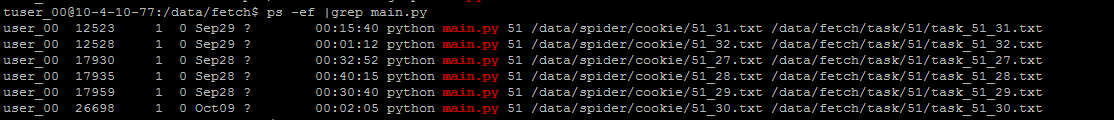
**抓取系统的现状以及常用操作**

# 机器部署情况（目录、文件、关键程序说明等描述）

## 216机器部署

1. 旧版spider程序，位于/data/spider目录之下，目前不在运行但是包含以往的简历数据文件
2. 新版本fetch程序，位于/data/fetch目录之下，主要运行51job六个账号进程和CJOL一个进程



目前cjol不在运行。。。可根据需要启动

1. Cookie更新响应的php文件，结合chrome helper插件用于更新cookie文件

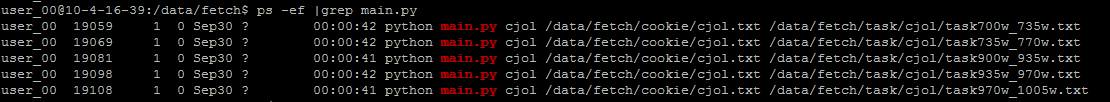
更新cookie的目录位于/data/spider/cookie之下，该目录可以在php文件之中进行修改。

1. 文件同步auto\_monitor.py,位于/data/spider/目录之下，用于更新其他两台机器上的cookie，监控本机的cjol.txt文件在文件修改之后上传到其他102、103两台机器

* 所监控的文件位于/data/spider/cookie/cjol.txt，上传其他两台机器的目录为/data/fetch/cookie/cjol.txt，可以根据需要在代码之中进行修改
* 目前所采用的ssh方式，程序之中涉及机器登录信息，可以在运行时将源码删掉，运行编译后的pyc文件；
* 或者采用其他的文件同步方式(比如从另外两台机器向本机的server请求然后更新)

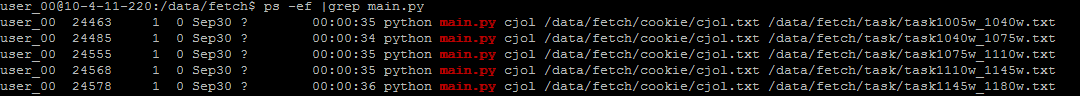
## 102机器部署

该机器主要用于部署新版的fetch程序，位于/data/fetch目录之下，主要运行CJOL五个进程



## 103机器部署

该机器主要用于部署新版的fetch程序，位于/data/fetch目录之下，主要运行CJOL五个进程



## 本机存储的部分文件

在靠近桌面的新机器上的保存了原有的智联抓取的简历和已经移动的CJOL的简历数据。

CJOL简历数据位于G:\cjol数据备份，移动了部分文件仍有大量数据存储于云端服务器

智联的简历位于E:\data目录之下

# 目前所发现问题

## 号段抽样

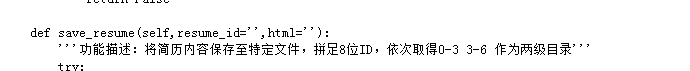
* Cjol号段跑到887万左右，后面的一些号段几乎没有发现有简历的存在，有必要进行抽样选取号段
* 51job号段区间相对较小，需要经常更换，否则就会出现命中率降低空跑的情况。

