

# 人人网爬虫技术文档

一、整体爬虫思路.....	2
1.1 爬虫目的.....	2
1.1 爬虫整体概述.....	2
1.2 爬虫思路框图.....	2
二、爬虫代码剖析.....	4
2.1 爬取人人网中用户的 ID.....	4
2.1.1 网页解析流程.....	4
2.1.2 详细代码解析.....	7
2.2 根据 ID 爬取用户信息.....	8
2.2.1 网页解析流程.....	8
2.2.2 详细代码解析.....	10
三、代码运行和示例.....	12
3.1 代码运行.....	12
3.1.1 爬 ID 代码运行方式.....	12
3.1.2 爬图片代码运行方式.....	12
3.2 示例.....	12
3.2.1 爬 ID 代码示例.....	12
3.2.2 爬图片代码示例.....	12

## 一、整体爬虫思路

### 1.1 爬虫目的

本爬虫项目主要为人脸算法提供人脸数据，因为人人网包含大量实名制用户且用户在其主页的相册中上传大量照片，我们可以针对每一个用户爬取其照片，每个用户为一个类作为训练集提供可靠的大量的人脸数据。

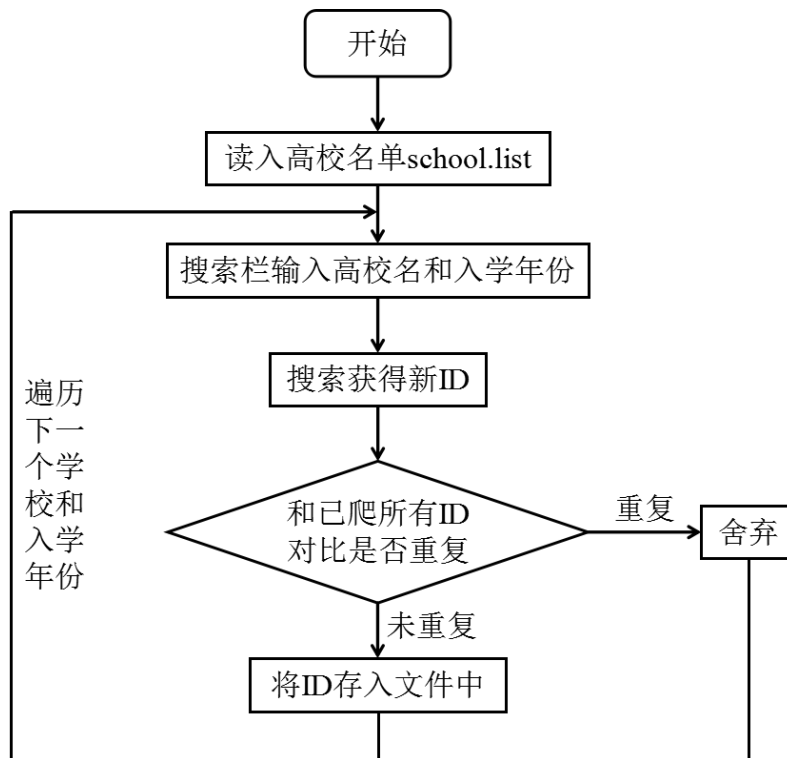
### 1.1 爬虫整体概述

人人网爬图主要分成两个部分：

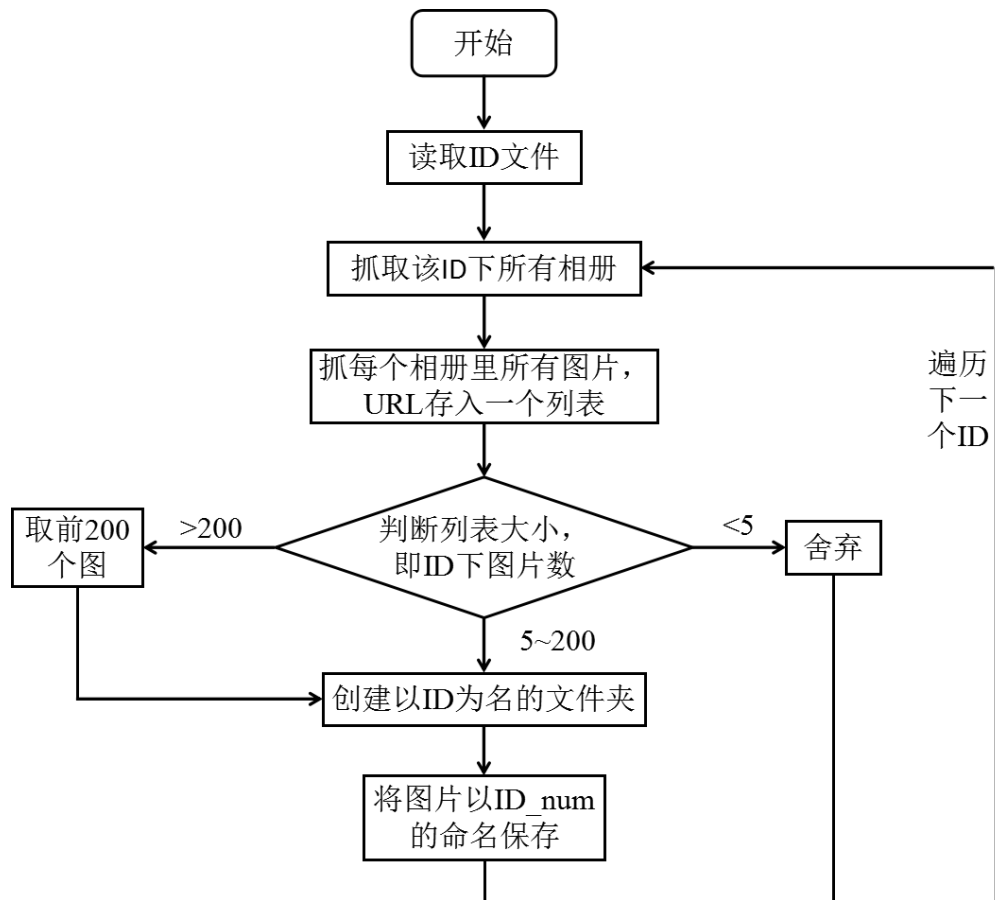
- (1) 爬取人人网中用户的 ID；
- (2) 根据 ID 爬取该用户的相册图片。

