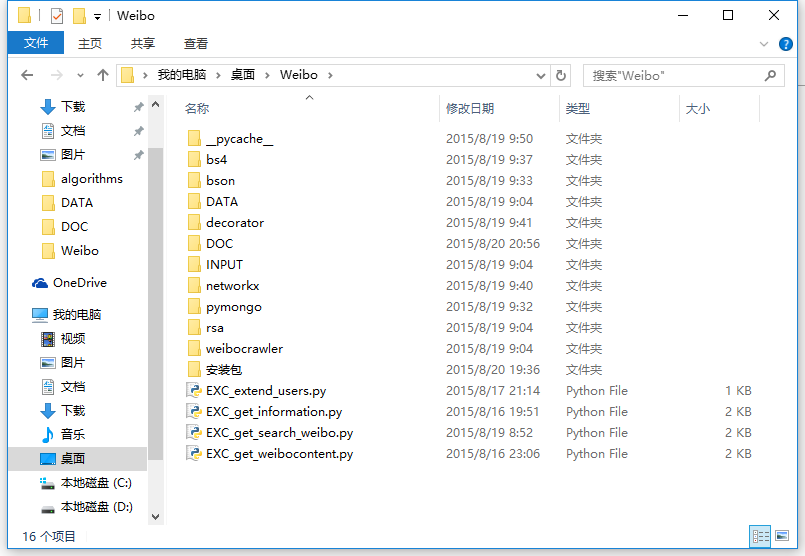
新浪微博爬虫分析工具使用指南

# 一、更新日志

|  |  |  |  |
| --- | --- | --- | --- |
| 日期 | 模块 | 内容 | 作者 |
| 2015/8/18 | 系统程序 | 系统开发、调试完成 | xcc |
| 2015/8/20 | 说明文档 | 说明文档编写完成 | xcc |
| 2015/8/21 | config.txt | 增加爬虫页面、睡眠时间可调 | xcc |
| DirOperator.py | 修改为：每次数据存储在数据库名字对应文件夹下 | xcc |
| 2016/12/5 | EXC\_nickname2userhref.py | 昵称到uid转化 | xcc |
|  |  |  |  |
|  |  |  |  |

# 二、文件结构介绍

Weibo文件夹包含新浪微博爬虫分析工具，本工具是由Python脚本语言开发，数据库采用MongoDB，所以需要安装并配置相关环境，这点会在“安装篇”详细介绍，具体使用及功能会在“使用篇”中详细介绍，下面先看一下文件夹内部的文件用途。



