

Summary of Chinese Celebrities Data Collecting & Cleaning

Na Zhang

Outline

1. Create Name List
2. Download Images

1. Create Name List

● Goal

✓ Find Chinese Celebrities

✓ Contain persons in

Mainland China

Hong Kong

Taiwan

7/30/2017

Summary of Data Collecting

3

Occupation includes

- Singers
- Actors/Actresses
- Models
- Host/Hostess
- Directors
- Athletes
- Dancers
- Entertainers
- Musician
-

7/30/2017

Summary of Data Collecting

4

How to find?

● Mainly Chinese Websites

- Use some famous Search Engines, like Google, Baidu.
- Baidu, largest Chinese search engine
- Search by keywords



Search Google or type URL



百度一下

7/30/2017

Summary of Data Collecting

5

Name List Sources

- Tencent entertainment 腾讯娱乐资料库
- 6Kstar 乐酷明星网
- Ttpaihang 天天排行榜
- StarRank 百度百科人气榜
- n63.com 明星写真馆
- Zhidao 百度知道
- TopBaidu 百度搜索风云榜
- Other sources...

7/30/2017

Summary of Data Collecting

6

(1) 腾讯娱乐资料库 (Tencent Entertainment)

○ http://ent.qq.com/c/dalu_star.shtml

The screenshot shows the Tencent Entertainment database interface. At the top, there's a navigation bar with '明星' (Stars) selected. Below it is a search bar and a '热门搜索' (Hot Search) section. The main content area is divided into two columns. The left column has '按地区检索' (Search by Region) with options like '全部明星', '内地明星', '港台明星', '亚洲明星', and '欧美明星'. Below that is '趣味查询' (Fun Search) with '同生日查询' (Search by Birthday) and '同星座查询' (Search by Zodiac). The right column shows '按地区检索: 中国大陆' (Search by Region: Mainland China) with a grid of letters A-Z and '0-9'. Below the grid, there are sections for '0-9', 'A', and 'B', each containing a list of stars and a '更多' (More) link.

7/30/2017

Summary of Data Collecting

7

(2) 乐酷明星网 (6Kstar)

○ <http://star.6k.com/diqu/zhongguo/>



The screenshot shows the 6Kstar website interface. At the top, there's a navigation bar with '明星档案' (Star Profiles) selected. Below it is a search bar and a '当前位置' (Current Location) section. The main content area is divided into two columns. The left column has '按职业' (Search by Profession) and '按地区' (Search by Region) with options like '中国内地', '港澳台', '亚洲', '欧美', and '其它'. Below that is '按字母' (Search by Letter) with a grid of letters A-Z. The right column shows a list of stars under the heading '中国内地明星' (Mainland China Stars). The list is organized by letter 'A' and contains names like 安又琪, 艾敬, 阿穆隆, A-OK乐队, 安战军, 安雯, 安琪, 安琥, 艾琳, 阿朵, 阿丘, 阿炳, 敖杨, 敖特根, 敖磊, 安亚平, 安雅萍, 安旭, 安霖, 安荣生, 安启虎, 安建, 安建.

7/30/2017

Summary of Data Collecting

8

(3) 天天排行榜 (ttpaihang)

○ <http://www.ttpaihang.com/subjects/starworld/>

中国最受欢迎的男明星新人榜

榜单统计时间: 2017-03-03 05:50:19

名次	得票占比	得票数	人气占比	人气值
TOP 1、邱泽	16.40%	16155票		182302
TOP 2、罗晋				
TOP 3、鹿晗				
TOP 4、炎亚纶				
TOP 5、吴亦凡 KRIS				
TOP 6、张艺兴				
TOP 7、陈赫				
TOP 8、黄子韬				
TOP 9、边伯贤				
TOP 10、张翰				
TOP 11、汪东城				
TOP 12、魏晨				
TOP 13、王雨				
TOP 14、张杰				
TOP 15、陈翔				
TOP 16、熊亦儒				
TOP 17、唐禹哲				
TOP 18、韩庚				
TOP 19、李佳航				
TOP 20、靳东				