### 1.2 爬虫思路框图

- (1) 爬取人人网中用户的 ID：



(2) 根据 ID 爬取该用户的信息

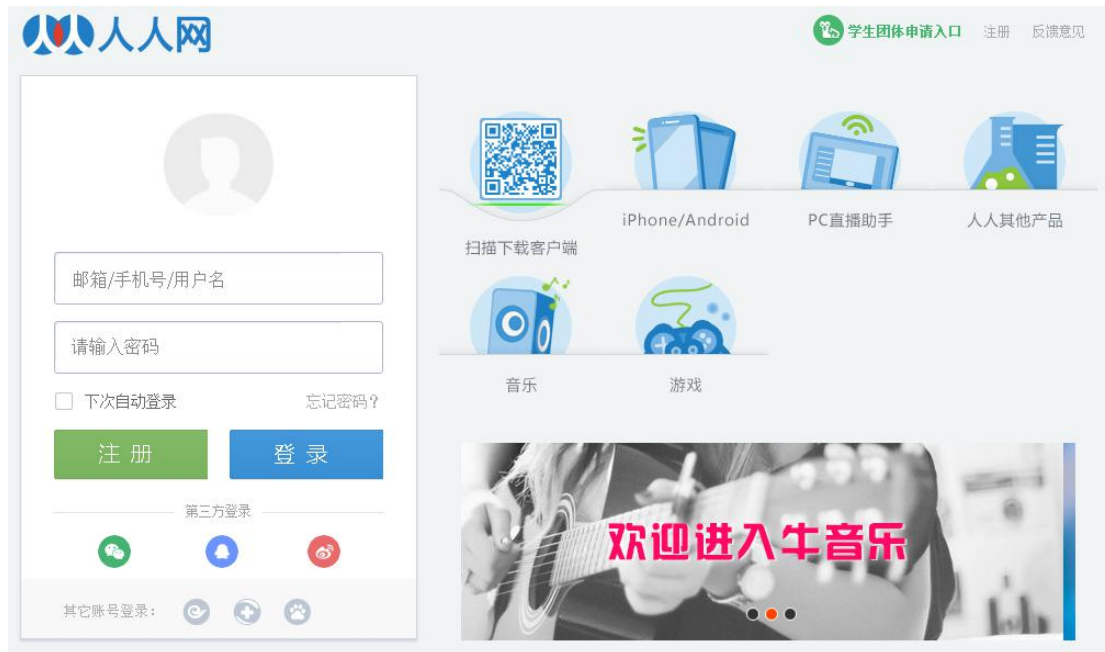


## 二、爬虫代码剖析

### 2.1 爬取人人网中用户的 ID

#### 2.1.1 网页解析流程

登录人人网主页：<http://www.renren.com>



首先需要注册一个账号，然后根据账号登录进去，主页如下，点击左侧个人主页：



进入个人主页后，点击上方搜索：



在搜索界面点击找人，根据下面提示的关键词来选择条件找出需要爬出哪一类人的信息

人人网

搜索好友，公共主页，状态

Q

全部

找人

相册

状态

分享

公共主页

更多»

搜索

高

热门搜索：baby 公主脸 月球 psy 找到1437个结果

筛选条件：使用筛选缩小搜索范围

请选择大学

院系

入学年

请选择高中

性别

☐ 仅显示非好友

工作单位

请选择初中

小学

家乡

年龄段

星座

更多搜索条件

如设置条件大学：“清华大学”，入学年：“2010”：

ren.com/s/all?from=opensearch&q=#qt=/tindex=2

人人网

搜索好友，公共主页，状态

Q

全部

找人

相册

状态

分享

公共主页

更多»