1、文件夹内四个py类型文件为实现四种功能的主程序：

EXC\_get\_information.py：获取用户基本信息

EXC\_get\_weibocontent.py：获取用户的微博信息

EXC\_extend\_users.py ：获取用户群体的相似扩展用户

EXC\_get\_search\_weibo.py：获取关键词搜索微博信息

2、DATA文件夹：

输出数据存放的文件夹，每一次数据存储在以当前数据库名字命名的文件夹中

user\_information.csv：用户信息表

weibo\_content.csv：用户微博内容表

recent\_weibo.csv：用户最近一条微博信息表

new\_users.csv：通过微博内容发现的新用户名单

extend\_users.csv：相似群体扩充表

search\_weibo.csv：关键词搜索结果列表

3、DOC文件夹：

存放说明文档，也就是你正在看的这个。

4、INPUT文件夹：

存放程序运行所需要的输入信息

config.txt：数据库表、集合配置文件

cookies.txt：存储微博cookies信息

SearchKeywords.txt：搜索微博的关键词

SeedUserList.txt：种子用户列表

5、安装包

需要安装的软件包

6、其余都是程序运行需要的库包，不出bug不用管。

# 三、安装篇

新浪微博爬虫分析工具依托Python环境，数据库采用MongoDB，所以需要安装软件，并配置相应环境，软件包在根目录下的“安装包”文件夹。

