**5.2.1数据分析**

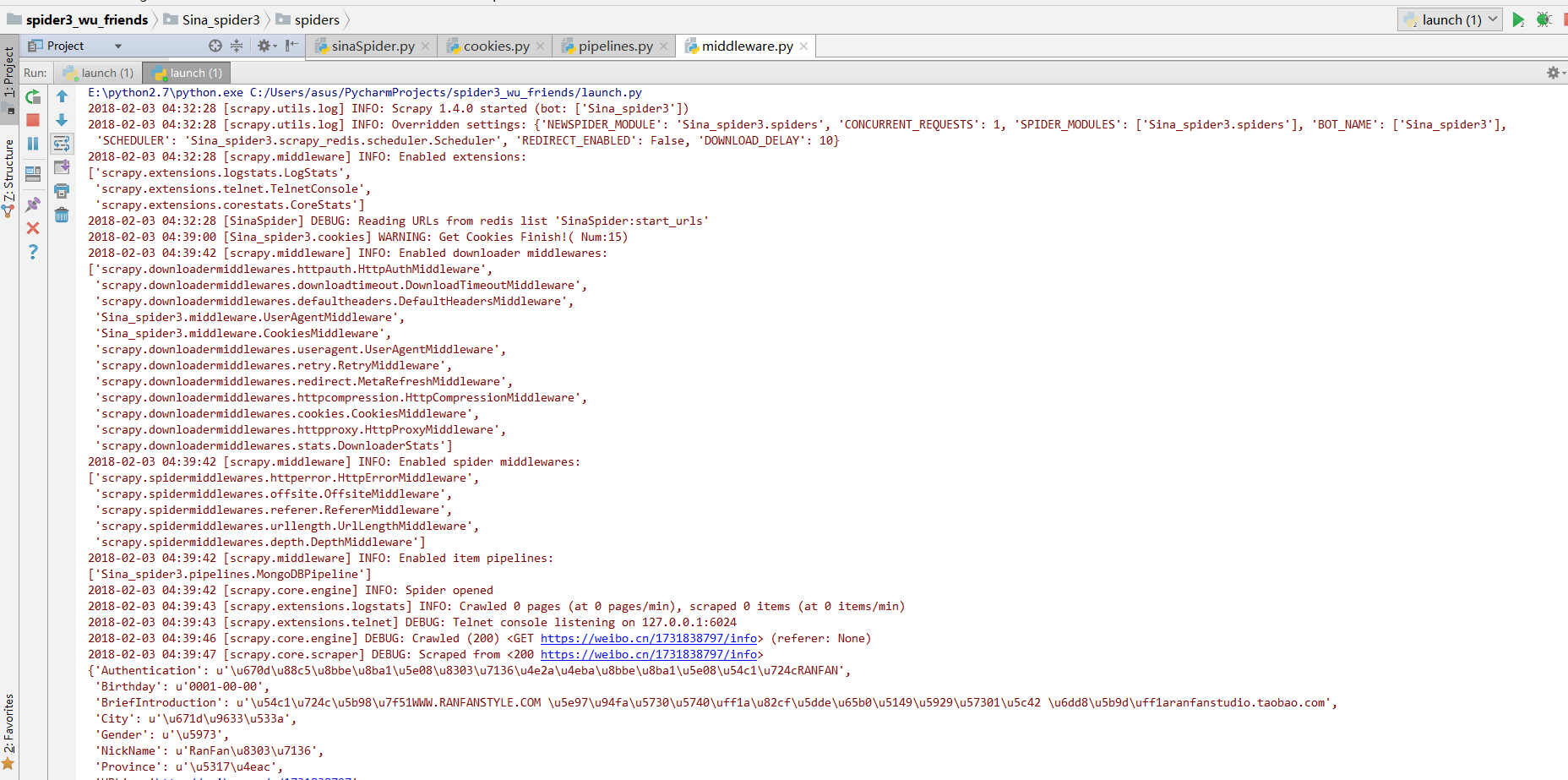
本论文是用Scrapy-Redis架构上的爬虫所爬取的数据，在Nosql类型的Mongodb数据库中的做简单的数据分析。爬了2天，因为写了暂停和继续功能，我基本是开一下关一下的，不过速度确实还不够快不快。我记得有个人 PHP 写的一天就能 1300w ！呢好啦，正式开始吧，上一章介绍了数据的E-R模型以及设计等等这里就不再啰嗦就简单介绍一下数据先：