搜索

热门搜索：baby 公主脸 月球 psy 找到103841个结果

筛选条件：大学:清华大学 × 大学入学年份:2010 ×

清华大学

院系

2010

请选择高中

性别

☐ 仅显示非好友

工作单位

请选择初中

小学


家乡

年龄段

星座


更多搜索条件

排序：默认 人气




董黄伟 人气3004  
21位共同好友

+ 关注好友



沈震 人气5186  
18位共同好友

+ 关注好友

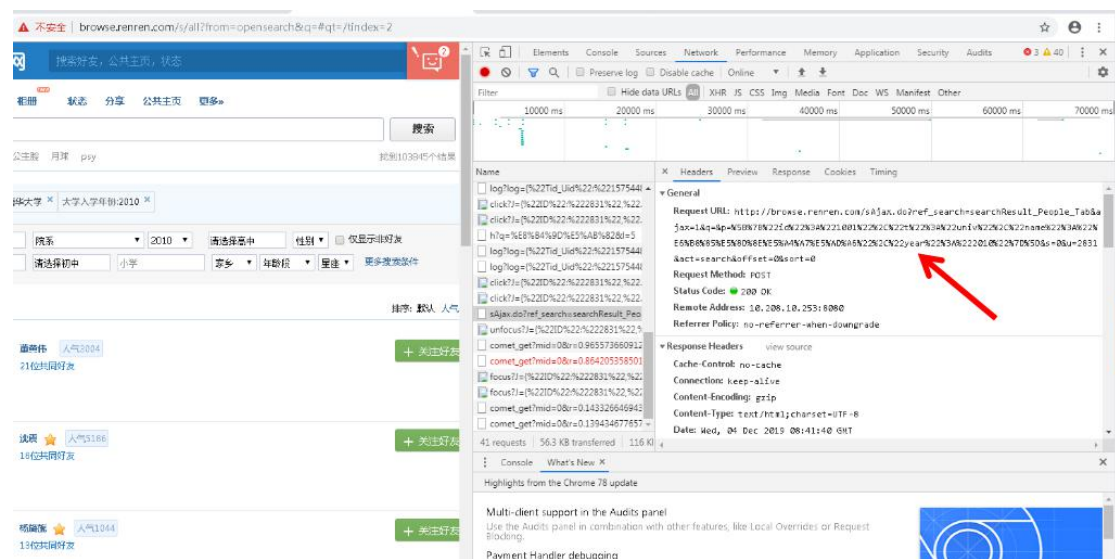


杨施施 人气1044  
13位共同好友

+ 关注好友

显示每页 10 个用户，按下 F12，打开 network，按下页面中的登录按键，可以找到

request\_url 链接，根据该链接可以直接用于后期直接根据关键词查找用户 ID。



解析成网址：

[http://browse.renren.com/sAjax.do?ref\\_search=&q=&p=%5B%7B%22t%22%3A%22univ%22%2C%22name%22%3A%22E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%22%2C%22id%22%3A%221001%22%2C%22year%22%3A%222011%22%7D%5D&s=0&u=2831&act=search&offset=0&sort=0](http://browse.renren.com/sAjax.do?ref_search=&q=&p=%5B%7B%22t%22%3A%22univ%22%2C%22name%22%3A%22E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%22%2C%22id%22%3A%221001%22%2C%22year%22%3A%222011%22%7D%5D&s=0&u=2831&act=search&offset=0&sort=0)

其中：网页中的“univ”和“year”后所跟的字符表示搜索的关键词，offset=0 中的“0”表示从第 1 个人开始所在的页面（一页 10 个人），如果是第 11 个人开始就是 offset=10。

经过测试，univ 即大学的标签后有包括汉字学校名（例：清华大学）和学校数字代码（例：1001），只修改汉字学校名链接并不会跳转，修改学校数字代码才会跳转，所以在获取不同用户的页面中，只需修改学校数字代码即可。bu

← → ↻ 不安全 | browse.renren.com/sAjax.do?ref\_search=searchResult\_People\_Tab&ajax=1&q=&p=%5B%7B%22t%22%3A%22univ%22%2C%22name%22%3A%22E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%22%2C%22id%22%3A%221001%22%2C%22year%22%3A%222011%22%7D%5D&s=0&u=2831&act=search&offset=0&sort=0

排序:默认 人气

1. **董黄伟** 人气3004  
21位共同好友  
关注好友
2. **沈震** ★ 人气5186  
18位共同好友  
关注好友
3. **杨琦璇** ★ 人气1044  
13位共同好友  
关注好友
4. **紫雪花GU** ★ 人气4462  
10位共同好友  
关注好友

查看该页面的源码如下图，

```

<html>
  <head>...</head>
  <body>
    <div class="list-mod list-follow">
      <div class="search_content_header public_header">...</div>