## Python安装配置

安装文件：python-3.4.3.msi

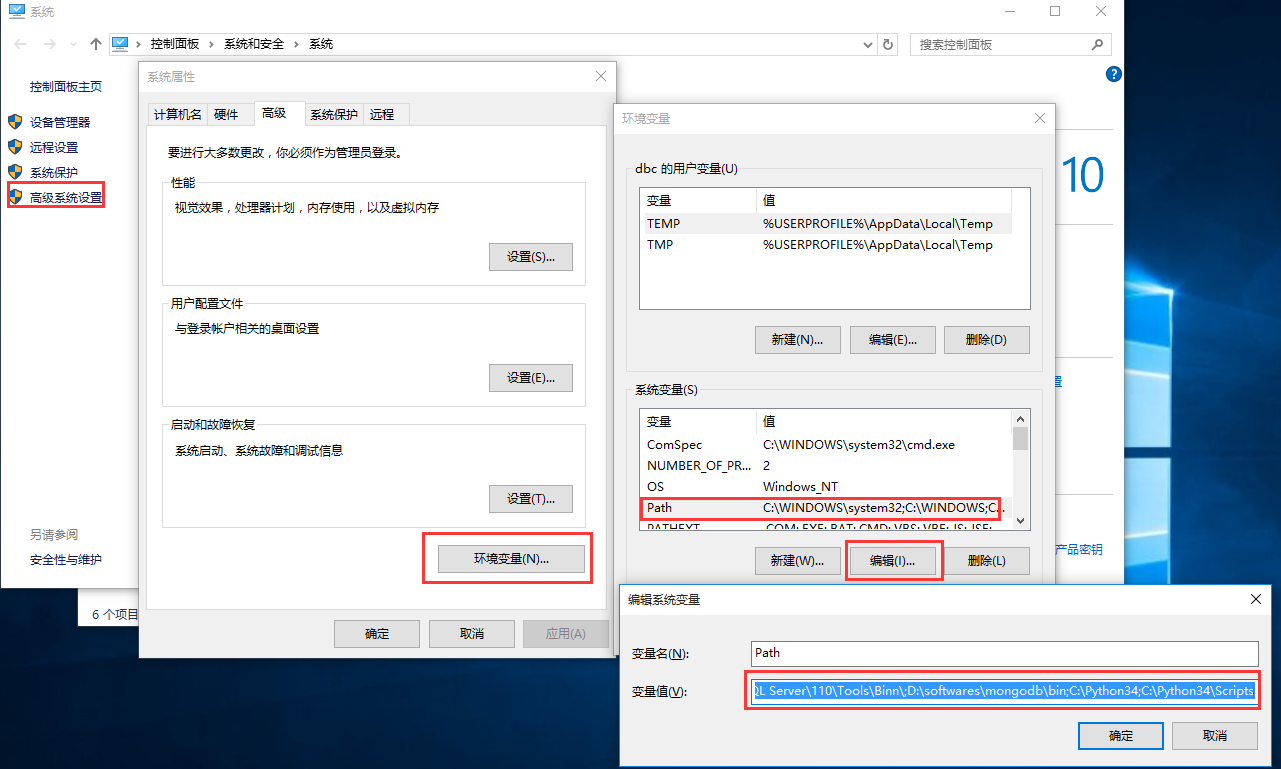
版本：3.4.3 32bit for Windows

配置环境变量（可选）：

Python默认安装目录：C:\Python34

easy\_install(自动安装第三方库包的工具)：C:\Python34\Scripts

具体步骤如下图所示：



安装完成后，会出现python IDLE，里面有python的编辑运行环境，后面运行程序可以在里面进行，或者在控制台窗口执行。

## MongoDB安装配置

1、MongoDB

安装文件：mongodb-win32-x86\_64-2008plus-2.6.7-signed.msi

配置环境变量：

找到安装目录，将其中bin文件夹内容的路径添加到环境变量

我的安装目录是：D:\softwares\mongodb\bin，所以将本目录添加到系统环境变量中。

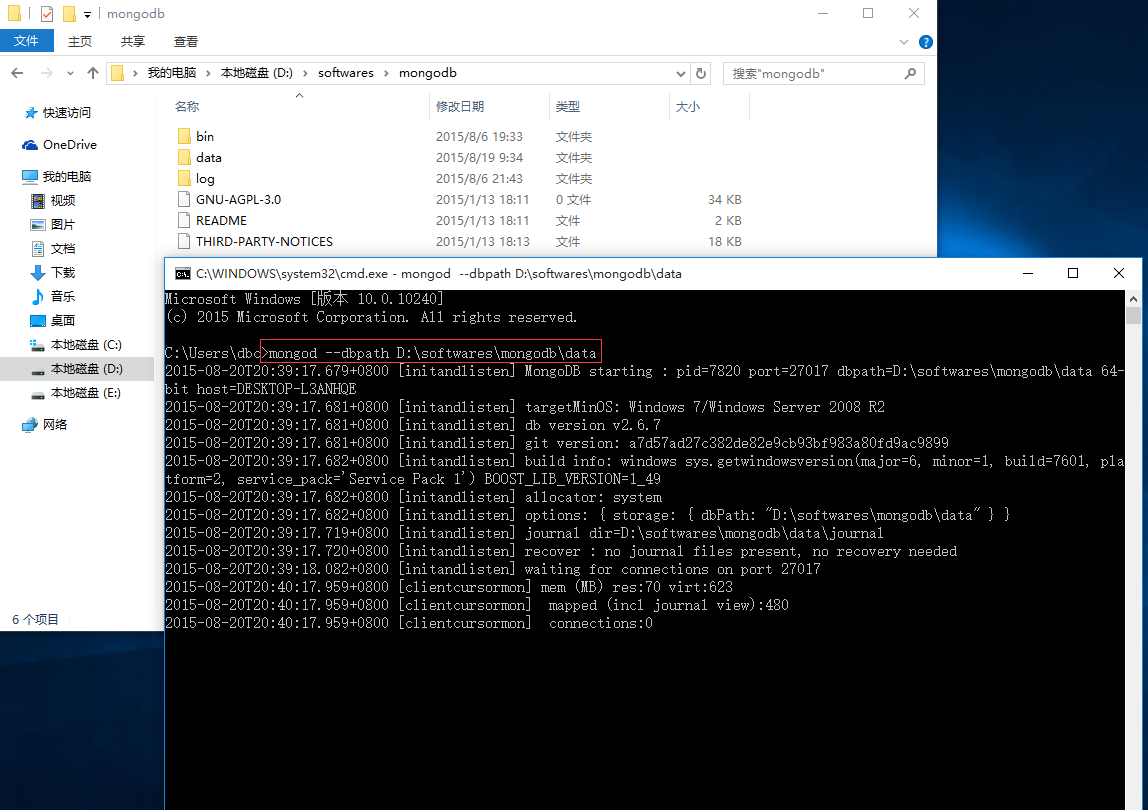
注意：

每次使用MongoDB数据库，需要打开MongoDB服务，也就是指定一个数据库文件夹，具体方法为：

在安装目录中新建一个data文件夹，在控制台窗口输入：

mongod --dbpath D:\softwares\mongodb\data

具体操作及效果如图所示：



出现上述效果则说明服务打开成功。

2.MongoDB可视化工具