7/30/2017

Summary of Data Collecting

9

(4) 百度百科人气榜 (star rank)

○ <http://baike.baidu.com/starrank>

本周榜 | 上周榜 | 赢取鲜花攻略

排名	明星	鲜花数	TOP粉丝
1	王俊凯	1146165	如宝今年五岁WR
2	易烊千玺	442513	yy1128lian
3	金秀贤	190948	mary200366
4	钟汉良	100079	零距离Sunflowe
5	马天宇	96708	吃饭饭睡睡觉耶
6	黄子韬	77136	Devil、心痛
7	曾艳芬	67495	李胜贤妻良母
8	郑智薰	60550	panwei53
9	金钟国	55383	savmc622

7/30/2017

Summary of Data Collecting

10

(5) 明星写真馆 (n63)

○ <http://www.n63.com/>



7/30/2017

Summary of Data Collecting

11

(6) 百度知道 (zhidao)

<https://zhidao.baidu.com/question/369594442.html>



明星名字大全 50

我需要大陆和港台的所有男女明星的名字, 这个任务是the360buy交给我的。
回答格式如下: 【越全越好 再加分!!!】

最佳答案

推荐于2016-08-05 17:09:50

大陆男明星

大陆女明星

港台男明星

港台女明星

邢佳栋 李学庆 高昊 潘粤明 李宇春 张靓颖 周笔畅 何洁 文龙 张殿菲 邓超 张杰 杨坤 沙 汤唯 张筱雨 韩雪 孙菲菲 张鼎 杨子 邓安奇 赵鸿飞 马可 黄 虎 印小天 于和伟 田亮 夏雨 江一燕 厉娜 许飞 胡灵 郝菲 弘 朱雨辰 丁志诚 黄征 张子 冰 魏晨 郭敬明 何晟铭 巫迪 谭维维 魏佳庆 张亚飞 李旭丹 李炜 罗中旭 张远 李立 释小 勇 张国强 玉米提 周觅 张丹 王蓉 汤加丽 汤芳 牛萌萌 范 谢和弦 陈道明 柳云龙 汪峰 纪敏佳 黄雅莉 叶一茜 马苏 张翰 杨洋 宋晓波 解小东 龚 为 柏栩栩 蒲巴甲 凌潇肃 李 姚笛 朱妍 真颖 陈西贝 冯家 钱泳辰 撒贝宁 徐峥 谭杰希 陶虹 徐静蕾 黄奕 董洁 巩俐

吴卓羲 游鸿明 胡宇崧 张震岳 陈国坤 张信哲 范逸臣 王绍伟 广仲 林文龙 赵又廷 刘德华 周 杰 狄龙 郭富城 光良 黄浩然 彭 锋 王喜 黄贯中 江华 贺一航 郑 凯 吴镇宇 哈狗帮 吴尊 张国荣 叶璇 唐宁 曾之乔 安以轩 杨丞琳 侯佩岑 同恩 陈松伶 吴奇隆 金城武 李圣杰 陈建州 秋生 罗嘉良 欧弟 马国明 范植 青云 黄子华 丁子高 童安格 王 陈晓东 潘玮柏 黄日华 张学友 古天乐 甄子丹 梁朝伟 房祖名 陈司翰 朱孝天 王子 敖犬 黄维

蔡依林 张韶涵 王心凌 徐若瑄 林志玲 王菲 S.H.E Twins 徐熙媛 桂纶镁 林依晨 陈乔恩 梁静茹 蔡诗芸 范玮琪 廖碧儿 张柏芝 李嘉欣 容祖儿 李玟 贾静雯 MaggieQ 林心如 朱茵 叶璇 唐宁 曾之乔 安以轩 杨丞琳 侯佩岑 同恩 陈松伶 文颂娴 梁凯蒂 林韦君 陈思璇 曹敏莉 乐基儿 郑雪儿 余诗曼 郑秀文 萧蔷 温碧霞 刘嘉玲 刘玉玲 林熙蕾 李若彤 张曼玉 关之琳 陈慧琳 萧淑慎 蔡少芬 萧亚轩 田丽 杨采妮 李丽珍