      <ol class="fl search_log" style="width: 100%; id="active_2012_module">
        <li>
          <p class="avatar Userimg">...</p>
          <div class="info">
            <dl>
              <dd>
                <strong> == $0
                  <a href="http://www.renren.com/profile.do?ref=searchresult_0&id=240135305&q=[p=[\u5866\,\,\"year\":\\"2010\"];]_s=0|u=2831&act=name&rt=user&in=0&ft=2&hh=1" target="_blank">董黄伟</a>
                </strong>
                <span></span>
                <span></span>
                <a class="user_lively" popval="{\"friendNum\":0,\"popValue\":3004,\"friendRank\":0,\"passive\":0,\"active\":0,\"viewCount\":0}">人气3004</a>
              </dd>
            </dd>
          <dd style="clear:left;">...</dd>
        </dl>
      </div>
      <ul class="actions">...</ul>
    </li>
    <li>...</li>
    <li>...</li>
    <li>...</li>
    <li>...</li>
  </body>
</html>

```

就可以检索出的某用户所对应的 ID 号。

### 2.1.2 详细代码解析

首先，post\_url 代表登录界面的网页，同时需要 post\_data 为账号密码信息，通过 session.post 请求访问人人网。

```

session = requests.session()
# 登录的表单 url
post_url = "http://www.renren.com/PLogin.do"
post_data = {"email": "*****", "password": "*****"}
# 使用 session 发送 post 请求，cookie 保存在其中
# 在使用 session 进行请求登陆之后才能访问的地址
session.post(post_url, data=post_data, proxies=proxy, headers=headers)

```

当添加了用户名和密码，成功访问人人网后，接下来，继续访问通过关键词搜索的用户列表页（每页十个用户）。

```

url =
'http://browse.renren.com/sAjax.do?ref_search=&q=&p=%5B%7B%22t%22%3A%22univ%22%2C%22'nam
e%22%3A%22{0}%22%2C%22id%22%3A%22{2}%22%2C%22year%22%3A%22{1}%22%7D%5D&s=0&u={3}&
act=search&offset={4}&sort=0'.format('清华大学', year_list[m], school_list[n], rand_user, count)
try:
    r = session.get(url, timeout=5, proxies=proxy, headers=headers)
except Exception as e:
    print(e)
    continue