用户发表的微博（\_id ，ID ，Content ，PubTime ，Co\_oridinates ，Tools ，Like ，Comment ，Transfer ）

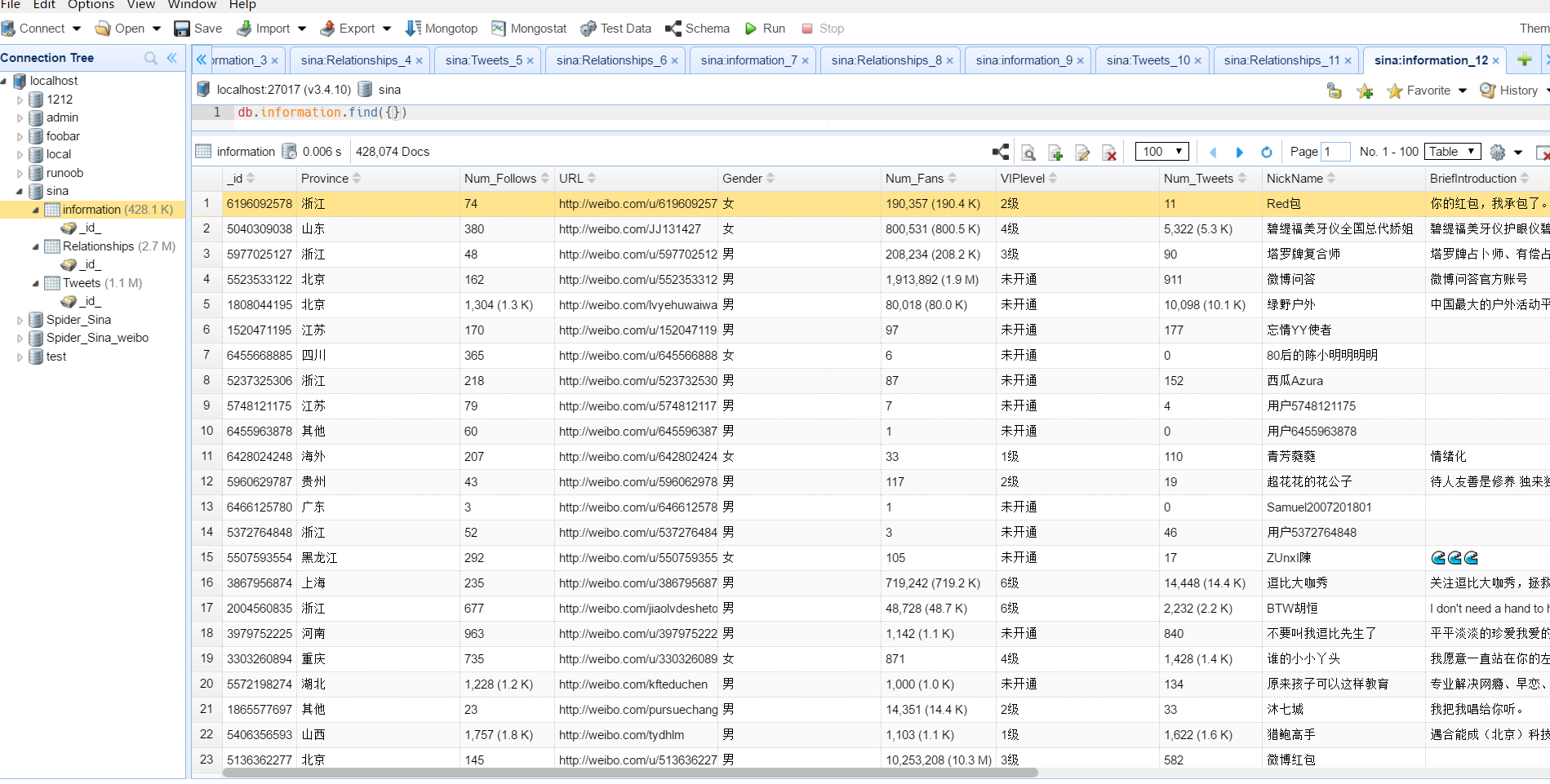
用户关系表（\_id，Host2，Host1）

用户个人信息（\_id，NickName，Gender，Province，City，BriefIntroduction，Birthday，Num\_Tweets:，Num\_Follows，Num\_Fans，SexOrientation，Sentiment，VIPlevel，Authentication，URL）

爬虫运行成功爬取数据中

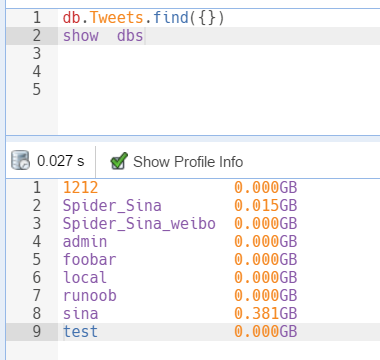


