**一、爬取内容选择：**

预爬取网站：

新浪微博、豆瓣、贴吧、知乎、CSDN、科学网、领英、小木虫、人人、QQ空间

【1】唯一标识码【2】昵称【3】性别年龄/生日【4】地点【5】教育/工作【6】头像【7】简介【8】关注列表【9】被关注列表【10】发布内容【11】标签【12】其他

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 唯一标识 | 昵称 | 性别 | 年龄生日 | 教育工作 | 头像 | 简介 | 关注列表 | 粉丝列表 | 发布内容 | 标签 |
| 新浪微博 | ID号 |  |  |  |  |  |  |  |  |  |  |
| 豆瓣 | 字符串 |  |  |  |  |  |  |  |  |  |  |
| 知乎 | 字符串 |  |  |  |  |  |  |  |  |  |  |
| 贴吧 | 访问需要登陆账号，只能查看非隐藏动态，而大部分人都选择隐藏。 | | | | | | | | | | |
| 领英 | 访问需要权限，必须加好友之后才能访问网页内容，不能大规模爬取。 | | | | | | | | | | |
| 人人 |
| QQ空间 |
| 科学网 | 访问需要登陆账号，发布内容多数都是长篇博客，标签化的信息不多。 | | | | | | | | | | |
| 小木虫 |
| CSDN |

1. 人人网，QQ空间，领英

访问需要权限，必须加好友之后才能访问网页内容，不能大规模爬取。

1. 百度用户（贴吧）

访问需要登陆账号，只能查看非隐藏动态，而大部分人都选择隐藏，可爬取内容不多。

1. 博客类（如CSDN）

访问需要登陆账号，【10】发布内容多数都是长篇博客，标签化的信息不多。

1. **豆瓣**

【1】ID号或者字符串

【2】【6】

【8】【9】能从这两个列表获得所有好友的ID号

【10】广播，一个豆瓣活动的日志，也有一些类似微博的短消息

【11】豆瓣小组，类似标签

【12】有大量的“喜欢”内容，如书、电影、音乐

1. **微博**

【1】ID号

【2】【3】【4】【5】【6】【7】

【8】【9】能从这两个列表获得所有好友的ID号

【10】微博，可判别哪些是热门微博

【11】从微博给出的选择

1. **知乎**

【1】知乎指定唯一ID字符串

【2】【3】【4】【5】【6】【7】

【8】【9】能从这两个列表获得所有好友的ID号

【10】提问、回答，可以提取回答的内容

【11】关注了话题作为标签，以及设置自己擅长

【12】有部分人已经和微博进行关联，可作为测试集

**二、可以爬取的内容：**

**微博**

主页url：[http://weibo.cn/\*\*\*/](http://weibo.cn/***/profile)



图1:微博界面



图2:源代码中可以得到ID

资料url: [http://weibo.cn/\*\*\*/info](http://weibo.cn/***/info)