```

根据源码生成出信息，通过正则表达式的方式查找出该页十个用户的 id 信息。

```
x = re.findall('<strong>.*id=(\d+)', str(r.text))
```

针对每个获得的 ID，进行去重操作，即判断该 ID 是否存在于 all\_dict 中，如果存在则输出 exists 字样，如果不存在则写出到文件中。

```
for i in range(len(x)):
#去重操作
    if all_dict.__contains__(x[i]):
        print(f"exists:" + x[i])
    else:
        f.write(x[i] + '\n')
        f_all.write(x[i] + '\n')
        all_dict[x[i]] = "1"
        print(f"add:" + x[i])
```

生成 ID 文本信息如下：



Ist0823 - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

539999273  
499146073  
431085809  
428561563  
426833787  
335996030  
324665873  
298978720  
294597883  
510105649  
732529337  
702252049  
357518322  
737148859  
382225849  
495650924

## 2.2 根据 ID 爬取用户信息

### 2.2.1 网页解析流程

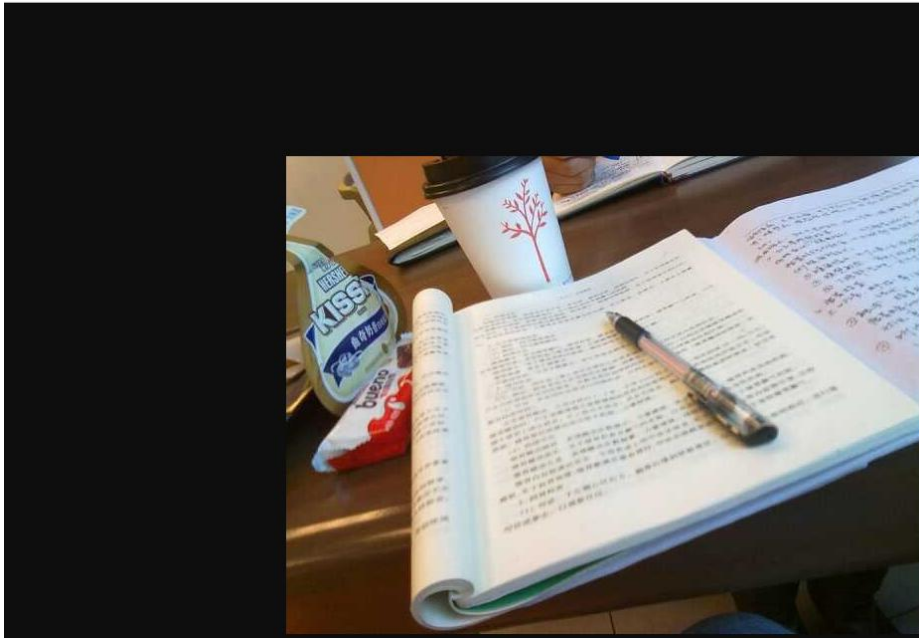
直接通过网站

<http://photo.renren.com/photo/382225849/albumlist/v7?offset=0&limit=40#>进入 ID 为 382225849 用户的相册页





fmn.rimg.com/fmn065//20111219//1345//p\_large\_UwwwA\_34e80001be711261.jpg



### 2.2.2 详细代码解析

根据 ID 信息，以及给出的相册链接，直接访问该用户相册主页：

```
album_url = 'http://photo.renren.com/photo/{0}/albumlist/v7?offset=0&limit=40#'.format(id)
try:
    r = session.get(album_url, timeout=2, proxies=proxy, headers=headers)
    Idpath = savepath+Id+"/"
    if not os.path.exists(Idpath):
        os.makedirs(Idpath)
    else:
        continue
except Exception as e:
    print(e)
    continue
```

根据源码信息直接查找出该用户的相册 ID

```
try:
    x = re.findall("albumId": "(.*?)" , str(r.text))
except Exception as e:
    print(e)
    continue
num = 0
```