## 号段进行管理

有必要对于cjol、51job已经抓取过的号段进行记录管理，尤其是51job相对零散分布的这种号段区间；有效管理已经抓取的号段，避免重复抓取。

## 文件移动（磁盘空间）

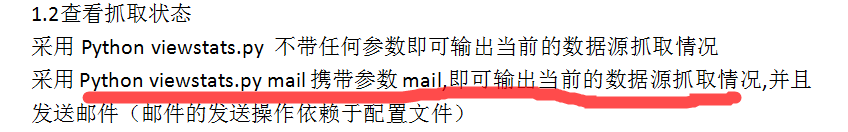
* 单纯从云端将文件移动到本地，可以采用多种方式 SSH、FTP、HTTP等，但是由于带宽的限制速度有些慢
* 从节约磁盘空间的角度考虑，对于目前没有压缩的HTML格式的简历文件进行手动压缩从内网进行转移，速度可观也可以暂时缓解磁盘空间的问题。
* 为了节约磁盘空间，减缓磁盘占用的增长率，那么可以调整抓取基类之中的简历保存函数，采用tarfile或者zipfile库进行文件的压缩存储。



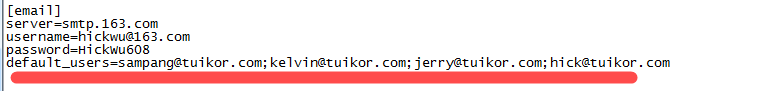
## 邮件的发送

1. 数据报告邮件的发送

数据报告邮件的发送采用《代码简要说明》里面查看任务状态以及邮件发送之中提到的：



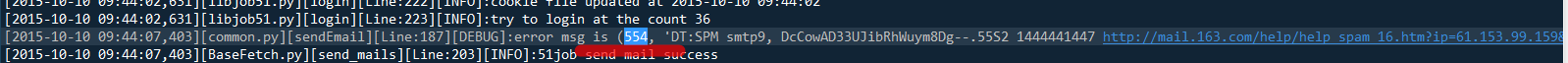
如需修改邮箱地址，可以直接修改/data/fetch/conf/basic.conf文件，无需重启，修改之后立即生效



为了方便操作，可以将该操作设定为系统定时任务，登录管理员增加contab任务即可。

1. 告警邮件的发送

告警邮件的发送偶尔会遇到554，反垃圾邮件



# 51job常见操作

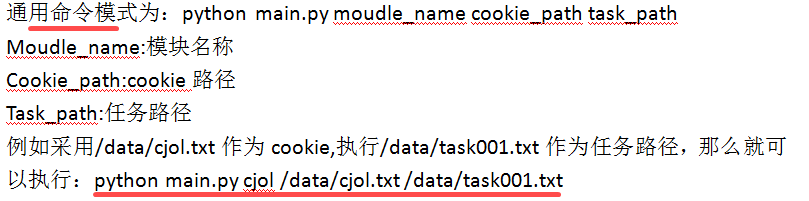
## 运行的号段以及号码信息

参考抓取情况txt

## 增加账号方法

增加账号进行抓取，新增加一个账号进行抓取：

* + - 1. 首先参考inti.php文件之中的ID和账号映射配置，确定一个不重复的新的ID号码，比如88；
      2. 为了方便登录,添加新增账号的信息到helper插件，需要在chrome插件之中进行配置账号快捷键,比如ALT+8，方便登录的一种方式，非必要环节
      3. 修改init.php文件，增加该账号和ID的映射关系，使得在浏览器登录之后可以在服务器特定位置创建cookie文件，比如ID为88则会创建cookie文件名为51-88.txt；
      4. 然后，而后创建一个对应任务文件，或抽样或手搜索确定一个没有抓取过的号段。
      5. 确定以上工作，参考执行方式，执行进程即可。



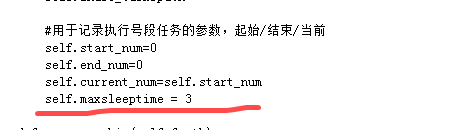
## 更换账号方法

更换账号无需增加新的ID，只是将原有的ID对应的账号信息进行替换即可。

1. 首先参考inti.php文件之中的ID和账号映射配置，确定一个需要更换的ID号码，比如66；
2. 更新账号的信息到helper插件，将原有的账号信息替换即可，无需更换快捷键
3. 修改init.php文件，修改账号和ID的映射关系，将ID对应的用户名修改为新的用户名即可；
4. 确定以上工作，在浏览器上进行登录即可更新远程的cookie文件，那么新的cookie对应新的账号。

