

**《Python语言程序设计》**

**课程设计**

**题 目：** 利用百度AI人脸识别抓取知乎高颜值图片

**学生姓名：** 林怡

**学 号：** 1120170725

**院 （系）：** 信息工程学院

**专业班级：** 软件工程

2018年 6月 12日

# 系统设计

1. **抓取数据**

知乎话题【[美女](https://www.zhihu.com/topic/19552207/newest)】下所有问题中回答所出现的图片。

1. **抓取工具**

Python 3，并使用第三方库 Requests、lxml、AipFace。

1. **人脸检测**

[AipFace](http://ai.baidu.com/docs#/Face-Detect/top)，由百度云 AI 开放平台提供，是一个可以进行人脸检测的 Python SDK。可以直接通过 HTTP 访问，免费使用。

1. **过滤条件**
2. 过滤所有未出现人脸图片（比如风景图、未露脸身材照等）
3. 过滤所有非女性（AIPFace gender不等于female；存在 AipFace 性别识别不准的情况）
4. 过滤所有非真实人物，比如动漫人物 （AipFace Human 置信度小于 0.6）
5. 过滤所有颜值评分较低图片（AipFace beauty 属性小于 45；AipFace 评分无任何客观性，仅供参考）
6. **算法逻辑**
7. 通过 Requests 发起 HTTP 请求，获取知乎话题【[美女](https://www.zhihu.com/topic/19552207/newest)】下的部分讨论列表
8. 通过 lxml 解析抓取到的每个讨论中 HTML，获取其中所有的 img 标签相应的 src 属性
9. 通过 Requests 发起 HTTP 请求，下载 src 属性指向图片（不考虑动图）
10. 通过 AipFace 请求对图片进行人脸检测
11. 判断是否检测到人脸，并过滤不合格图片
12. 将过滤后的图片保存到本地文件系统，
13. 返回第一步，继续
14. **结果保存**

保存在当前位置下image文件夹中，每张图片命名为 【颜值 + 作者 + 问题名 + 序号】。

# 使用说明

1. **运行准备**

安装 [requests](http://docs.python-requests.org/zh_CN/latest/user/quickstart.html)、[lxml](http://lxml.de/installation.html)、[baidu-aip](http://ai.baidu.com/docs#/Face-Python-SDK/top) 库，都可以通过 pip 安装。