图3:个人信息页面

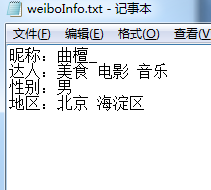


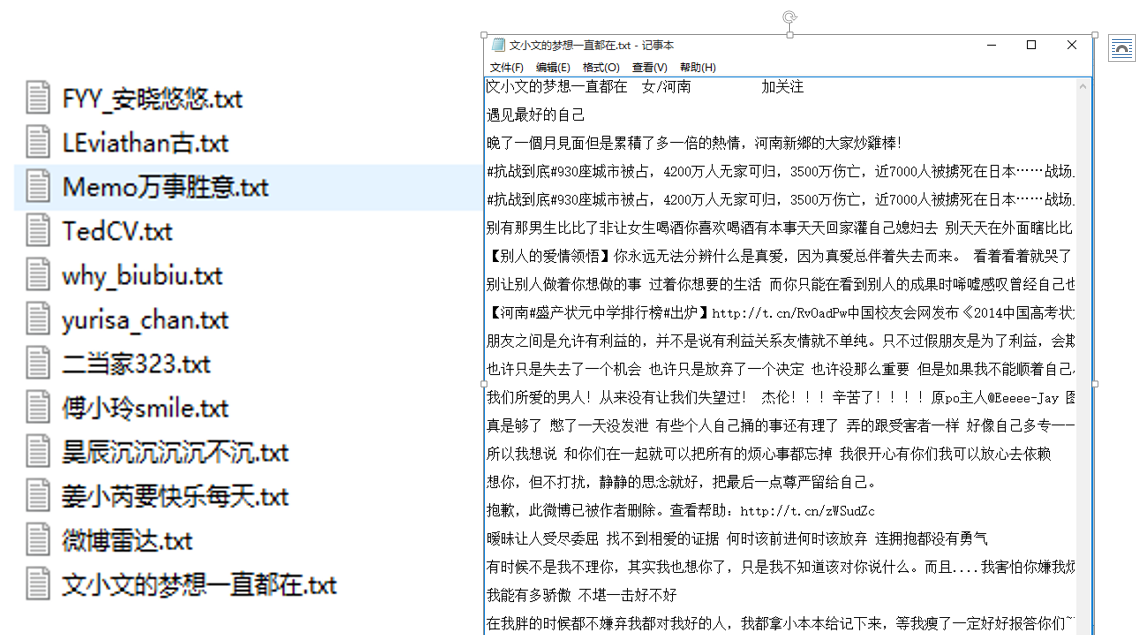
图4:爬取下的信息样例

粉丝url：http://weibo.cn/\*\*\*/fans

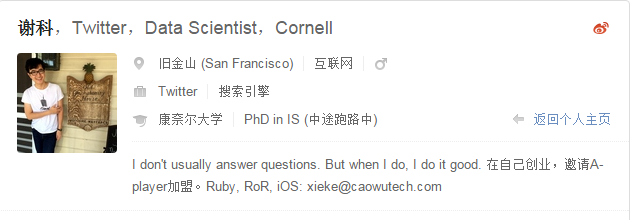


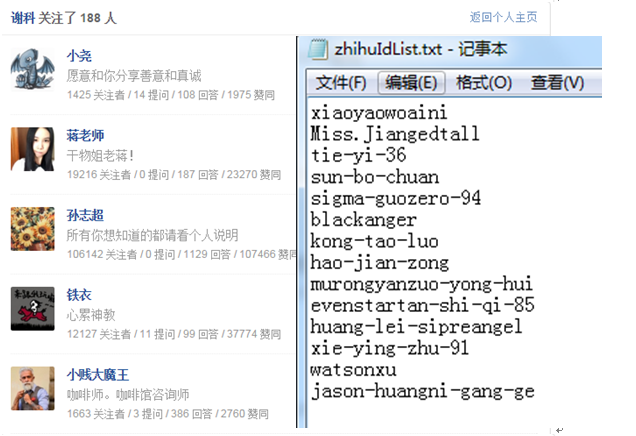
图5:粉丝页面及样例

微博url：[http://weibo.cn/\*\*\*/profile](http://weibo.cn/***/profile)

****图5:爬取到微博样例

**知乎：**

****

****

****

****

**豆瓣：**

****

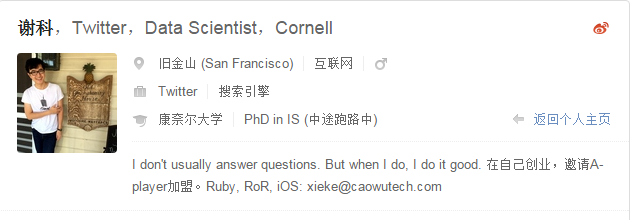
****

****

****

**三、可融合内容**

1.知乎的微博链接

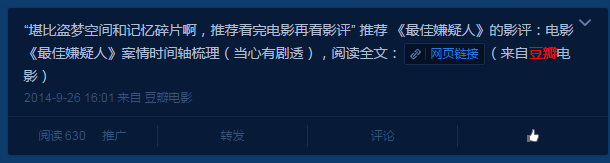
****

****

2.微博的知乎、豆瓣微博

这个只能说明对相关内容感兴趣，不代表微博主人和所发内容作者是一个人。

****

****

**四、爬取方法：**

**（一）主要技术**

1.使用cookies或提交表单两种方法模拟登陆

2.使用requests得到html文件

3.用re、lxm、BS4解析html文件

**（二）爬取策略（以微博为例）：**

1.进入一个ID的关注列表，得到其好友的ID链接

2.进入每个ID链接中，得到好友的ID号

3.放到字典里，作为键？写入文件中？为每个ID号建立文件？（去重）

4.按照深度遍历的方法，递归得到大量的ID号

5.按照ID号的顺序爬取每个ID所需要的内容

**五、可能遇到的问题：**

**爬取频率。**

**如何突破网站对于爬虫的限制。**

策略：

1模仿真人登陆：

设定sleep时间。

2更换登陆的认证信息：

准备多个账号，将其Cookie放到列表中。每个ID爬取完更换cookie。(**验证码**)

3其他…