## 修改sleep时间

为了避免51job每小时2000的限制，适当调整sleep时间，直接修改libjob51.py文件之中的maxsleeptime参数即可。



## 任务完成或者中途更换任务方法

1. 任务完成：任务完成之中进程会自动退出，需要重新制定任务文件之后将进程进行重启即可
2. 中途更换任务：中途更换任务，任务本身没有完成因为命中率或者其他原因需要更换，此时的操作不同的是需要将原来进程kill掉，然后修改任务文件重新启动。

# Cjol常见操作

## 已完成号段信息

参考抓取情况txt

## 任务完成或者中途更换任务方法

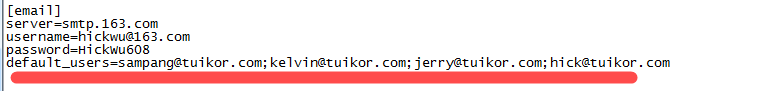
该处的操作和51类似，操作一样。

1. 任务完成：任务完成之中进程会自动退出，需要重新制定任务文件之后将进程进行重启即可
2. 中途更换任务：中途更换任务，任务本身没有完成因为命中率或者其他原因需要更换，此时的操作不同的是需要将原来进程kill掉，然后修改任务文件重新启动。

# 其他常见的操作

## 修改通知邮箱地址

如需修改邮箱地址，可以直接修改/data/fetch/conf/basic.conf文件，无需重启，修改之后立即生效。



需要注意的问题：

* 如果是分布式多台机器跑，报警可能会出现重复的邮件，避免该问题可以调整配置文件，比如附属的机器不设置邮箱地址。
* 另外邮件发送函数里面具有缺省的配置，当发送数据报告邮件并且如果此时邮箱地址列表为空，那么启用缺省的邮箱地址。