结合用户 ID 和相册 ID 直接通过链接直接查找出该相册页面：

```
pic_url = r'http://photo.renren.com/photo/{0}/album-{1}/v7'.format(id,x[i])
#print(pic_url)
try:
    r2 = session.get(pic_url, timeout=2, proxies=proxy, headers=headers)
except Exception as e:
```

```
print(e)
continue
```

访问该相册，并且找出该相册下所有图片的链接，将其所有的图片 URL 存入列表中：

```
try:
    y = re.findall("url": "(.*?)", str(r2.text))
except Exception as e:
    print(e)
    continue
for j in range(len(y)):
    Url = '/'.join(y[j].split('\n'))
    urlist.append(Url)
    num += 1
```

为保证相片质量和爬取速度，针对用户相片数小于 5 的不进行爬取，针对相片数大于 200 的只爬取其前 200 张。

```
if num > 5:
    if num > 200:
        urlist = urlist[:200]
        capture_pic(urlist, lddpath, ld)
    else:
        os.rmdir(ldpath)
```

根据图片链接下载图片并保存

```
def capture_pic(urlist, lddpath, ld):
    count = 0
    for url in urlist:
        picpath = lddpath + ld + '_' + str(count) + '.jpg'
        with open(picpath, "wb") as f:
            try:
                picture = requests.get(url, proxies=proxy, timeout=3, headers=headers)
                f.write(picture.content)
                print("%s 下载成功" % picpath)
                count += 1
            except Exception as e:
                print(e)
                continue
```

## 三、代码运行和示例

### 3.1 代码运行

代码所在桌面：远程连接 10.208.121.22, 堡垒机，在堡垒机输入远程桌面 IP：  
192.168.11.130, username: admin, password: 123456

