本章将就实现的网络爬虫进行测试运行，并展示爬取的数据。

**5.1测试环境**

开发语言：python2.7

需要的Python模块：scrapy、requests、lxml、pywin32、pymongo、json、httplib、base64、redis

开发环境：8G内存、i5-4200H处理器、三星SSD 850 EVO 120G

操作系统：Windows10专业版系统、操作系统安装在三星SSD上

网络带宽：联通15M带宽

数据库：Redis、Mongodb

IDE：Pycharm

绘图工具：Visio2013

统计工具：Execl2013

浏览器：Google Chrome版本 64.0.3282.119（正式版本） （64 位）

数据库可视化工具：mongobooster、RedisStudio-en-0.1.5

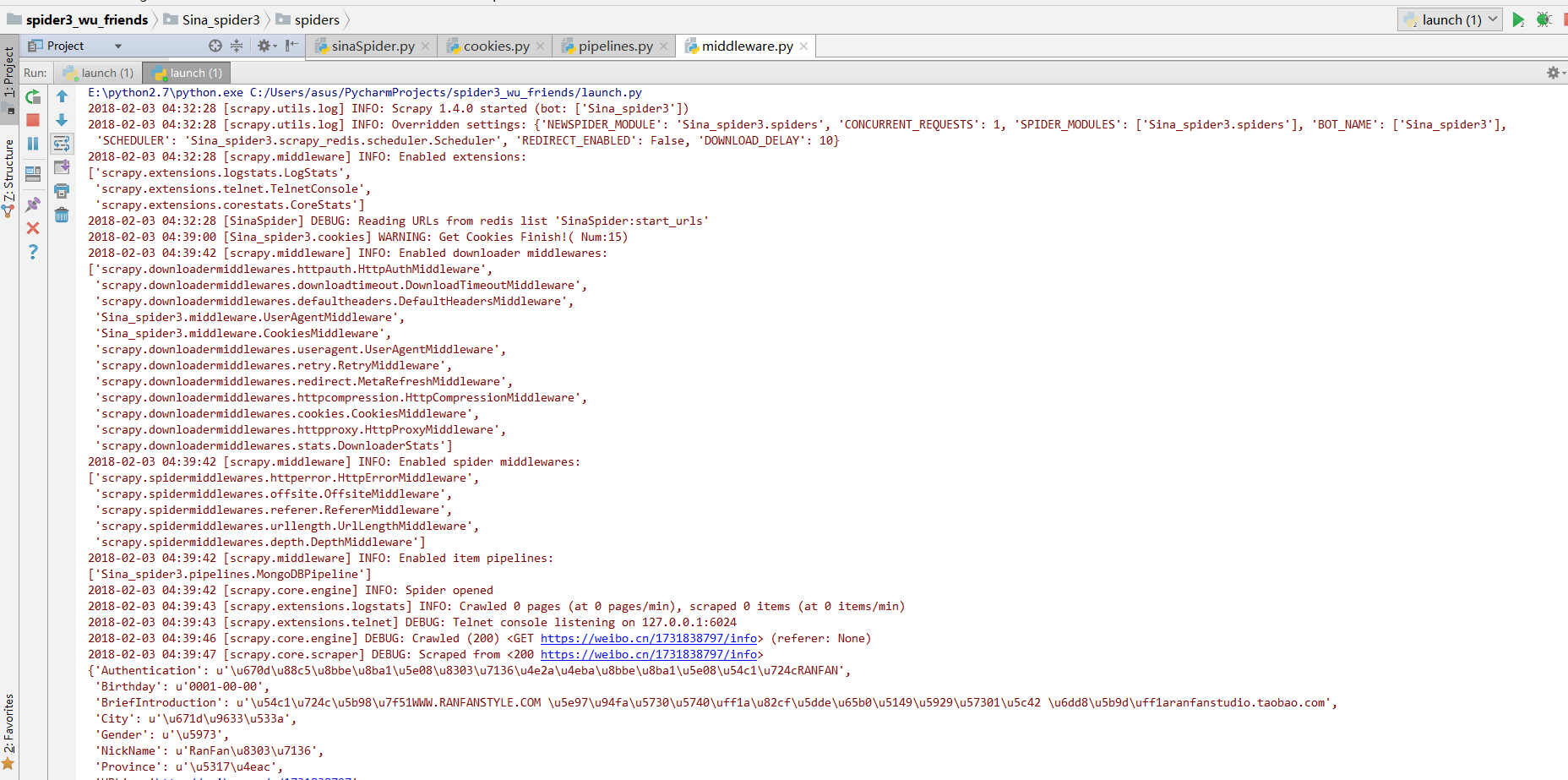
爬虫框架： Scrapy，使用 Scrapy+Redis 、Redis

**5.2运行状态及测试**

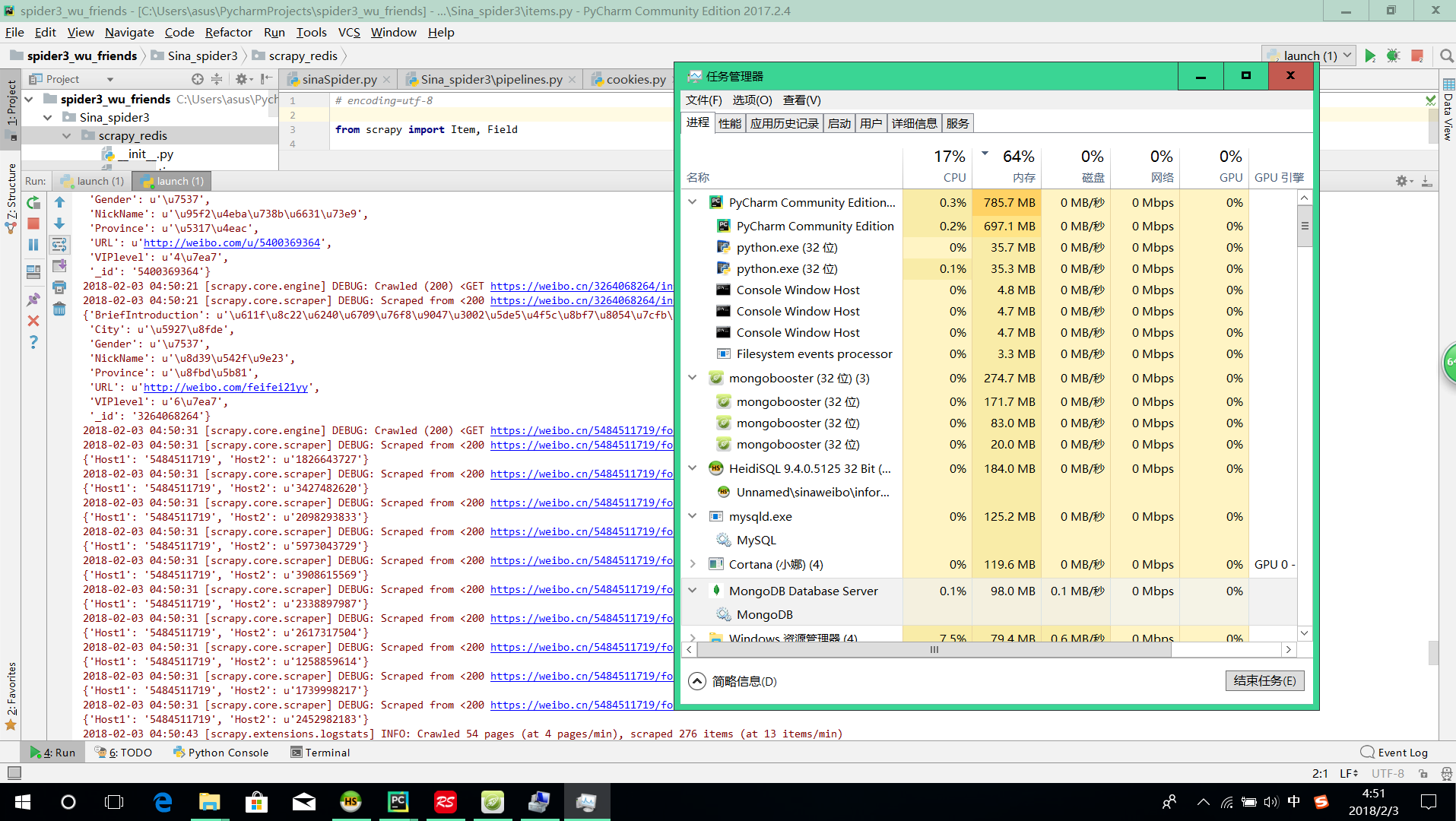
经测试，爬虫抓取微博的速度可以达到12个小时爬取400W条数据，具体要视网络情况。需要定期更换Cookies中的账号信息。建议大量预存。比如我在淘宝网上买了25个账号作为预存在Cookies上。

