.sina.com.cn	/	2021-04-12T05:38:39.347Z	13			
ALF	1618205919	.weibo.com	/	2021-04-12T05:38:39.653Z	13			
Apache	42.94.176.191_1586669542.950876	.sina.com.cn	/	Session	37			
Apache	9694981611926.66.1586669543390	.weibo.com	/	Session	36			
SCF	AuV8RXFsdag7wyTQhN9BY1c2jYwoSsh9aSMPFJ3qPuVSkGdtWg-HhvHXt0M_...	.sina.com.cn	/	2030-03-14T15:46:49.340Z	91	✓		
SCF	AuV8RXFsdag7wyTQhN9BY1c2jYwoSsh9aSMPFJ3qPuVSLV_bgo1CMav8_z_3a...	.weibo.com	/	2030-04-10T05:38:40.653Z	91	✓		
SINAGLOBAL	223.72.78.127_1578287353.957277	.sina.com.cn	/	2038-01-19T03:00:02.826Z	41			
SINAGLOBAL	8019930077197.102.1580695948502	.weibo.com	/	2030-01-31T02:12:28.000Z	41			
SSOLoginState	1586669919	.weibo.com	/	Session	23			
SUB	_2A25ztkODeRhGeBN71EX-CbNzDWIHXVQ4k3GrDV_PUNbm9AKLVqgkW9N...	.sina.com.cn	/	Session	93	✓		
SUB	_2A25ztkwDeRhGeBN71EX-CbNzDWIHXVQ4k34rDV8PUNbmtANLXCmkW9...	.weibo.com	/	Session	93	✓		
SUBP	0033WrSXqPxfM725Ws9jqgMF55529P9D9WhpGqrQUxTF9q65i-NzbDFS5JpX...	.sina.com.cn	/	Session	128			
SUBP	0033WrSXqPxfM725Ws9jqgMF55529P9D9WhpGqrQUxTF9q65i-NzbDFS5JpX...	.weibo.com	/	2021-04-12T05:38:40.653Z	128			
SUHB	0cS3mxNR6teeGy	.weibo.com	/	2021-04-12T05:38:40.653Z	18			
ULV	1584373613461:4:3:3::1584355172250	.sina.com.cn	/	2021-03-11T15:46:53.000Z	37			
ULV	1586669543409:9:2:1:9694981611926.66.1586669543390:1586407350564	.weibo.com	/	2021-04-07T05:32:23.000Z	67			
UM_distinctid	1713627536b44d-020b6f4b912792-c383f64-144000-1713627536cc5b	.sina.com.cn	/	2020-09-30T14:31:19.000Z	72			
UOR	www.baidu.com,blog.sina.com.cn,	.sina.com.cn	/	2021-01-05T05:09:15.000Z	34			
UOR	,,graph.qq.com	.weibo.com	/	2021-04-12T05:38:44.000Z	17			
U_TRS1	0000007f.38ae5849.5e12c0f8.78e4510d	.sina.com.cn	/	2030-01-03T05:09:14.403Z	41			

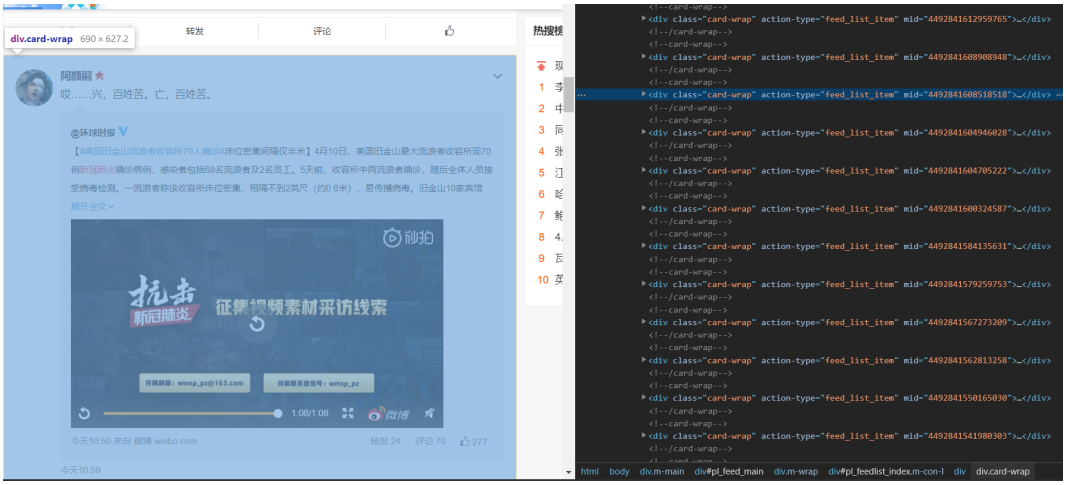
注意事项;

cookie会过期

编码阶段

1. 编程语言java
2. 第三方工具包:
    1. jsoup : 发送请求+html文档解析
    2. fastjson: json数据操作

3. 解析html文档，找到所需数据
    1. 使用chrome浏览器
    2. 几种可能的情况:
        1. 原创
        2. 转发，空白文本
        3. 转发，有文本
    3. 使用转发+有文本的微博格式来定位数据



4. 数据列表

```
<!--card-wrap-->
<div class="card-wrap" action-type="feed_list_item" mid="4492841621378426">…</div>
<!--/card-wrap-->
<!--card-wrap-->
<div class="card-wrap" action-type="feed_list_item" mid="4492841612959765">…</div>
<!--/card-wrap-->
<!--card-wrap-->
<div class="card-wrap" action-type="feed_list_item" mid="4492841608908948">…</div>
<!--/card-wrap-->
```