7/30/2017

Summary of Data Collecting

12

(7) 百度搜索风云榜 (top baidu)

http://top.baidu.com/buzz?b=258&c=9&fr=topbuzz_b454_c9

The screenshot shows the Baidu Buzz website interface. On the left, there is a navigation menu with categories like '娱乐名人' (Entertainment Celebrities), '女演员' (Actresses), '男演员' (Actors), '女歌手' (Female Singers), '男歌手' (Male Singers), '主持人' (Hosts), '体坛人物' (Sports Figures), '美女' (Beautiful Women), '帅哥' (Handsome Guys), '选秀歌手' (Talent Show Singers), and '欧美明星' (Western Stars). The '主持人' (Hosts) category is selected and highlighted in blue. The main content area displays a list of hosts with their names, photos, and search popularity scores. The top three entries are:

排名	关键词	相关链接	搜索指数
1	张大大	简介 贴吧 视频	67242
2	蕾娜	简介 贴吧 视频	4357
3	董贝宁	简介 贴吧 视频	3015

7/30/2017

Summary of Data Collecting

13

(8) Other sources

athletes: http://blog.sina.com.cn/s/blog_5f05e98f0102x6yt.html

<http://www.xuexila.com/news/1243489.htm>

历届奥运会 (1984——2016) 中国金牌榜及冠军运动员统计 (2016-08-22 0)

标签: 体育 分类: 体育及其他资料

一、1984年第23届奥运会 (2016-08-22 0)

顺序	时间	2016中国运动员名单
第01枚	1984年07月	2016中国运动员名单
第02枚	1984年07月	项目:游泳
第03枚	1984年07月	领队:王路生
第04枚	1984年07月	副领队:许琦
第05枚	1984年07月	教练:何新中,叶瑾,张亚东,李志忠,么正杰,李雪刚,崔登荣,金伟,徐国义,朱志根,韩林岩,刘海涛,吕森,金浩
第06枚	1984年08月	医生:张乐伟
第07枚	1984年08月	科研人员:贾蕾,巴震
第08枚	1984年08月	翻译:陆一帆
第09枚	1984年08月	管理人员:程浩,王强
第10枚	1984年08月	女运动员:

中国体育人物

展开



宁泽涛

泳坛新男神
及亚洲飞鱼

傅园慧

奥运摘铜的
洪荒少女

张继科

仅用15个月
成就大满贯



谌龙

里约夺冠实
现大满贯



王一梅

中国女排最
强主攻之一



鲍春来

前中国羽毛
球运动员

中国运动员

展开



惠若琪

新生代球迷
追捧的偶像

谢杏芳

已退役羽毛
球运动员

郎平

排球运动员
教练员

7/30/2017

Summary of Data Collecting

14

How to get the list?

- Copy names from webpage to document (.txt/.doc/.xls/...)
 - Format the list
 - Manually
 - Automatically: write code
- Check name list
 - remove band name, repeated names, non Chinese...
 - Write code + manually

7/30/2017

Summary of Data Collecting

15

- Problems?
 - More than one persons have same name
 - One person has more than one names
 - Name has incorrect characters
 - Person does not exist
 - There are non-Chinese persons
 - Is not a name (bands,...)

7/30/2017

Summary of Data Collecting

16

About Name List

- Assign each name an ID number
- 10,109 names in total



7/30/2017

Summary of Data Collecting

17

2. Download Images

- Method
 - Automatically download images from Internet by running python script
- Procedures
 1. Analyze pattern of websites
 2. Preprocess name list (It depends)
 3. Write python script
 4. Read name list and download images
 5. Save images