爬虫爬取的在mongodb中的数据



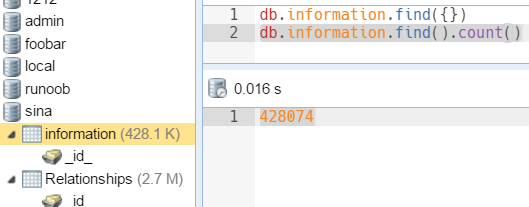
在Mongodb的可视化工具MongoBooster中写SQL语句做简单的数据分析：

>show dbs



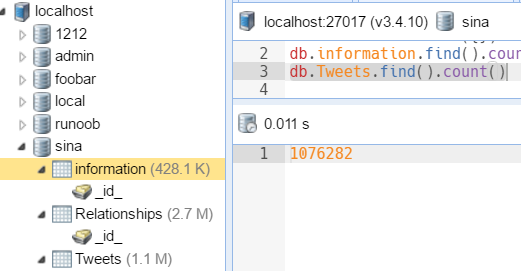
**1.统计各个表的数据总共多少条:**

>db.information.find().count()



用时0.016s information表一共428074条数据

> db.Tweets.find().count()

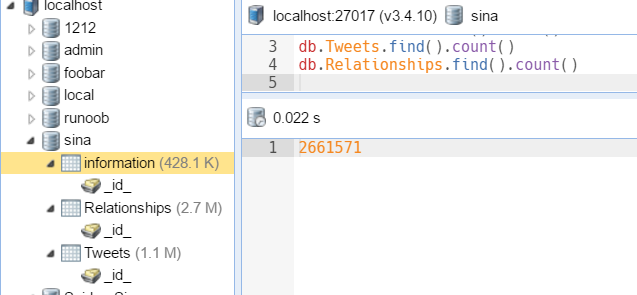


用时0.011s Tweets表一共1076282条数据

>db.Relationships.find().count()