5. 详细信息

用户名&id

```
<a href="//weibo.com/5915215161?refer_flag=1001030103_" class="name" target="_blank" nick-name="重庆检察" suda-data="key=tblog_search_weibo&amp;value=seqid:1586670305619057212113|type:1|t:0|pos:2-0|q:%E6%96%B0%E5%86%A0%E8%82%BA%E7%82%8E|ext:cate:31,mpos:1,click:user_name">重庆检察</a>
```

微博内容：

```
    <p class="txt" node-type="feed_list_content" nick-name="重庆检察">

        <a href="https://s.weibo.com/weibo?q=%23%E5%85%A8%E5%9B%BD%E7%A1%AE%E8%AF%8A%E6%96%B0%E5%9E%8B%E8%82%BA%E7%82%8E%E7%97%85%E4%BE%8B%23" target="_blank">#全国确诊新型
            <em class="s-color-red">肺炎</em>病例#
        </a> 【
        <a href="https://s.weibo.com/weibo?q=%2331%E7%9C%81%E5%8C%BA%E5%B8%82%E6%96%B0%E5%A2%9E%E6%96%B0%E5%86%A0%E8%82%BA%E7%82%8E%E7%A1%AE%E8%AF%8A%E7%97%85%E4%BE%8B99%E4%BE%8B%23" target="_blank">#31省区市新增
            <em class="s-color-red">新冠</em>
            <em class="s-color-red">肺炎</em>
            确诊病例99例#
        </a>
        97例为境外输入】4月11日0–24时，31个省区市和新疆生产建设兵团报告新增确诊病例99例，其中97例为境外输入病例，2例为本土病例（黑龙江2例）；无新增死亡病例；新增疑似病例49例，均为境外输入病例（上海43例，黑龙江3例，内蒙古2例，吉 •
        <a href="//weibo.com/5915215161/ICXeyD57s?refer_flag=1001030103_" action-type="fl_unfold" target="_blank">
```

```
            展开全文
            <i class="wbicon">c</i>
        </a>
    </p>
```

发表时间

```
<p class="from">
    <a href="//weibo.com/5915215161/ICXeyD57s?refer_flag=1001030103_"
target="_blank" suda-
data="key=tblog_search_weibo&amp;value=seqid:1586670305619057212113|type:1|t:
0|pos:2-
0|q:%E6%96%B0%E5%86%A0%E8%82%BA%E7%82%8E|ext:cate:31,mpos:1,click:wb_time">
 今天10:59
    </a>
     来自
    <a href="//weibo.com/" rel="nofollow">iPhone客户端</a>
</p>
```

转发，点赞，评论

```
<div class="card-act">
    <ul>
        <li>
            <a href="javascript:void(0);" action-type="feed_list_favorite"
suda-
data="key=tblog_search_weibo&amp;value=seqid:1586670305619057212113|type:1|
t:0|pos:2-
0|q:%E6%96%B0%E5%86%A0%E8%82%BA%E7%82%8E|ext:cate:31,mpos:1,click:fav">
                收藏
            </a>
        </li>
        <li>
            <a href="javascript:void(0);" action-
data="allowForward=1&amp;mid=4492841629314474&amp;name=重庆检察
&amp;uid=5915215161&amp;suda-
data=key%3Dtblog_search_weibo%26value%3Dseqid%3A158667030561905721211 3%7Cty
pe%3A1%7Ct%3A0%7Cpos%3A2-
0%7Cq%3A%25E6%2596%25B0%25E5%2586%25A0%25E8%2582%25BA%25E7%2582%258E%7Cext%
3Acate%3A31%2Cclick:do_repost,mid:4492841629314474" action-
type="feed_list_forward" suda-
data="key=tblog_search_weibo&amp;value=seqid:1586670305619057212113|type:1|
t:0|pos:2-
0|q:%E6%96%B0%E5%86%A0%E8%82%BA%E7%82%8E|ext:cate:31,mpos:1,click:repost,mi
d:4492841629314474">
                转发
            </a>
        </li>
        <li>
            <a href="javascript:void(0);" action-
data="pageid=weibo&amp;suda-
data=key%3Dtblog_search_weibo%26value%3Dweibo_h_1_p_p" suda-
data="key=tblog_search_weibo&amp;value=seqid:1586670305619057212113|type:1|
t:0|pos:2-
0|q:%E6%96%B0%E5%86%A0%E8%82%BA%E7%82%8E|ext:cate:31,mpos:1,click:comment"
action-type="feed_list_comment">
```

```
                评论 1</a>
        </li>
        <li>
            <a title="赞" action-data="mid=4492841629314474" action-type="feed_list_like" href="javascript:void(0);" suda-data="key=tblog_search_weibo&amp;value=seqid:1586670305619057212113|type:1|t:0|pos:2-0|q:%E6%96%B0%E5%86%A0%E8%82%BA%E7%82%8E|ext:cate:31,mpos:1,click:like,mid:4492841629314474,act:add">
                <i class="icon-act icon-act-praise"></i>
                <em>1</em>
            </a>
        </li>
    </ul>
</div>
```

注意事项：

每次循环获取数据之后暂停一会，防止被服务器判定为不正常访问

```
private static int INTERVAL_PAGE=1000;
private static int INTERVAL_DAY=3000;
private static int INTERVAL_KEYWORD=10000;
```

数据清洗

1. 文本检查：
    1. 过滤掉文本长度小于5或者大于500的记录；
    2. 过滤掉文本中没有包含关键词的记录
2. 用户名检查：
    1. 过滤掉用户名包含下列字段的记录
        1. 新闻
        2. 日报
        3. 央视