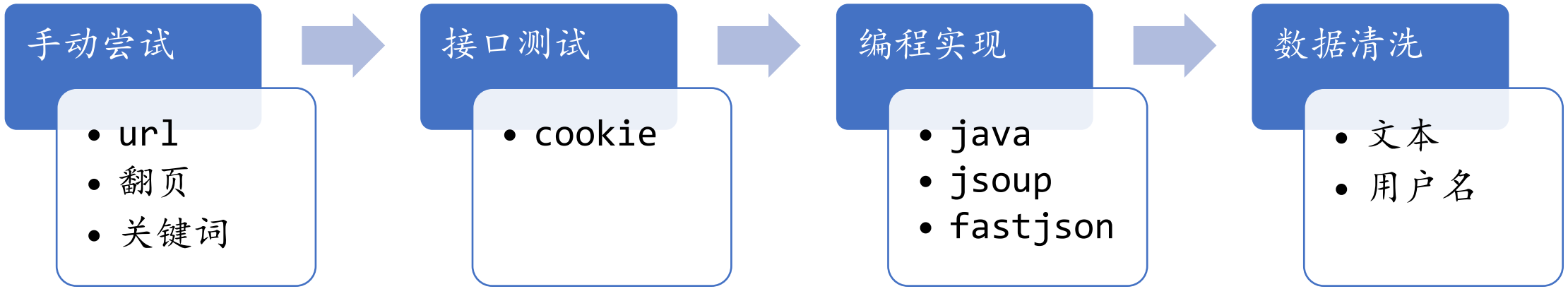
微博数据爬取

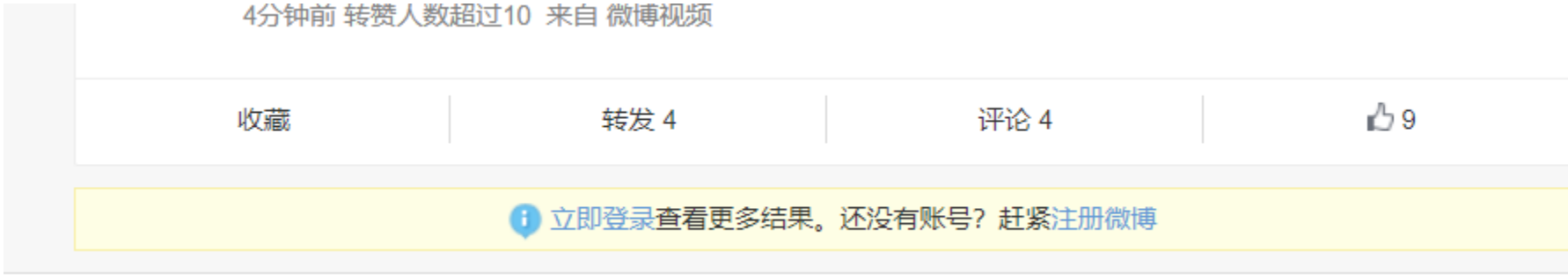
交通7 曹志坚

整体流程


手动尝试
url
翻页
关键词
接口测试
cookie
编程实现
java
jsoup
fastjson
数据清洗
文本
用户名

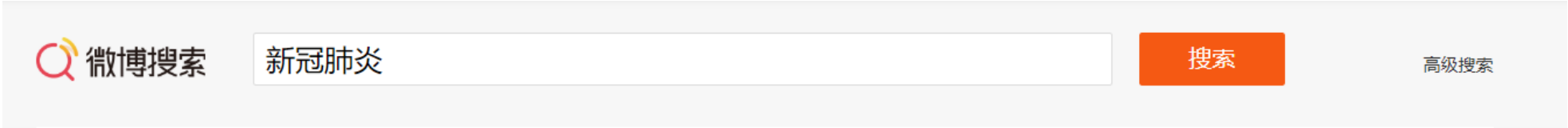
1.手动尝试

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