因为MongoDB没有图形化的操作界面，在控制台中通过命令行查看、更改数据库内容并不方面，所以建议安装MongoDB可视化工具-MongoVUE

安装文件：MongoVUE.exe

破解工具：Installer.msi

安装成功，打开添加一个本地账户，就可以对本地的MongoDB数据库进行查看，更新工作了。

# 四、使用篇

## 1、使用前准备

主要是初始化INPUT文件夹中必要内容

1. **cookies.txt：**

新浪爬虫登录最简单直接的方式就是利用微博在浏览器中的cookies来解决，有了cookies，程序就能冒充浏览器来获取微博数据，下面是获取cookies的方法：

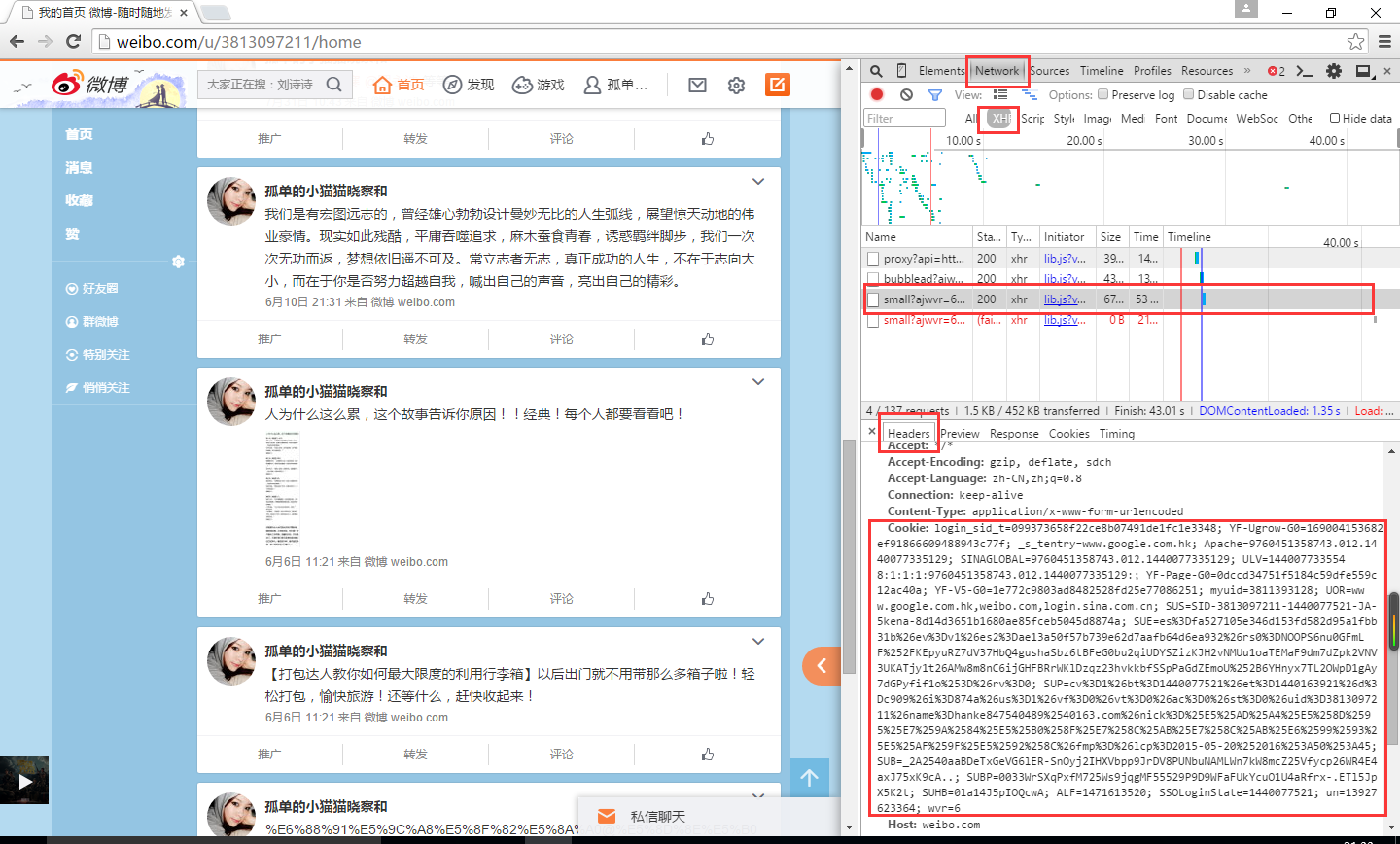
（1）浏览器手动登录微博（最好别用自己的账号，被封了就不好了）🡺

（2）右键“审查元素”，出现右边所示 🡺

（3）选择“Network”，筛选中选择：XHR选项，如果没有出现内容，刷新网页 🡺

（4）选择一个文件，在下面“header”选项中就有cookies信息了 🡺

（5）复制cookies信息，粘贴进cookies.txt文件



1. **config.txt：**

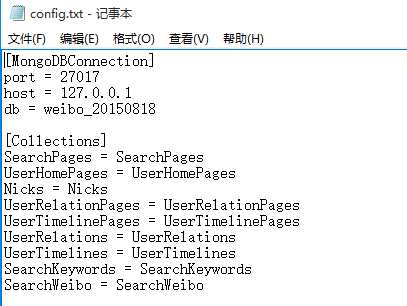
config.txt中储存了关于数据库的一系列信息，

（1）[MongoDBConnection]：

本地数据库port、host不用修改；db为指定的数据库名称，每次使用，需要自己手动修改

（2）[Collections]

