weinxin.sogou.com 帐号抓取与匹配

Added by 候贤旭, last edited by 候贤旭 on Aug 06, 2014

一、系统准备：

1. ubuntu 12.04等Linux操作系统
2. Python 2.7
3. Redis2.8.13

二、Python 模块安装：

1. sudo pip install beautifulsoup4
2. sudo pip install redis
3. sudo pip install MySQLdb 若失败的话，用 sudo apt-get install python-mysqldb 安装

三、文件结构：

```
├── config.py --> redis、mysql、ip代理等参数设定
├── connect_with_proxy_ip_and_fake_ua.py --> 利用代理和Fake-ua 来连接网络（基于urllib2模块）
├── crawl_weixin_sogou_main.py – > 启动帐号匹配程序
├── craw_weixin_info_from_homepage.py -> 从weixin.sogou.com上每个微信号主页获取微信帐号信息
├── help_func_for_user_account_crawl.py --> 解析网页的相关函数
├── mysql_to_redis.py – > 将本地mysql的weibo_id, weibo_name 和 account_id，写入Redis
├── new_user_account_crawl.py --> 利用关键词抓取新的帐号信息
├── tweet_crawl.py --> 对所有帐号的最近三条tweet进行抓取
├── output_all_weixin_info_from_redis_to_file.py --> 将redis里的数据导入本地文件中
├── get_latest_date_to_redis.py --> 从抓取的TWEET获取每个帐号最近TWEET时间
└── user_account_verification.py --> 对所有帐号进行匹配，并抓取帐号的其它信息
```

四、修改配置文件 config.py

1. Redis: 在本地搭建服务，即host='localhost'
2. 设置公司MySQL服务器地址及用户名和密码
3. 设置代理IP及Fake_ua
4. 设置weixin.sogou.com最大的搜索页（目前为20）
5. 设置每个帐号TWEET的抓取条数（目前为3）
6. 设置程序运行的线程数（目前为10）

五、任务分类及运行：（直接将文件拷到本地）

1. **帐号匹配**：
   运行：python crawl_weixin_sogou_main.py
   a. crawl_weixin_sogou_main.py
   b. mysql_to_redis.py
   c. user_account_verification.py
   d. config.py
   e. connect_with_proxy_ip_and_fake_ua.py
   f. help_func_for_user_account_crawl.py
   g. craw_weixin_info_from_homepage.py

2. **帐号补充**：
   运行：python new_user_account_crawl.py
   a. new_user_account_crawl.py
   b. config.py
   c. connect_with_proxy_ip_and_fake_ua.py
   d. help_func_for_user_account_crawl.py

3. **tweet抓取**：
   运行：python tweet_crawl.py --> 在本地生成目录"TWEET_HTML/yyyymmdd"，存放帐号的TWEET信息
   a. tweet_crawl.py
   b. config.py
   c. connect_with_proxy_ip_and_fake_ua

4. 最近TWEET日期获取：
   运行：python get_latest_date_to_redis.py TWEET_HTML/yyyymmdd/
   a. get_latest_date_to_redis.py
   b. config.py

四、抓取流程图

Labels None

1 Comment

林晨 about 2 hours ago

添加部署相关的内容：
1. 系统准备 (安装redis_server, mysqldb)
2. Python模块安装
3. 拷贝代码: /var/wwww
4. 修改配置文件
5. 运行