代码所在位置：C:/Users/lbh/所有爬虫备份/人人网爬虫项目/

运行前请仔细阅读目录下 readme.txt

#### 3.1.1 爬 ID 代码运行方式

双击 run\_crawl\_id.bat 即可。

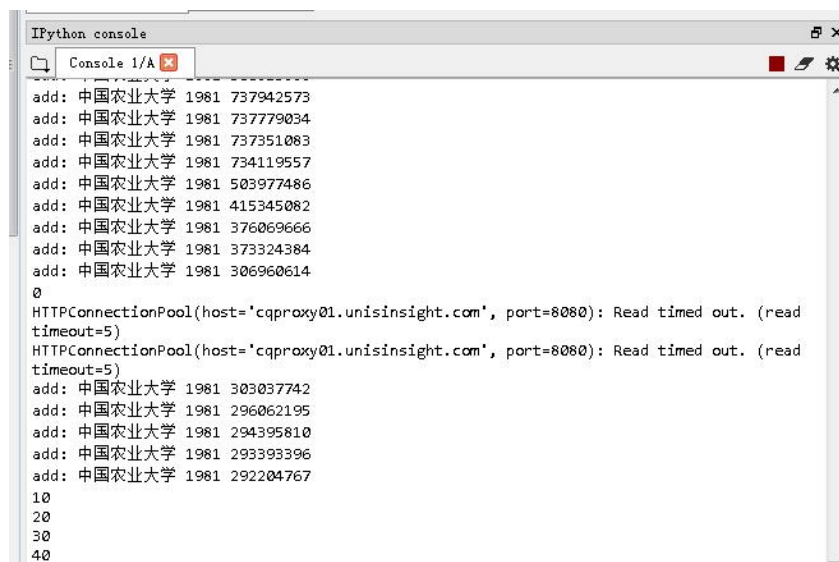
#### 3.1.2 爬图片代码运行方式

右键 run\_crawl\_pic.bat 选择编辑，在弹出窗口里修改图片保存路径和 ID 列表文件路径，  
保存后双击这个 bat 文件即可运行。yu

### 3.2 示例

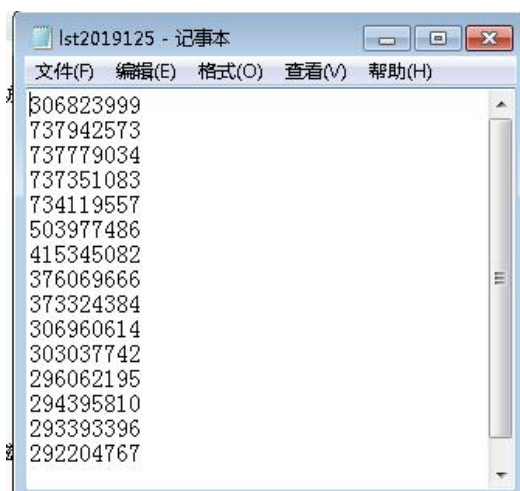
#### 3.2.1 爬 ID 代码示例

爬 ID 代码输出样式如下图所示，其中 add 代表在新添加 ID。



```
IPython console
Console 1/A
add: 中国农业大学 1981 737942573
add: 中国农业大学 1981 737779034
add: 中国农业大学 1981 737351083
add: 中国农业大学 1981 734119557
add: 中国农业大学 1981 503977486
add: 中国农业大学 1981 415345082
add: 中国农业大学 1981 376069666
add: 中国农业大学 1981 373324384
add: 中国农业大学 1981 306960614
0
HTTPConnectionPool(host='cproxy01.unisinsight.com', port=8080): Read timed out. (read
timeout=5)
HTTPConnectionPool(host='cproxy01.unisinsight.com', port=8080): Read timed out. (read
timeout=5)
add: 中国农业大学 1981 303037742
add: 中国农业大学 1981 296062195
add: 中国农业大学 1981 294395810
add: 中国农业大学 1981 293393396
add: 中国农业大学 1981 292204767
10
20
30
40
```

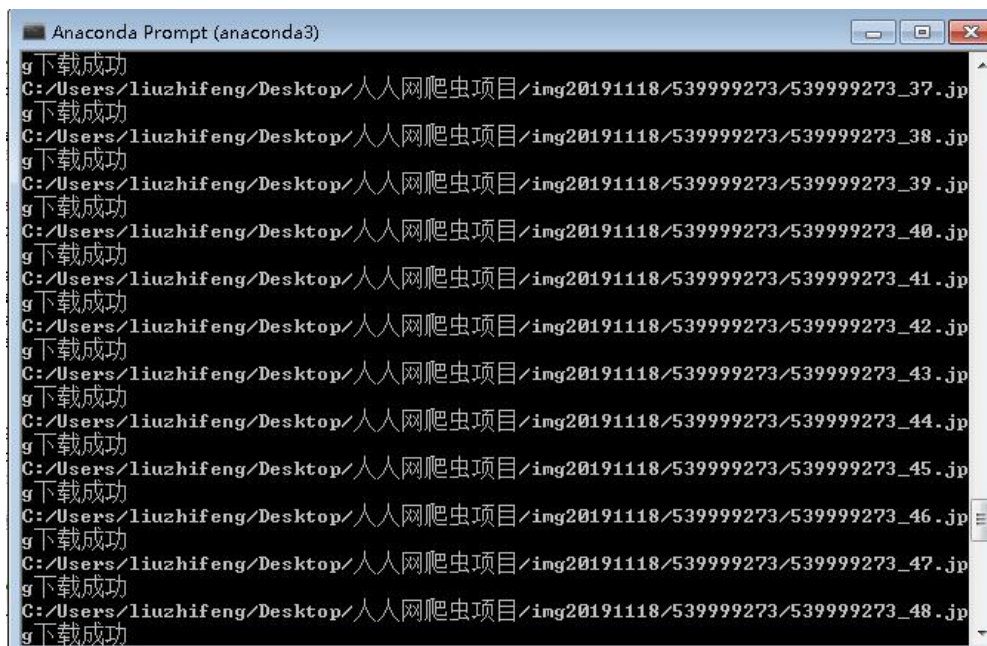
ID 最后保存为下图所示：



```
lst2019125 - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
306823999
737942573
737779034
737351083
734119557
503977486
415345082
376069666
373324384
306960614
303037742
296062195
294395810
293393396
292204767
```

#### 3.2.2 爬图片代码示例

爬 ID 代码输出样式如下图所示，其输出下载路径并提示下载成功。



图片保存为下图所示，其上级文件夹路径为该用户的 ID 号，图片保存格式为 ID 号+第几张图的编号。