数据库中的表信息，不用修改



1. **SeedUserList.txt**

种子用户列表

1. 用处：

如果需要如下三个功能，则需要初始化这个文件：

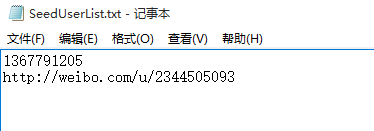
获取用户基本信息

获取用户微博信息

获取扩展相似用户

1. 格式：

支持微博UserID或者主页链接两种格式，如，



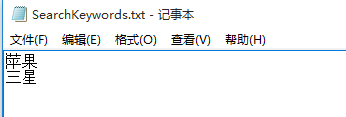
1. **SearchKeywords.txt**

搜索关键词列表

（1）用处：

如果希望进行关键词搜索，则需要在此文件输入搜索关键词

（2）格式：



## 2、获取用户基本信息

1、说在前面：

需要运行的文件是：EXC\_get\_information.py

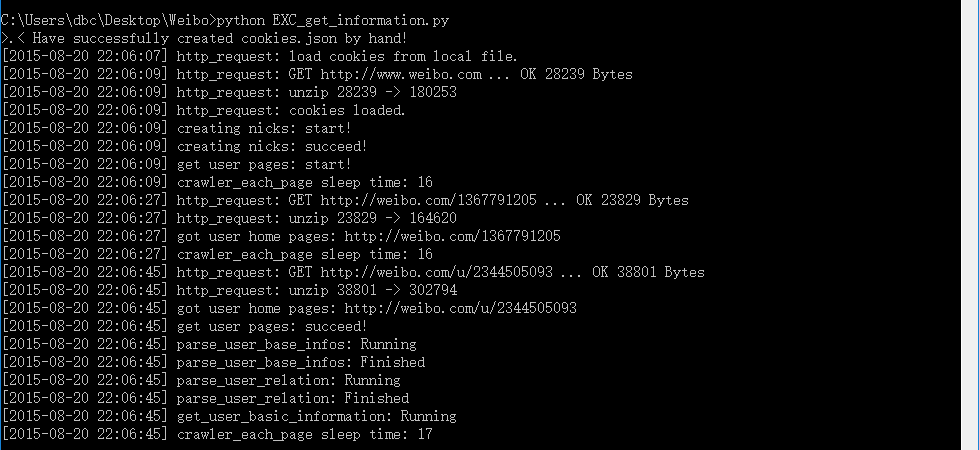
运行这一步的前提是已经将用户添加进了INPUT文件夹中的SeedUserList.txt

2、运行方法：

（1）python IDLE 打开运行

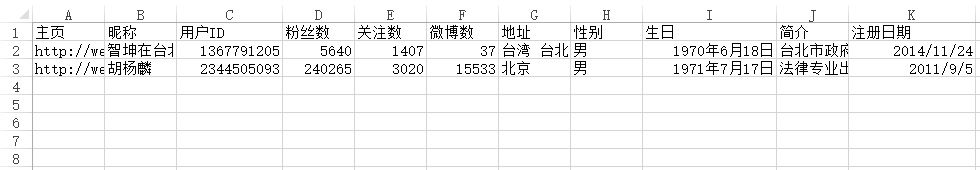
（2）命令行

控制台窗口，进入到Weibo文件夹内，输入python EXC\_get\_information.py，看到类似如下日志输出，则说明运行正常



3、运行结果

存放在DATA文件夹中user\_information.csv文件中，格式如下：



## 3、获取用户微博信息

1、说在前面：

需要运行的文件是：EXC\_get\_weibocontent.py

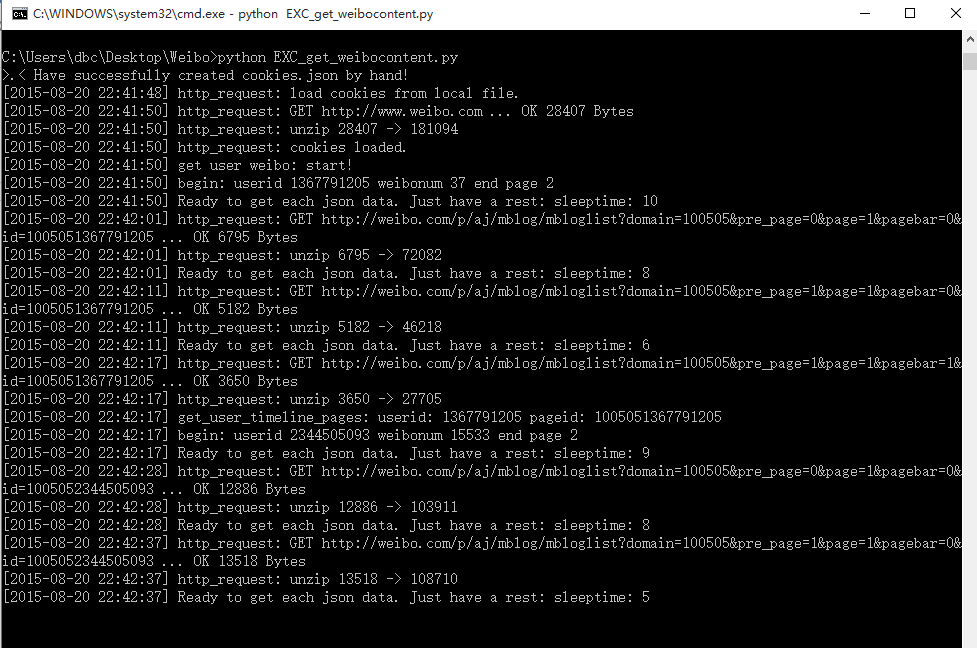
运行这一步的前提是已经成功运行过EXC\_get\_information.py程序，这是因为爬取微博内容需要用户的基本信息（pageID）作为基础。