MongoDB性能良好。足以处理该量级的数据。下图为爬虫运行时的控制台

截图:



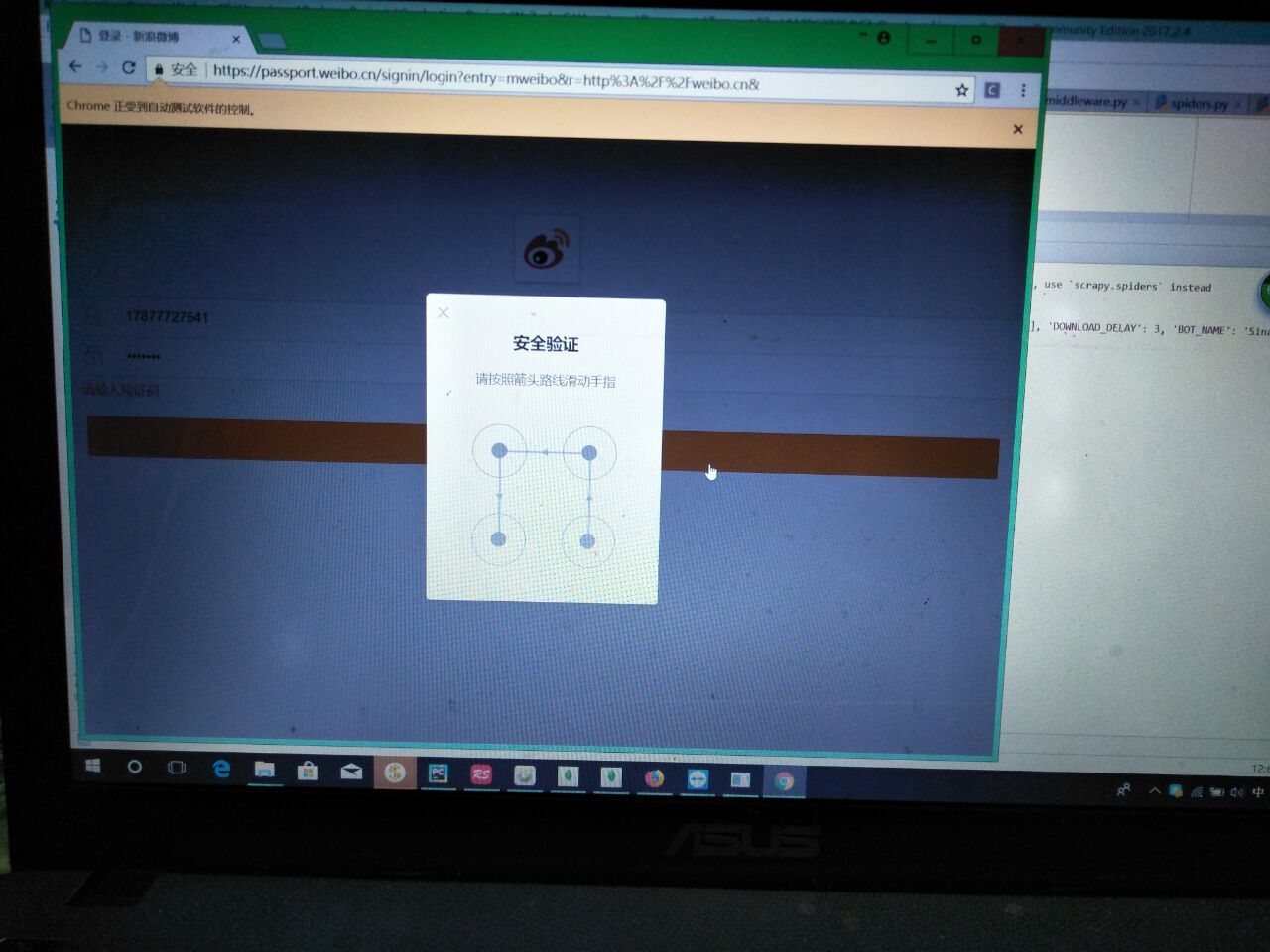
运行成功爬虫正在爬取



爬虫系统运行中各个状态