1.手动尝试

- 登录后
  - 有“高级搜索”入口
  - 每次搜索最多可以呈现50页记录，每页20条左右



1.手动尝试

- 高级搜索url
  - 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：

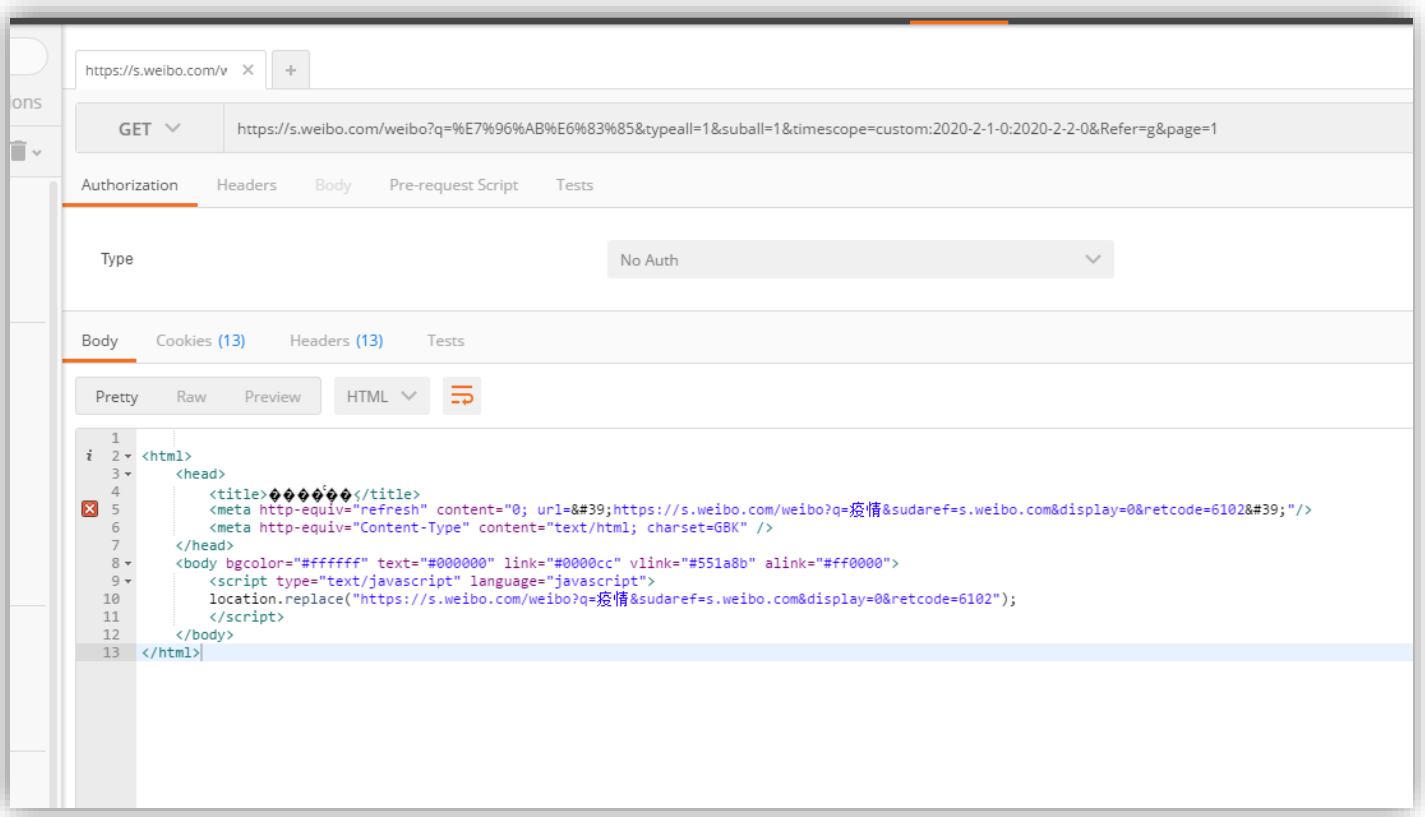
https://s.weibo.com/weibo?q=新冠肺炎&typeall=1&suball=1&timescope=custom:2020-04-02-1:2020-04-12-11&Refer=g