2、运行方法：

（1）python IDLE 打开运行

（2）命令行

控制台窗口，进入到Weibo文件夹内，输入python EXC\_get\_weibocontent.py，看到类似如下日志输出，则说明正常爬取进行中



3、运行结果

存放在DATA文件夹中，共有三个文件中，分别如下：

1. weibo\_content.csv

种子用户微博信息，默认每个用户抓取微博数为45条（一页），理论上可以将用户全部微博抓取，修改位置：

config.txt >> [CrawlParameter] >> UserWeiboPages

`

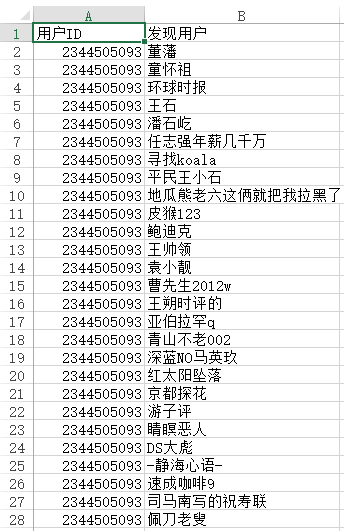
1. recent\_weibo.csv

最近一条微博信息



1. new\_users.csv

从微博内容中提取出来的新用户名字，与种子用户有转发互动



## 4、获取扩展相似用户

1、说在前面：

需要运行的文件是：EXC\_extend\_users.py

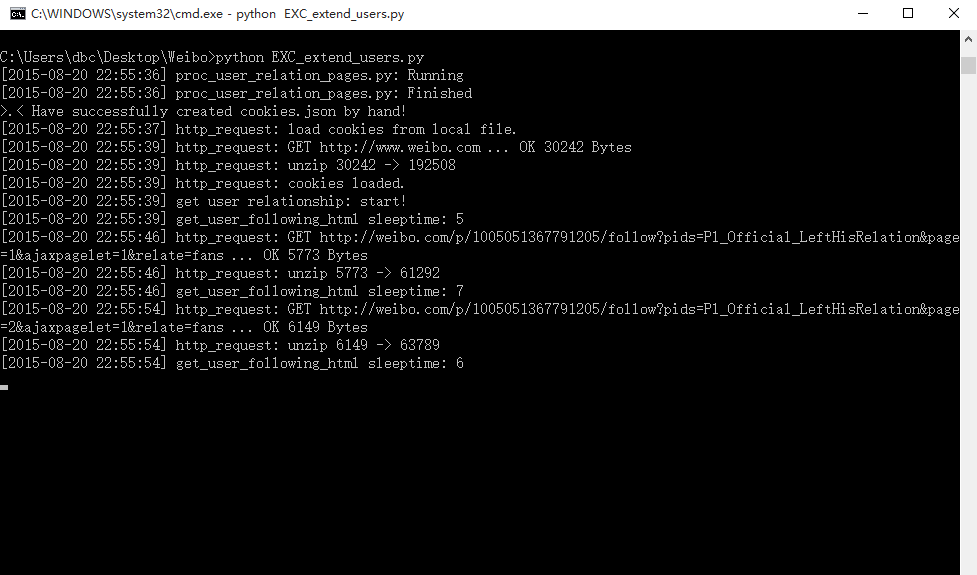
运行这一步的前提是已经成功运行过EXC\_get\_information.py程序，这是因为爬取微博内容需要用户的基本信息（pageID）作为基础。