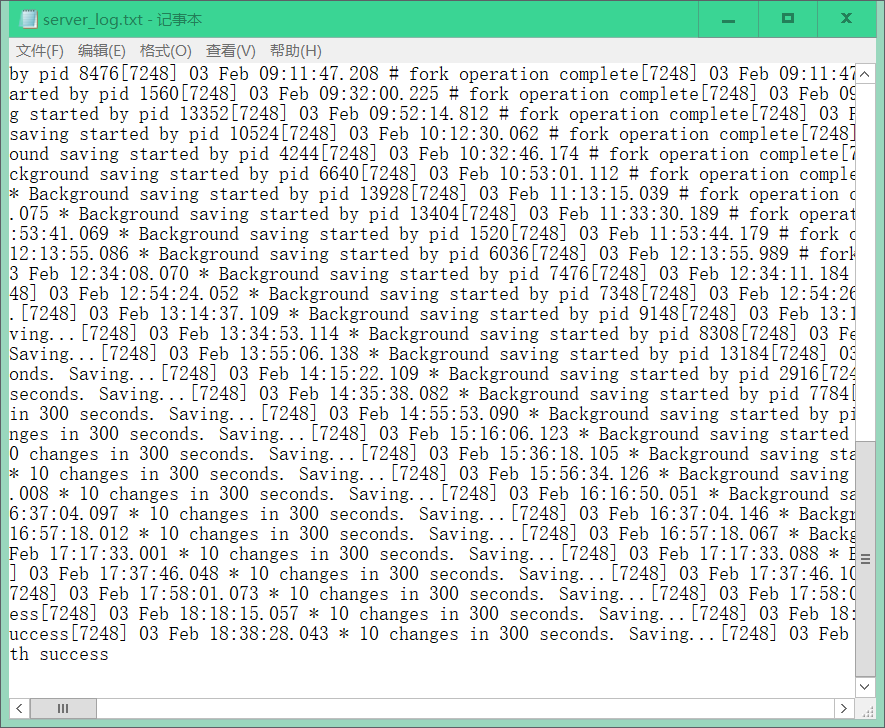
**5.3测试过程其他组件过程展示**

运行程序，登录移动版web微博，手动打码过程。

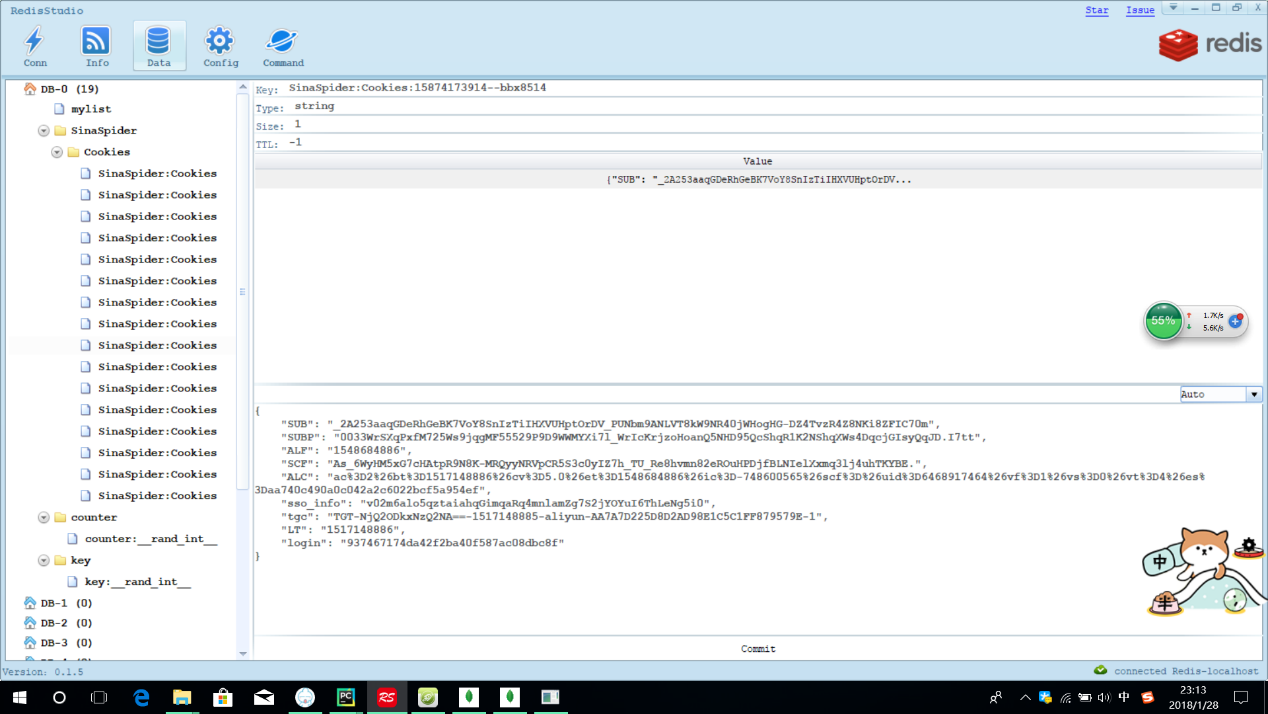