7/30/2017

Summary of Data Collecting

18

Image sources

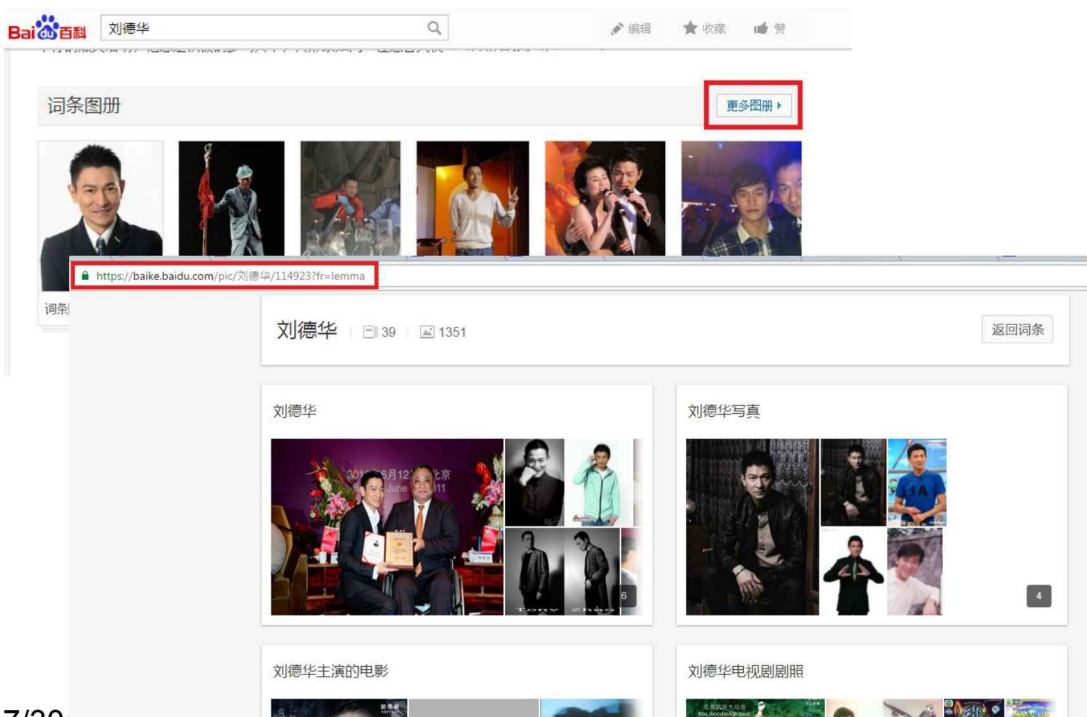
- Baidu Baike 百度百科
- Baidu Tieba 百度贴吧
- Mingxing 明星图库
- n63 明星写真馆
- Tupianzj 图片之家
- Google Images 谷歌图片

7/30/2017

Summary of Data Collecting

19

(1) 百度百科 <https://baike.baidu.com>



7/30/2017

Summary of Data Collecting

20

- Pattern of URL

<http://baike.baidu.com/item/刘德华>

<https://baike.baidu.com/pic/刘德华/114923?fr=lemma>

```
for i in range(0, len(text)-1):
    print '%d/%d\n' % (i+1, len(text))
    subject_no = text[i].split()[0].strip()
    subject_name = text[i].split()[1].strip()
    save_dir = 'F:/zn Chinese Celebrities/bd images/' + str(subject_no) + subject_name.decode('utf-8')
    URL = 'http://baike.baidu.com/item/' + subject_name

    try:
        page = requests.get(URL)
        m = re.search(urllib.quote(subject_name) + '/[0-9]+\?fr=lemma', page.content)
        number = m.group(0).split('?')[0].split('/')[1]
        updated_URL = 'http://baike.baidu.com/pic/' + urllib.quote(subject_name) + '/' + number + '?fr=lemma'
        if not os.path.isdir(save_dir):
            os.makedirs(save_dir)
        call(["image-scraper", "-s", save_dir, updated_URL])
    except Exception:
        print subject_name + ' skipped.\n'
        continue
```

7/30/2017

Summary of Data Collecting

21

- Install dependency python packages before running the script

- "imagescraper"

- Download and install

- <https://github.com/sananth12/ImageScrapper/>

- go to your chosen folder and then apply the following command:

- ```
python setup.py install
```

- Other dependency packages

- setproctitle / future/ requests/ selenium/ lxml ...