2、运行方法：

（1）python IDLE 打开运行

（2）命令行

控制台窗口，进入到Weibo文件夹内，输入python EXC\_extend\_users.py，看到类似如下日志输出，则说明正常爬取进行中



3、运行结果

存放在DATA文件夹中extend\_users.csv文件中，包括扩充用户及权重信息

## 5、获取关键词搜索微博

1、说在前面：

需要运行的文件是：EXC\_get\_search\_weibo.py

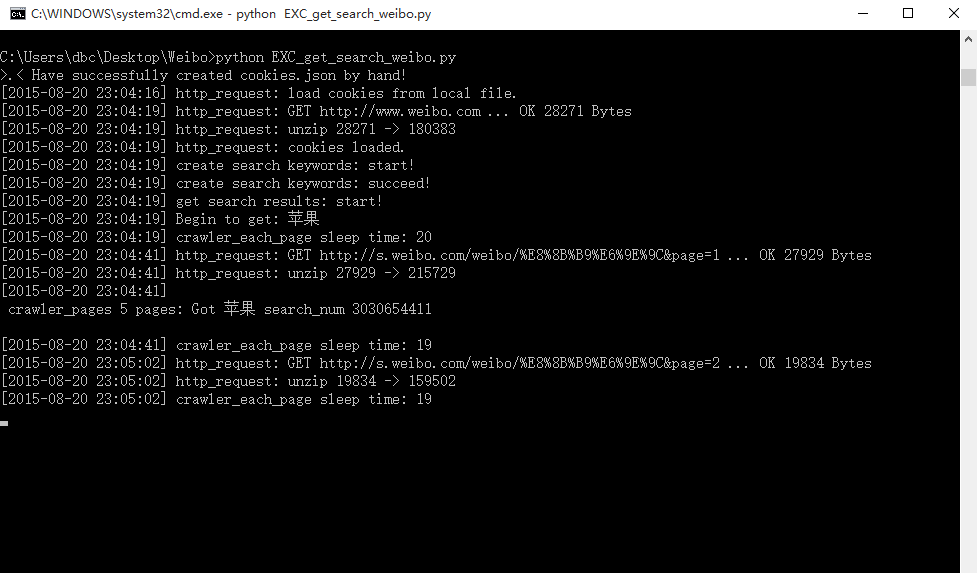
运行这一步的前提是已经将搜索关键词添加进了INPUT文件夹中的SearchKeywords.txt

2、运行方法：

（1）python IDLE 打开运行

（2）命令行

控制台窗口，进入到Weibo文件夹内，输入python EXC\_get\_search\_weibo.py，看到类似如下日志输出，则说明爬取正常进行。



3、运行结果

存放在DATA文件夹中search\_weibo.csv文件中，格式如下：



默认抓取每个人关键词五页的搜索结果，每页越为50条搜索结果，可修改：

config.txt >> [CrawlParameter] >> KeywordWeiboPages

自己指定要抓取的页码数

## 6、昵称转化为uid

1、说在前面：

需要运行的文件是：EXC\_nickname2userhref.py

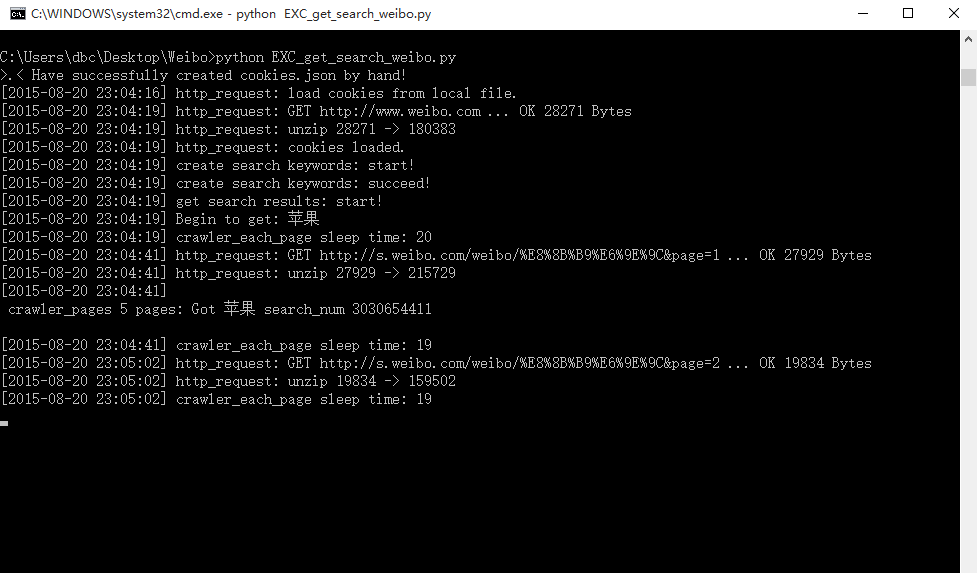
运行这一步的前提是已经将昵称添加进了INPUT文件夹中的SearchKeywords.txt