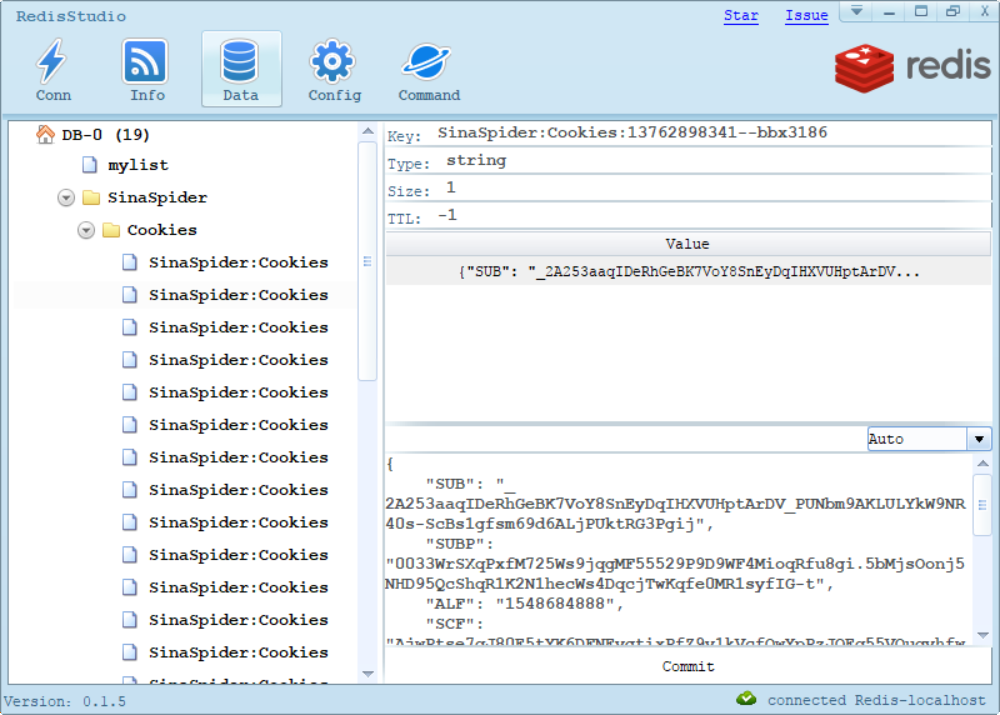
手动打码

以下是爬虫爬取的过程中Redis服务器的日志server\_log状态。



Redis的服务器日志