7/30/2017

Summary of Data Collecting

22

## ● Problems

When encounter the following situations, it fails.

❖ One name shows more than one persons

Baidu Baike search results for '宋佳'. The search results show four entries: '中国大陆80后女演员', '中国大陆60后女演员', '中国科学院地理科学与资源研究所助理研究员', and '青年女作家'. A blue callout bubble points to the list of links below, stating 'not follow the pattern'.

① [baike.baidu.com/item/宋佳/17175785#viewPageContent](http://baike.baidu.com/item/宋佳/17175785#viewPageContent)

① [baike.baidu.com/item/宋佳/5812#viewPageContent](http://baike.baidu.com/item/宋佳/5812#viewPageContent)

① [baike.baidu.com/item/宋佳/10484754#viewPageContent](http://baike.baidu.com/item/宋佳/10484754#viewPageContent)

① [baike.baidu.com/item/宋佳/17648669#viewPageContent](http://baike.baidu.com/item/宋佳/17648669#viewPageContent)

7/30/2017

Summary of Data Collecting

23

## ❖ No album exist

中文名: 张经纬  
职业: 电影编剧、导演  
国籍: 中国  
主要成就: 剧本《上帝的苹果》获得化装奖  
出生地: 香港  
代表作品: 《天地孩子: 交错的战事》  
籍贯: 广东深圳南山区

目录

1. 剧本
2. 导演作品

**剧本**

第一部剧情长片剧本《上帝的苹果》获得「2001年中港台电影编剧创作大赛」优胜奖。

第二部剧本《天水围》获得2005年香港亚洲电影投资会的最佳剧本奖。此剧本由许鞍华拍成了两部电影《天水围的日与夜》和《天水围的夜与雾》。

**导演作品**

2003年导演的处女作《天地孩子: 交错的战事》参加了香港电台的短片大赛。

2007年入选CINEK第一屆主展「征策」,完成纪录片《歌舞升平》,并以此片入选第18届国际电影节最高单元、韩国首尔独立电影节非竞赛单元、第20届法国飞越国际电影节等。

2002年—2008年执导纪录长片《音乐人生》,2009年7月15日该片在香港开始了特别放映,该片荣获第46届台湾电影金马奖 最佳纪录片、最佳音效、最佳剪辑奖。

2010年4月16日,张经纬凭《音乐人生》在第29届香港电影金像奖中获颁最佳导演奖。

张经纬: 编剧、导演、制片人、人物

## ❖ There is no such person in this website

7/30/2017

Summary of Data Collecting

24



- 6,979 persons / 184,580 images
- Image quality: big noise
  - Many persons in one subjects
  - Complex background
  - Blur
  - Illumination
  - Big pose angle
  - Irrelative images



Summary of Data Collecting

## (2) 百度贴吧: <http://tieba.baidu.com/>



- Pattern of URL

- <http://tieba.baidu.com/f?kw=张学友&ie=utf-8&tab=album>

- Use similar python script and environment

```
for i in range(0, len(text)-1):
 print '%d/%d\n' % (i+1, len(text))
 subject_no = text[i].split()[0].strip()
 subject_name = text[i].split()[1].strip()
 save_dir = 'F:/zn Chinese Celebrities/bd images/' + str(subject_no)
 URL = 'http://tieba.baidu.com/f?kw='+ urllib.quote(subject_name) + '&ie=utf-8&tab=album'
 try:
 if not os.path.isdir(save_dir):
 os.makedirs(save_dir)
 call(["image-scraper", "-s", save_dir, URL])
 except Exception:
 print subject_name + ' skipped.\n'
 continue
```

7/30/2017

Summary of Data Collecting

27

- Problems

When encounter the following situations, it fails.