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

2.接口测试

- 工具 postman



2.接口测试

- 直接使用url获取不到数据
- 解决方案：html请求中添加**cookie**

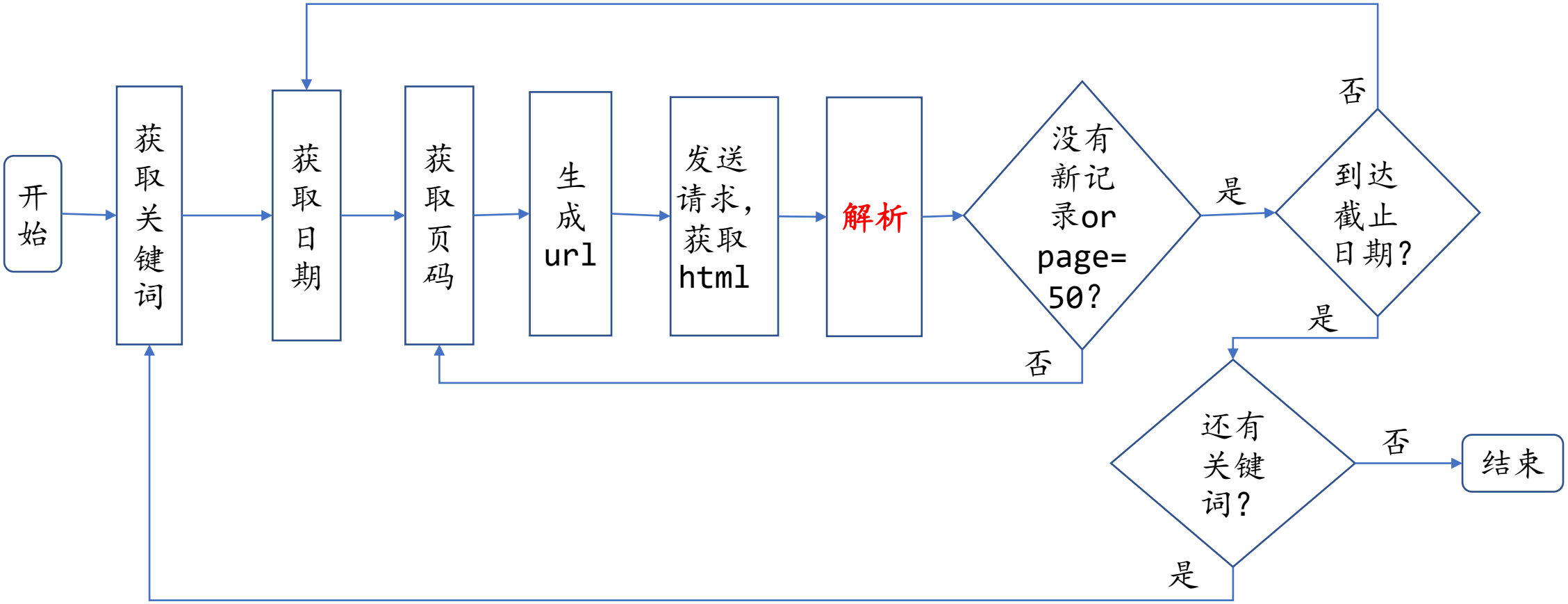
Name	Value	Domain	Path	Expires / Max-Age	Size	HttpOnly	Secure	SameSite
ALF	1618205919	.sina.com.cn	/	2021-04-12T05:38:39.347Z	13			
ALF	1618205919	.weibo.com	/	2021-04-12T05:38:39.653Z	13			
Apache	42.94.176.191_1586669542.950876	.sina.com.cn	/	Session	37			
Apache	9694981611926.66.1586669543390	.weibo.com	/	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	_2A25zltkODeRhGeBN71EX-CbNzDWIHXVQ4k3GrDV_PUNbm9AKLVqgkW9N...	.sina.com.cn	/	Session	93	✓		
SUB	_2A25zltkwDeRhGeBN71EX-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			

3.编程实现

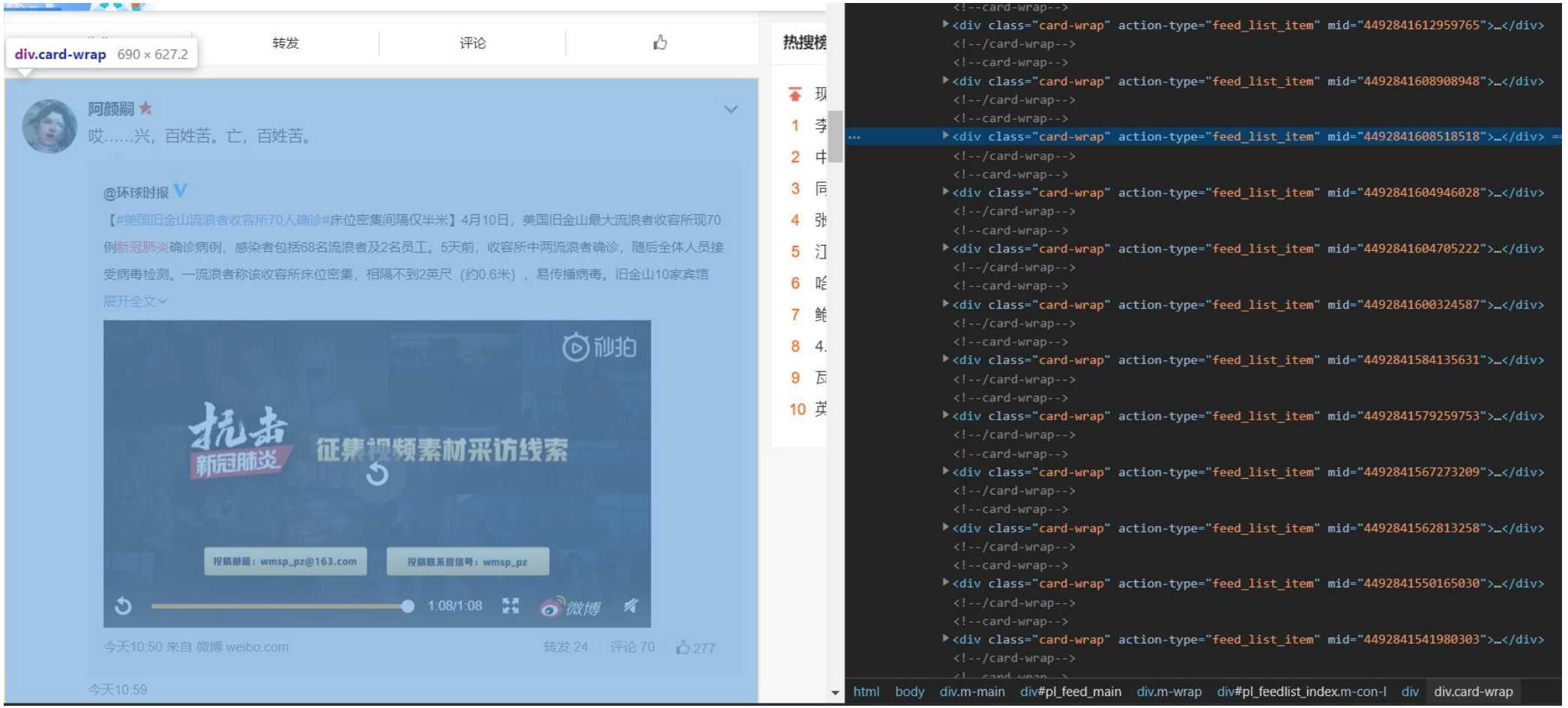
- 编程语言：java
- 第三方工具包
  - jsoup:发送请求+html文档解析
  - fastjson:json数据操作