用时0.022s Relationships表一共2661571条数据

所以我们爬虫爬取的数据有4165927条数据！

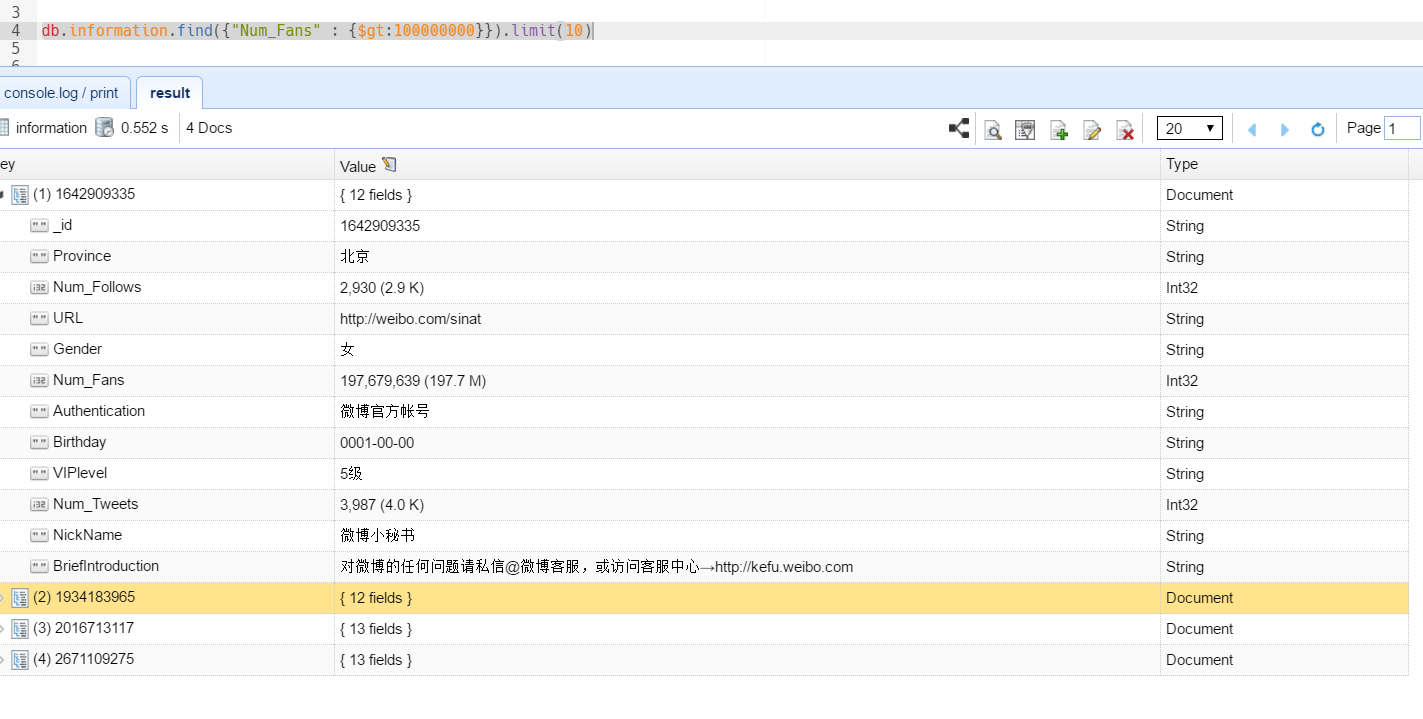


**2.查询粉丝数量大于一亿：**

这个最简单也最老套了吧，相信有一大堆人做过。

SQL 语句：

>db.information.find({"Num\_Fans" : {$gt:100000000}}).limit(10)



>db.information.find({"Num\_Fans" : {$gt:100000000}}).count()

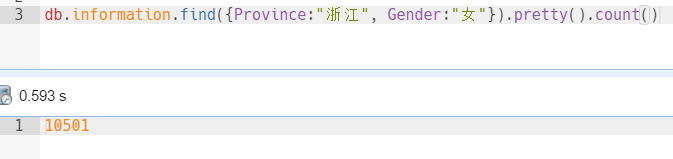
Result:1.027s 4

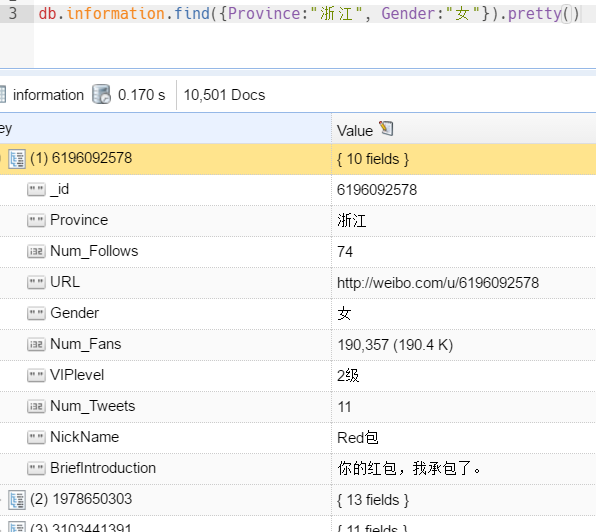
用时1.027s 总共4人

3.**查询浙江女用户情况：**

>db.information.find({Province:"浙江", Gender:"女"}).pretty()