2、运行方法：

（1）python IDLE 打开运行

（2）命令行

控制台窗口，进入到Weibo文件夹内，输入python EXC\_nickname2userhref.py，看到类似如下日志输出，则说明爬取正常进行。



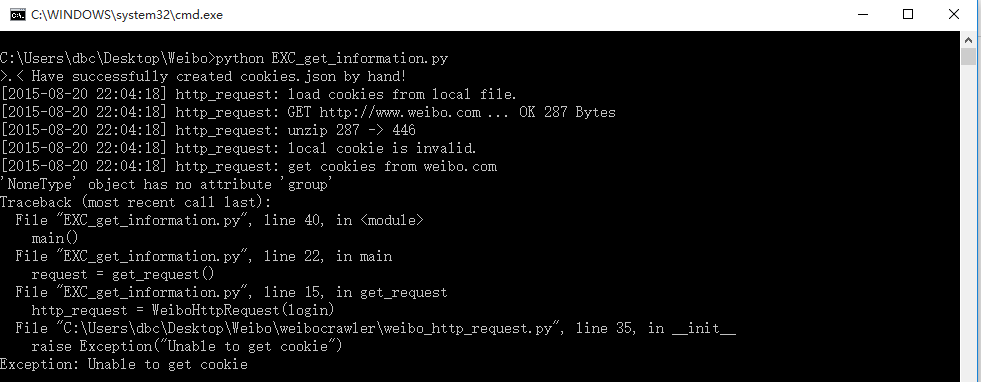
3、运行结果

存放在DATA文件夹中nick2href\_result.csv文件中。

# 五、常见故障

## 1、Exception: Unable to get cookie

解决方案：将最新的cookie更新到cookies.txt，具体方法参考第四章第一节中cookies.txt介绍部分

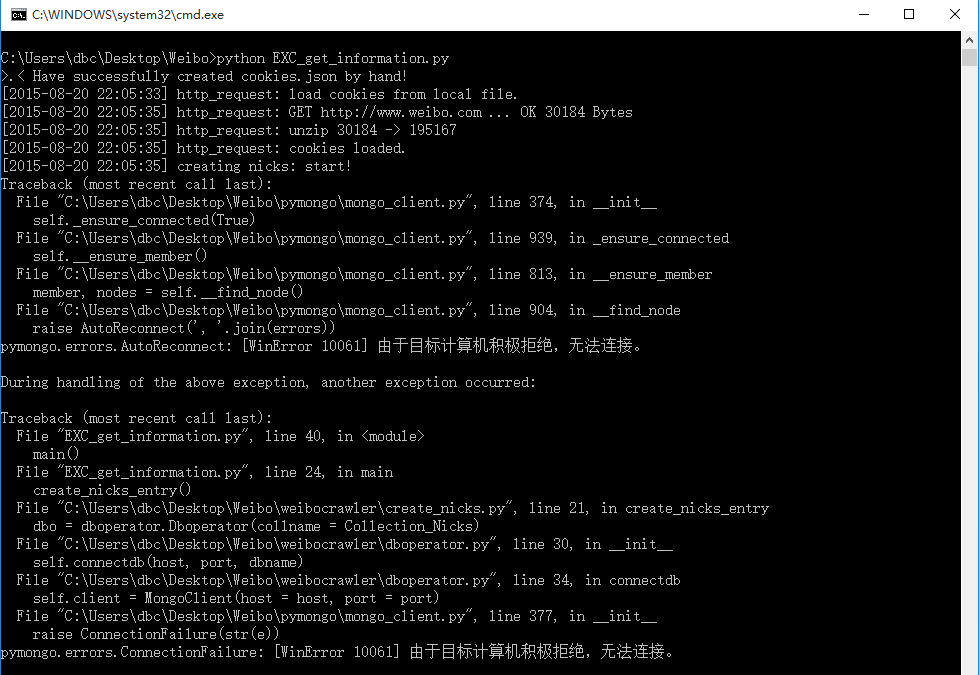


## 2、目标计算机积极拒绝，无法连接

出现此问题的原因是没有打卡MongoDB服务

解决方法：在控制台窗口输入：

mongod --dbpath D:\softwares\mongodb\data



## 3、频繁访问被封号

由于微博的反爬虫机制，同一IP，同一账号频繁访问会有被封号的可能，因此要修改访问微博的睡眠时间，睡眠时间越长，被封的风险就越小，但是效率就会变低，自己可以根据实际情况灵活调节，修改位置：config.txt 中CrawlParameter部分，详细含义如下：

KeywordSleeptime = 15

获取关键词搜索微博的睡眠时间设置，默认为15秒

UserPagesSleeptime = 15

获取用户信息的睡眠时间设置，默认为15秒

UserinfoSleeptime = 15

获取用户简介信息的睡眠时间设置，默认为15秒

UserWeiboSleeptime = 15

获取用户微博信息的睡眠时间设置，默认为15秒

UserRelationSleeptime = 15

获取用户关系信息的睡眠时间设置，默认为15秒

## 4、程序成功运行，数据大部分为空

校园网已经连接上，但是没有账号登录，这时候也可以获取数据，不过抓取的都是字节数比较少的空白页面，所以解析时候没有数，断网会直接有提醒，程序停止运行。

出现数据大部分为空的情况下，首先检查下网络吧。