1. **百度AI**

申请百度云[人脸检测](https://cloud.baidu.com/product/face)服务，要求登录百度账号。

创建新应用，将申请后的AppID、 ApiKey、 SecretKey 填写到代码中。

<http://ai.baidu.com/docs#/Face-Python-SDK/top>

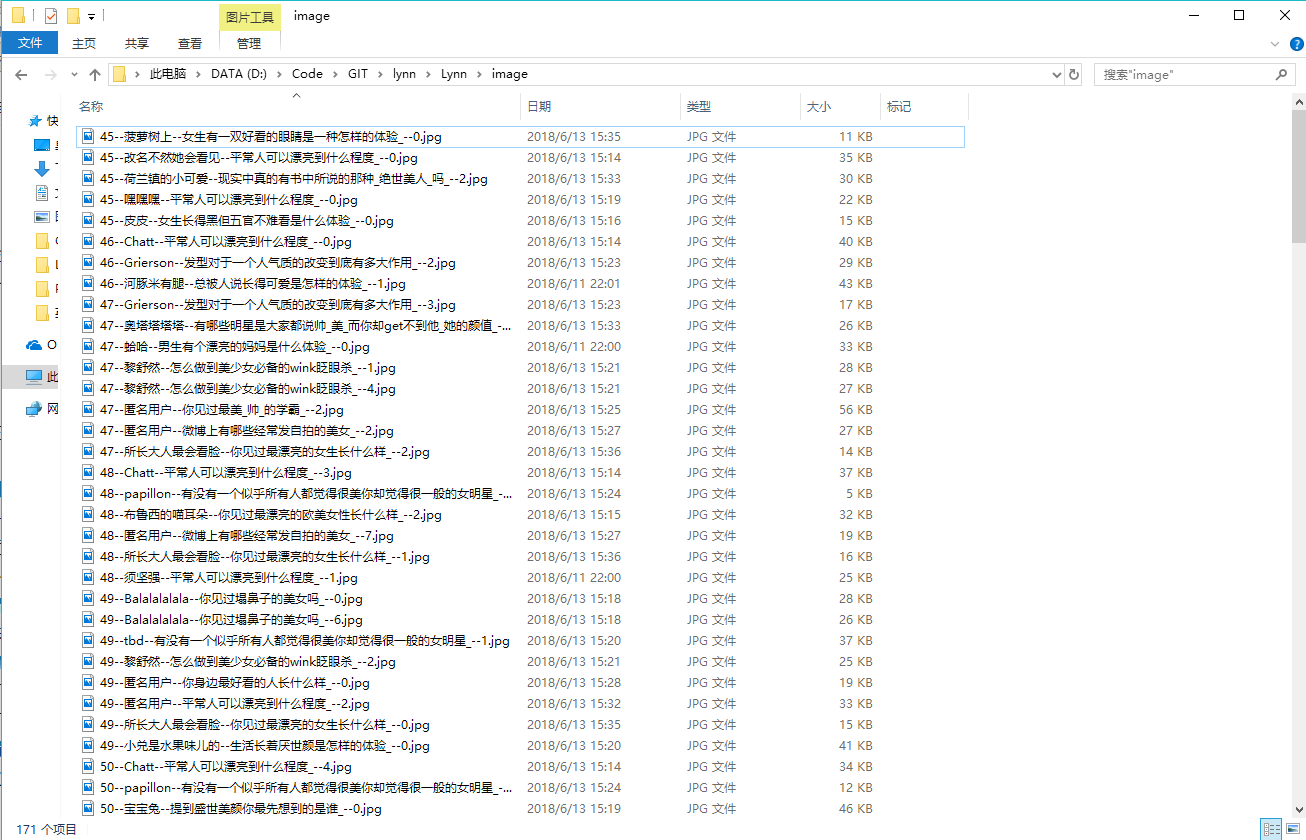
1. **请求失败**

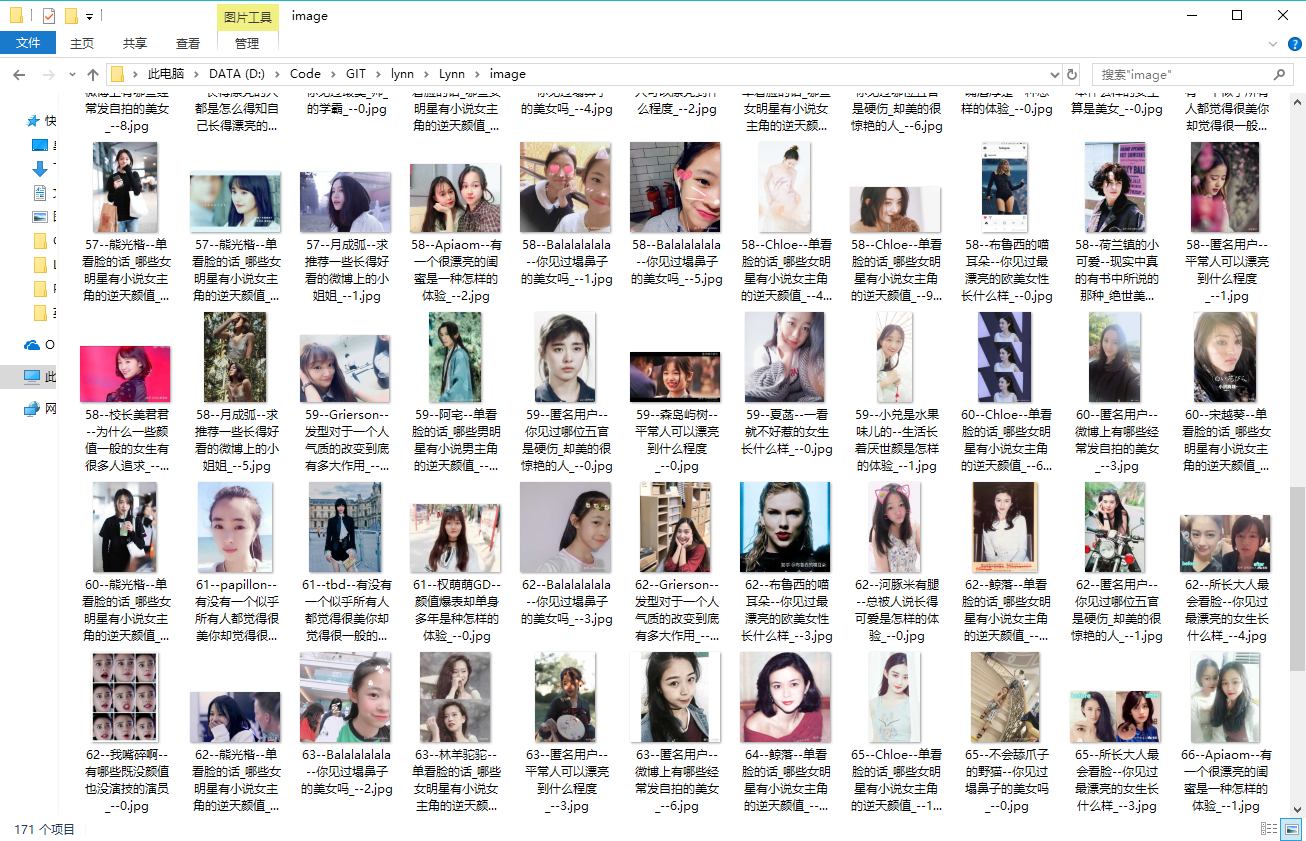
若请求知乎失败，返回如下。需填写 AUTHORIZATION，找一个知乎链接点进去，打开开发者工具，查看 HTTP 请求 header。



# 系统测试

运行成功，爬取部分图片如下。





# 代码

<https://github.com/Lynn0101/Lynn.git>