>db.information.find({Province:"浙江", Gender:"女"}).pretty().count()





4.**性取向情况：**

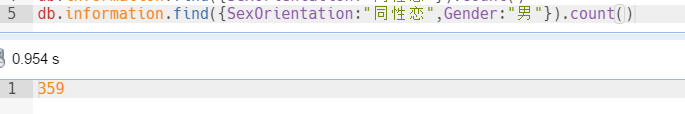
>db.information.find({SexOrientation:"同性恋"})

> db.information.find({SexOrientation:"同性恋"}).count()

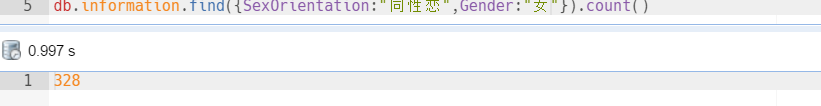
用时：0.635s 总共：687

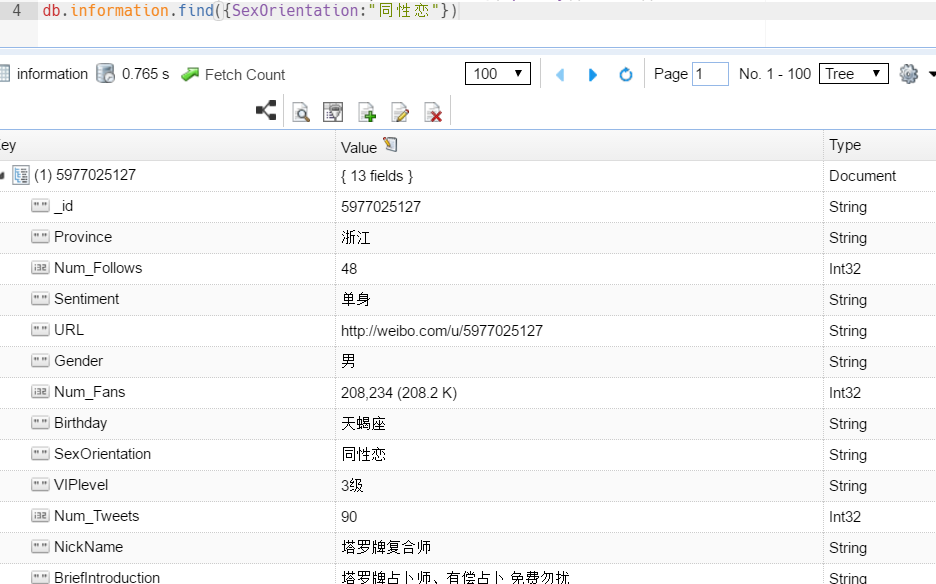
男同性恋情况：

>db.information.find({SexOrientation:"同性恋",Gender:"男"}).count()



女同性恋情况：





1.6**数据的导入导出：**

MongoDB的导出功能

mongodb提供了导入和导出的功能，分别是MongoDB下载目录下的mongoexport.exe和mongoimport.exe文件 ，具体的导出命令格式如下：

mongoexport -h dbhost -d dbname -c collectionName -o output

参数说明:

-h 数据库地址

-d 指明使用的库

-c 指明要导出的集合

-o 指明要导出的文件名（文件支持多种格式，如txt,wps,xls等）

在管理员权限CMD进入shell下：

mongoexport –host 127.0.0.1 -d sina -c Tweets -o Tweets.csv

MongoDB的导入功能

导入数据可以使用命令：

mongoimport -h dbhost -d dbname -c collectionname input

参数说明:

-h 数据库地址

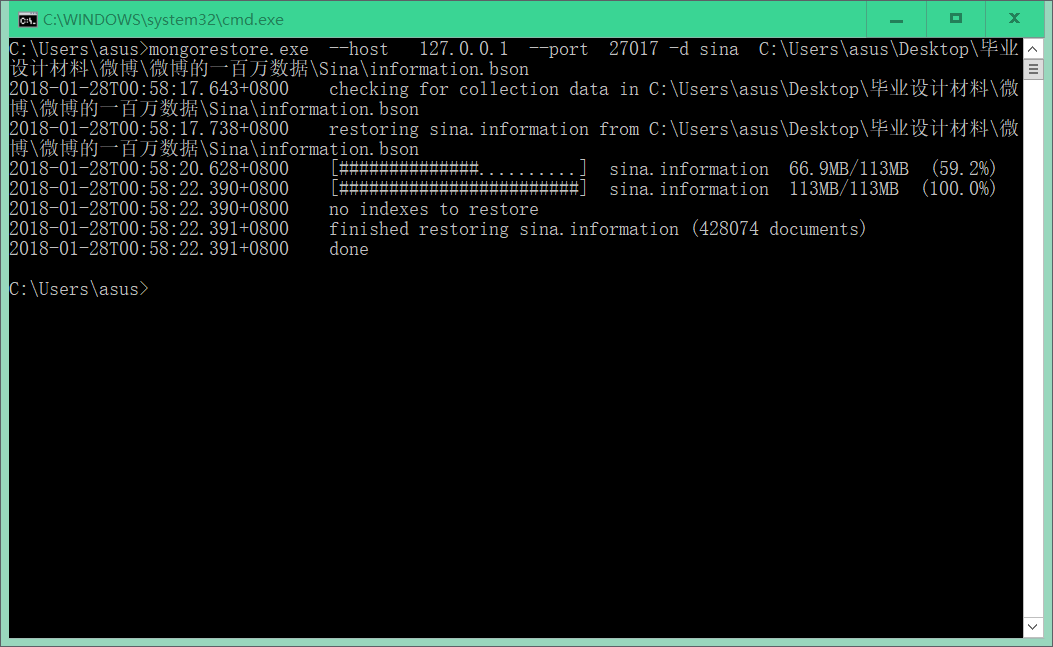
-d 指明使用的库

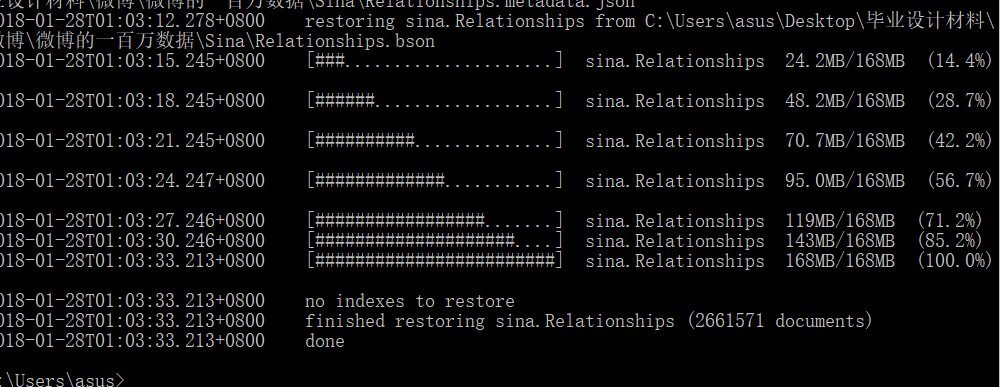
-c 指明要导入的集合（mongodb本身支持隐式创建，故事先无需创建集合）

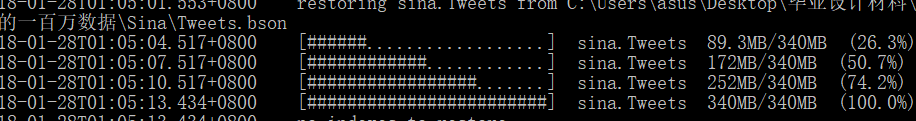
input 文件的地址

在管理员权限CMD进入shell下：

比如：Mongoimport --host 127.0.0.1 –port 27017 -d sina -c C:\user\asus\Desktop







用Excel就能打开cvs文件：