fastjson

3.编程实现——程序流程图


开始
获取关键词
获取日期
获取页码
生成url
发送请求，获取html
解析
没有新记录or page=50?
是
否
到达截止日期?
否
是
还有关键词?
否
结束
是

3.编程实现——html解析

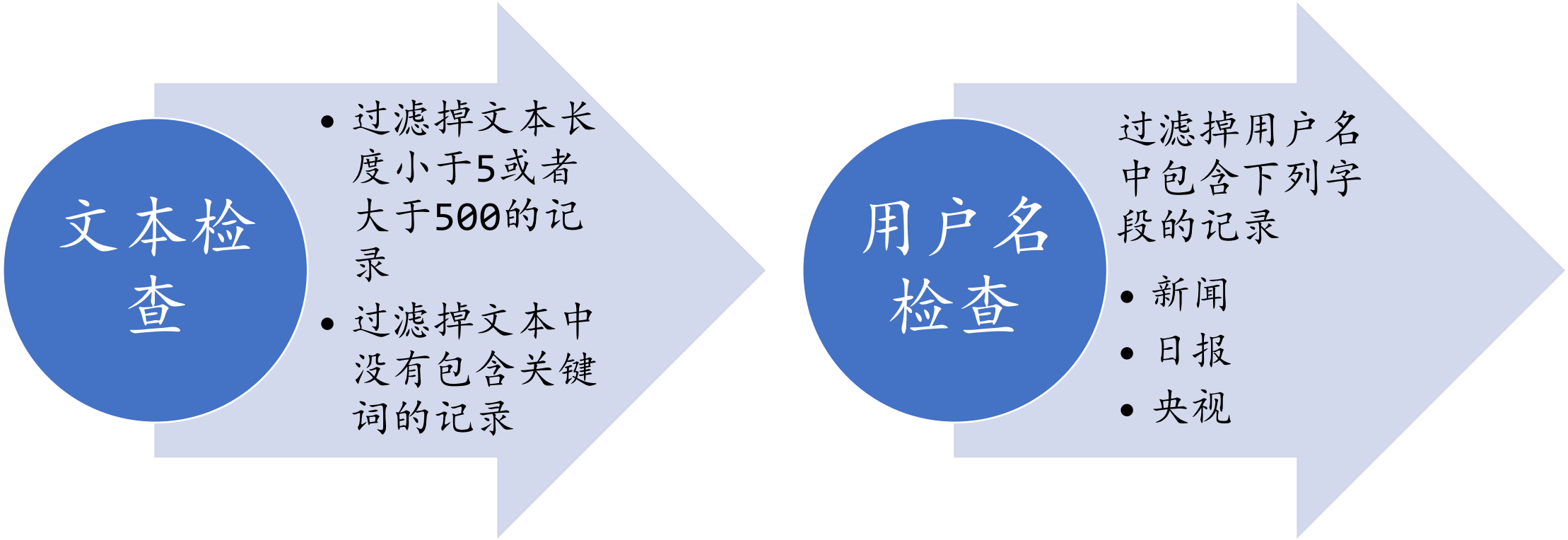


3.编程实现

- 注意事项
  - 每次向服务器发送请求之后让线程休眠一段时间，防止操作过于频繁被服务器判定为异常操作

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

4.数据清洗


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

谢谢，欢迎批评指正！