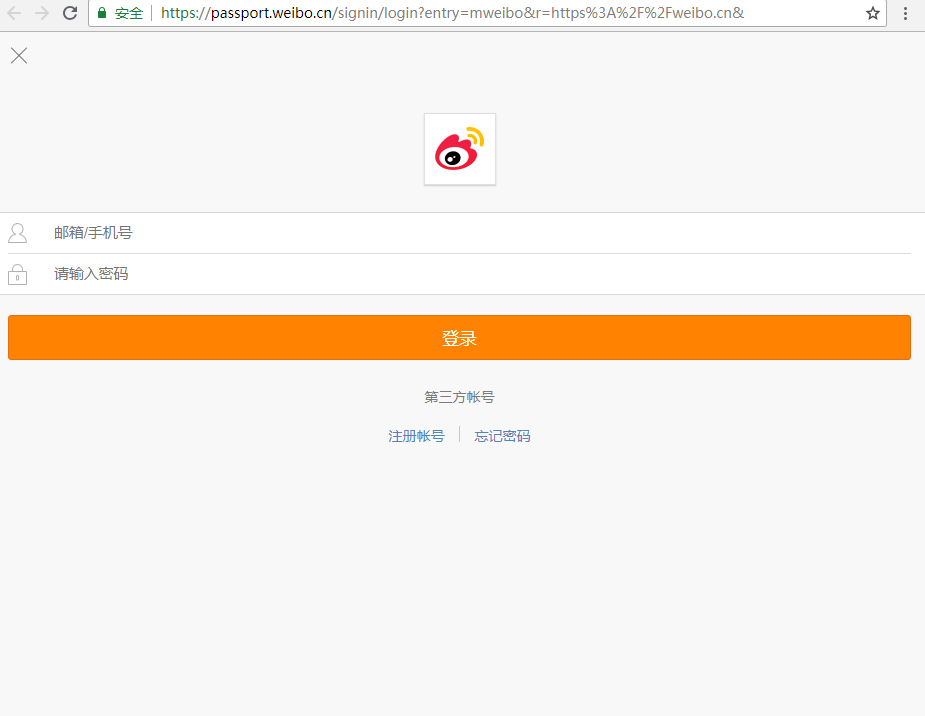
微博爬虫分析

1. 爬虫目标：
2. 爬取用户信息，关系
3. 爬取用户微博
4. 微博URL信息：

爬取微博信息需要登录微博账号，且设置合理的爬取策略，与采取一定的反反爬虫技术。例如：cookie使用，每次爬取间隔时间，并发数，反反爬虫过程中的user-agent、ip-proxy等。

**Login：**

url:[**https://weibo.cn/login/**](https://weibo.cn/login/)



获取cookies，业务爬取时使用，避免登录跳转重定向。需要检查cookie失效与重登机制。

实现方式采用selenium+headless driver（PhantomJS/Chrome/Firfox…）模拟登录操作。

**Info：**

url：[**https://weibo.cn/%s(ID)/info**](https://weibo.cn/%25s(ID)/info)

可获取信息：

ID

NickName

Area：可拆分province city

Birthday

Sex Orientation

Brief Introduction

Authentication 等，具体根据需求定制，可以关联其他页面获取额外信息。



**Tweets：**

**url：https://weibo.cn/%s(ID)/profile?filter=1&page=%num**

参数说明：%s为ID号

Filter=1表示原创、0表示全部、2表示图片

%num页码

可获取信息：

ID

Tweets ID

Content

Public Info:可才分Time、Terminal

Likes、Comments、Forwards

其他信息：tweets、follows、fans等





**Follow：**

**url：https://weibo.cn/%s(ID)/follow**

**Fan:**

**Url:** **https://weibo.cn/%s(ID)/fans**

可获取ID号的Follws和Fans信息：

Follow ID、Fan ID

****

1. 业务分析与问题：

理论上根据提供的一组种子ID，通过不断爬取follows、fansID可以爬取全网微博数据。

If not in IDs

Crawl Tweets

Crawl Infos

Follows ID

Fans ID

IDs

Relations

Pipeline

另外可以，自定义指定爬取某个ID用户的相关信息，具体策略（只爬取一层，或对其follow、fans做二次爬取分析）

难点：

爬取全网数据庞大，中间肯定会涉及到账号、ip被封等情况，如何容错、断线重爬，以及每时每刻ID信息都会有更新如何去重（判断是否进行二次爬取等），不论是基于url的判重还是基于爬取内容的判重都会涉及亿以上数据，又或者加上时间戳，增量入库，后续其他分析自行处理(闭环、死循环如何解决)。

其他问题肯定也很多，待分析。