Redis在运行中

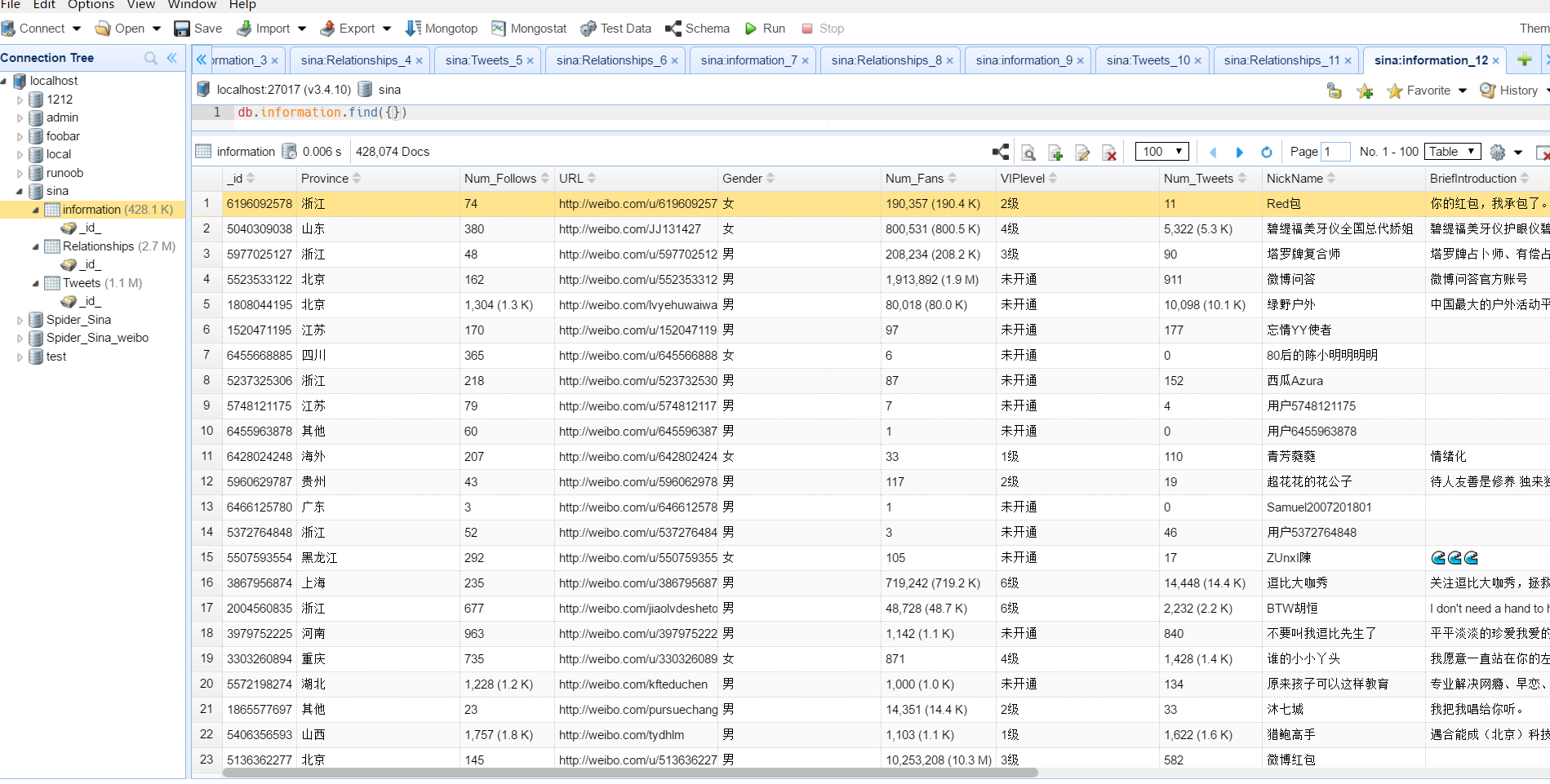


Redis中存储的Cookies

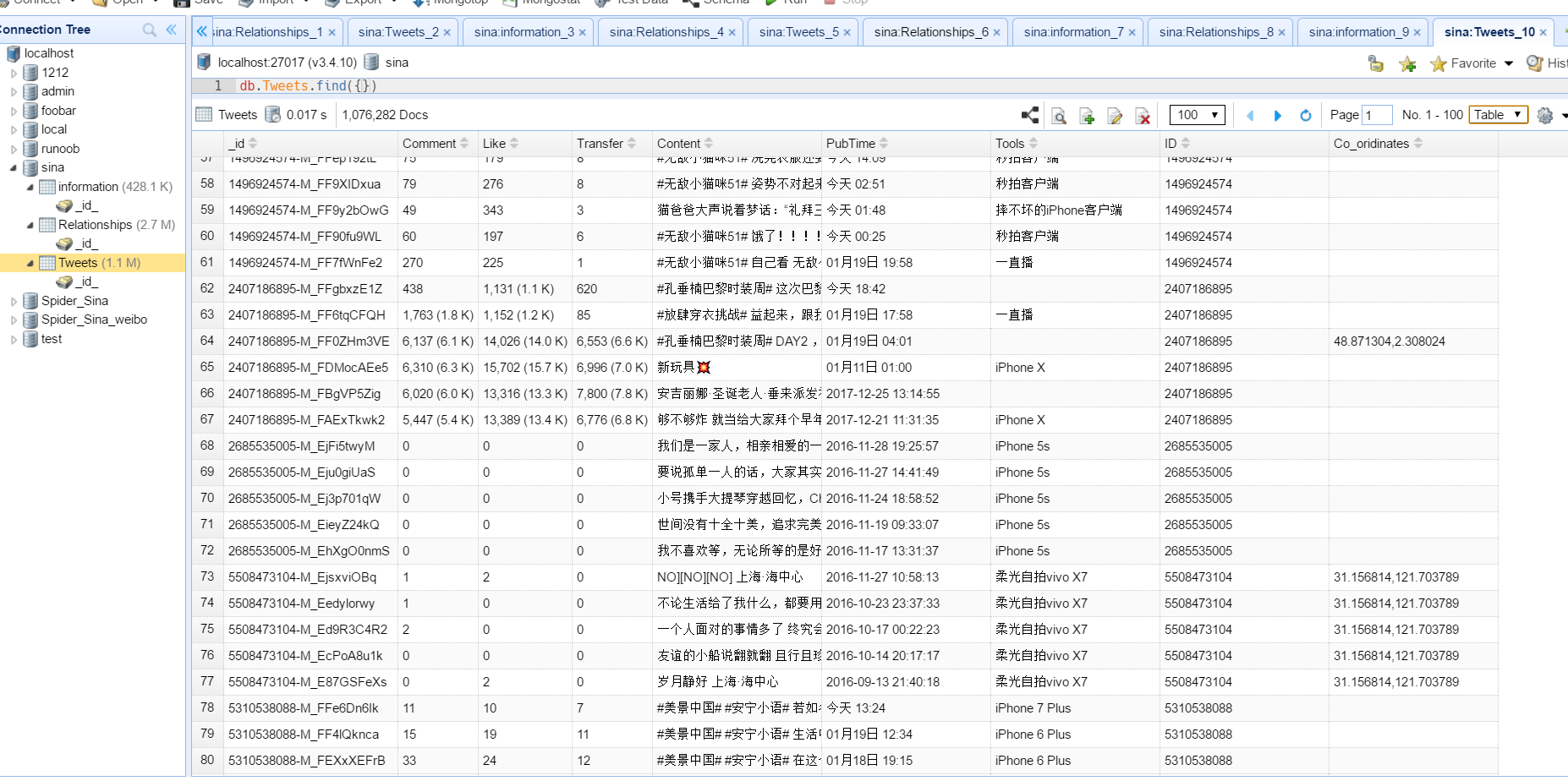