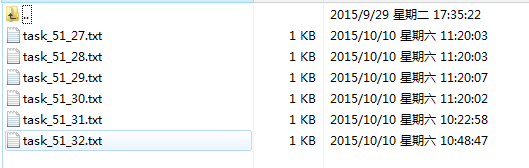


## 查看号段命中率

指定一个任务在执行时候，需要实时关注号段的简历抓取情况，简单分为以下三种：

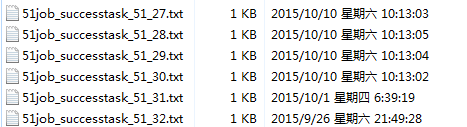
1. 从任务文件的更新时间看经常是否在进行ID执行

由于程序每执行一个ID均会更新任务文件，所以如果一个任务长期没有更新，那么或者进程退出或者cookie失效在等待登录



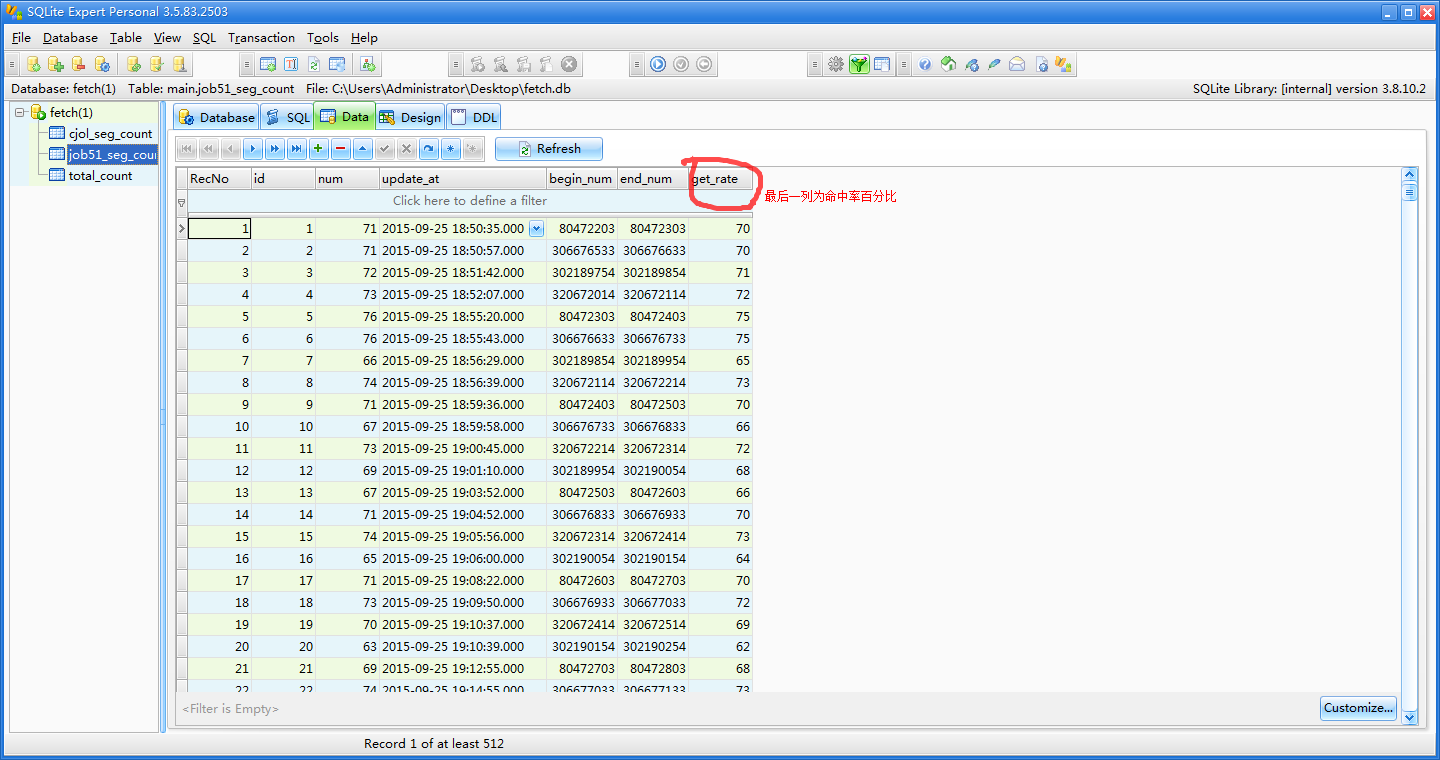
1. 从抓取简历成功记录的successXX.txt文件来看最近一个简历的抓取到的时间

程序每抓取成功一个简历就会记录一个计数到上述文件，并且同步更新到数据库总数计数



1. 查看数据库文件，可以看到号段的分段命中率

操作稍微繁琐，将远程的数据库拷贝出来，使用图形化的sqllite管理工具翻看数据库表格，可以看到分段的命中率。



# 程序改进点

## 邮件发送函数

目前并没有针对邮件内容的类型配置不同的通知邮箱地址，另外邮件的发送机制并没有重试机制，偶尔出现的554错误。

## 执行状态查看的脚本

执行状态查看的脚本里面，只是出现了简单的统计数字，后续可以增加上每个任务执行的账号、最近依次抓到简历的时间等信息

## 分号段的命中率查看

分号段的命中率查看依赖于数据库文件，可以尝试增加一个脚本利用程序读取数据库，设计时可以初步参考最近一个小时内的命中率记录、最近一天的方式或者也可以进行简单的叠加统计计算一个大的号段命中率

## 文件同步

分布式机器运行程序，需要将cookie同步更新到其他的目录，目前采用ssh方式，可以尝试在其他机器向主服务器请求的方式进行更新

## 任务调度

任务调度方式上，暂时没有实现，实现方式上推荐多进程模式或python微线程（操作比较繁琐），不推荐原始的Python线程效率很低。