1. 各文件功能说明

E:\pubmed\pubmed\spiders\author.py author spider 爬取信息存储到author.json文件

E:\pubmed\pubmed\spiders\authorPair.py authorPair spider 爬取信息存储到authorPair.json

E:\pubmed\extractEmail.py 从author.json及authorPair.json提取邮件信息并发送邮件

E:\pubmed\emailList.txt 提取到的邮箱列表

E:\pubmed\Time.ini 记录上次爬虫时间

1. 程序流程

* 读取Time.ini文件获取上次爬虫时间。第一次时间需手动写入，例如爬取2000/01/01到现在的文献，应写入2000 01 01 分三行写入
* 根据读取结果，重置start\_urls（开始爬虫网址列表）
* 执行爬虫命令，生成author.json及authorPair.json文件。其中author.json不包含“email”key，邮箱信息在authorInformation中；authorPair.json包含“email”key，可直接提取对应value
* 重写Time.ini文件。将当前系统时间写入Time.ini
* 提取email-add信息。其中author.json不包含“email”key，邮箱信息在authorInformation中；authorPair.json包含“email”key，可直接提取对应value
* Email-add信息去重，邮箱列表写入emailList.txt
* 读取emailList.txt，发送邮件

1. 操作说明
2. 安装python2.7，安装scrapy

* 安装Anaconda2-4.0.0-Windows-x86\_64.exe
* pip install scrapy

1. 将pubmed工程拷贝到任意目录。例如E:\
2. 打开dos窗口，切换路径到pubmed工程路径

例如：依次执行e: cd e:\pubmed

1. 执行命令python execute.py

PS:若爬虫关键词、第一次爬虫时间需修改，直接修改author.py及authorPair.py中start\_urls即可

C:\Users\todaysoft\AppData\Roaming\Tencent\Users\1094253356\QQ\WinTemp\RichOle\{79PW`6LDOV$S8}{C[Z9BHN.png