**5.4数据抓取结果展示与分析**

在程序运行过后得到名为”sina”的数据库，该数据库下保存有三张表，分别为

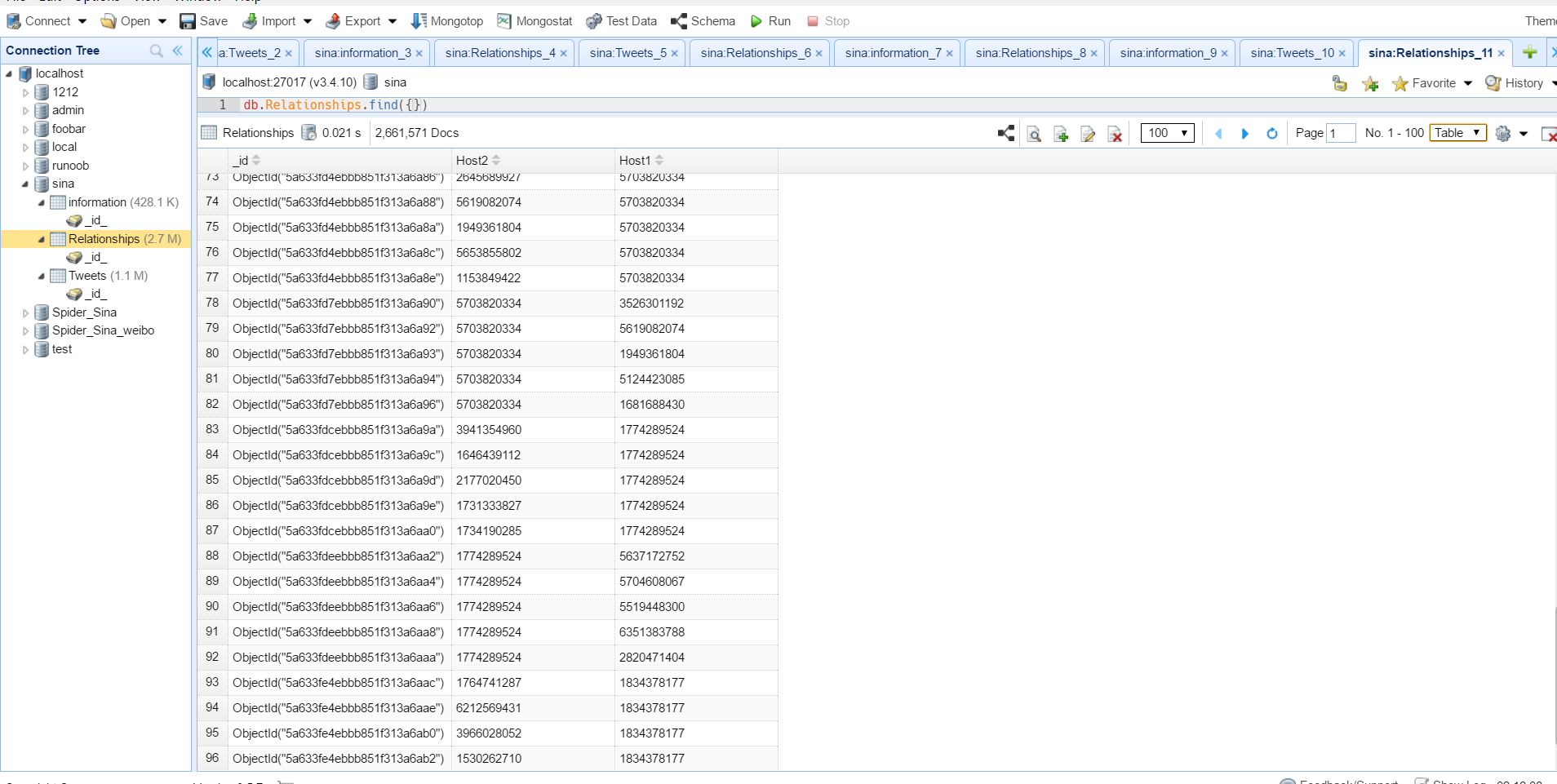
Information 、Tweets 、Relationships。实现存储了用户个人信息Information，用户发表的博客Tweets,微博用户关系表Relationships。



Information表



Tweets表



Relationships 表