### 运行环境：

Windows 7/10, mysql 5.6, python 3.6,

所需第三方python模块：requests

### 流程介绍：

从qq空间爬取照片主要分为两个步骤。

第一个步骤是获取qq空间相册和照片的url。这个步骤通过向QQ空间服务器发送特定的Http请求，数据会以json格式返回。需要注意的是，这一步骤需要模拟登陆QQ空间，服务器会通过QQ账号检测用户的访问信息。如果请求频率太快，就会被QQ的反爬虫策略识别，暂时屏蔽该账号的访问。获取的数据除了url还有照片的一些其他相关信息，包括名称，拍摄时间，所属相册等。所有的信息存储到了一张mysql数据库表中。一条记录对应一张照片。此步骤对应URL下载器脚本getinfo.py

第二个步骤是根据已获取的URL下载照片。读取数据库，下载照片，照片名称对应数据库中的编号。所有照片按照所属qq放在一个文件夹，文件夹名称就是qq号。此步骤对应下载器python程序，downloader.py.

#### 数据库表设计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名称 | 名称 | 类型 | 可空 | 备注 |
| id | 主键id | bigint | Not null | 主键,自增 |
| url | 照片url | varchar |  |  |
| width | 照片宽度 | int |  |  |
| height | 照片高度 | int |  |  |
| pic\_name | 照片名称（用户设置） | Varchar(255) |  |  |
| liked\_count | 照片被点赞次数 | int |  |  |
| comment\_count | 照片被评价次数 | int |  |  |
| upload\_time | 上传时间戳 | Varchar(255) |  |  |
| modifiy\_time | 修改时间戳 | Varchar(255) |  |  |
| shoot\_time | 照片拍摄时间戳 | Varchar(255) |  |  |
| album\_id | 所处相册id | Varchar(255) |  |  |
| album\_name | 所处相册名称 | Varchar(255) |  |  |
| qq\_number | 用户qq号 | Varchar(255) |  |  |
| is\_downloaded | 是否已下载 | bit |  | 默认为0 |
| gmt\_downloaded | 下载时间 | datetime |  | 默认为空 |
| gmt\_create | 记录创建时间 | datetime |  |  |
| gmt\_modified | 记录修改时间 | datetime |  |  |
| is\_deleted | 是否删除 | bit |  | 默认为0 |

文件夹中qzone\_photo\_struct.sql是建表的sql语句，在mysql中运行就可以了。

### URL下载器使用步骤

1. 浏览器登录QQ空间
2. 获取在qq空间域名的cookie（可以通过chrome,IE等浏览器的开发者工具获取），粘贴到cookie文件中。
3. 配置configuration1.txt 文件中的配置数据
4. Windows环境下，cd 命令切换到getinfo.py程序所在目录（请确保configuration1.txt,cookie文件也在此路径下）
5. 运行命令python getinfo.py arg11 arg12,其中（arg12为起始QQ号，arg12为向后遍历qq号次数， 比如从qq 5678910开始，采集10000个qq号，

命令为: python getinfo.py 5678910 10000）

#### configuration1.txt 配置项目说明

|  |  |  |  |
| --- | --- | --- | --- |
| 项目名称 | 说明 | 默认值 | 备注 |
| timeOutLimit | 最大http请求等待时间（S） | 10 | 可以根据网速调整 |
| maxSleepTime | 每次采集一个qq后最大等待时间 | 100 | 为了限制请求频率，防止被反爬虫，建议该值不要低于50 |
| dbhost | 数据库主机名 | localhost | 如果使用远程数据库，该值应为服务器ip |
| dbuser | 数据库登录用户名 |  |  |
| dbpasswd | 数据库登录密码 |  |  |
| db | 数据库名称 |  |  |
| port | 数据库服务端口 |  |  |
| charset | 数据库编码 | utf-8 |  |
| theheaders | http请求头 |  |  |

照片下载器使用步骤

1. 按需求，配置configuration2.txt文件
2. Windows环境下，cd 命令切换到downloader.py程序所在目录（请确保configuration2.txt文件也在此路径下）
3. 运行命令 python downloader.py arg11 arg12 (其中arg11是起始照片数据库id, arg12是向后遍历下载张数)

#### Configuration2.txt 配置项目说明

|  |  |  |  |
| --- | --- | --- | --- |
| 项目名称 | 说明 | 默认值 | 备注 |
| timeOutLimit | 最大http请求等待时间（S） | 10 | 可以根据网速调整 |
| dbhost | 数据库主机名 | localhost | 如果使用远程数据库，该值应为服务器ip |
| dbuser | 数据库登录用户名 |  |  |
| dbpasswd | 数据库登录密码 |  |  |
| db | 数据库名称 |  |  |
| port | 数据库服务端口 |  |  |
| charset | 数据库编码 | utf-8 |  |
| path | 照片下载路径 |  | 注意，该路径两端要加单引号 |

### 多人协作

整个照片爬取流程的速度瓶颈在于第一部分URL的下载。这部分需要QQ用户登录，如果请求频率太快容易被封号。可以多人用自己的qq号运行url下载器，提前划分好采集qq号段，并行工作。最好在远程服务器建立一个公共的数据库，大家的数据通过公共数据库共享，这样有利于进一步数据处理和操作。

此程序写的比较匆忙，没有经过很全面的测试，运行过程中可能会出现一些问题。如果大家在使用的过程中有问题或者发现Bug，请联系我。

李鑫 QQ:1136438564