- ❖ One name shows more than one persons
- ❖ No album exist
- ❖ The person does not exist in this website

7/30/2017

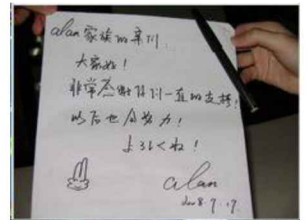
Summary of Data Collecting

28

● 5,330 persons / 70,233 images

● Image quality: big noise

- Many persons in one subject
- Complex background
- Cartoon images
- Wrong identity
- Blur
- Illumination
- Big pose angle
- Irrelative images



7/30/2017

Summary of Data Collecting

29

### (3) 明星图库: <http://www.mingxing.com/ziliao/>

www.mingxing.com/ziliao/

M明星 娱乐 mingxing.com 明星网 | 资料 图库 剧照 星闻 | 电影 电视剧 独家策划

明星资料 内地明星 港台明星 日韩明星 欧美明星

内地明星

www.mingxing.com/gangtai/wuqianyu/tupian.html

M明星网 写真 壁纸 剧照 现场 八卦新闻 明星库

吴千语 主页 图片 星闻 个人资料

2010年, 拍摄Tempo纸巾电视广告, 获颁百鸣奖成为天马电影出品有限公司合约艺人。

www.mingxing.com/neidi/chenlong/tupian.html

M明星 娱乐 mingxing.com 明星网 | 资料 图库 剧照 星闻 | 电影 电视剧 独家策划

陈龙 演员/歌手

出生: 中国 星座: 处女座 身高: 186cm 体重: 75kg

个人主页 成名史 详细资料 最新消息 写真图片 影视剧照 电影作品 电视剧

陈龙图片大全/写真图片/生活照/活动照片

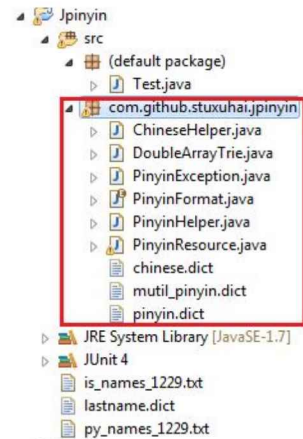
7/30/2017

Summary of Data Collecting

30

- Pattern of URL
  - Mainland China
    - <http://www.mingxing.com/neidi/chenlong/tupian.html>
  - Hong Kong & Taiwan
    - <http://www.mingxing.com/gangtai/lizongsheng/tupian.html>

- Name in the URL: pinyin (拼音)
- Need to convert Chinese Characters into Pinyin.
- Method: Java + Eclipse + Jpinyin
  - Jpinyin: Java Class Library



 [pinyin] pinyin\_names\_1229\_mingxing.txt

7/30/2017

Summary of Data Collecting

31

- Use similar python script and environment

✓ “imagescraper”

```

for i in range(0, len(text)):
 print '%d/%d\n' % (i+1, len(text))
 subject_no = text[i].split()[0].strip()
 subject_name = text[i].split()[1].strip()
 save_dir = 'F:/zn Chinese Celebrities/all images/' + str(subject_no)
 URL = 'http://www.mingxing.com/neidi/' + urllib.quote(subject_name) + '/tupian.html'
 #URL = 'http://www.mingxing.com/gangtai/' + urllib.quote(subject_name) + '/tupian.html'
 try:
 if not os.path.isdir(save_dir):
 os.makedirs(save_dir)
 call(["image-scraper", "-s", save_dir, URL])
 except Exception:
 print subject_name + ' skipped.\n'
 continue

```

7/30/2017

Summary of Data Collecting

32



## ● Problems

- ❖ names in URL are not always in strict conformity with pinyin



7/30/2017

Summary of Data Collecting

33

## ❖ pinyin names are not correct

- Polyphone
- Some words as last name have different pinyin

曾 : zeng (姓)  
ceng (普通)



7/30/2017

Summary of Data Collecting

34

## ❖ For bands, fail to download by individual names

- tfboys

www.mingxing.com/neid/tfboys\_eupian.html

M明星 娱乐 mingxing.com 明星网 | 资料 图库 剧照 星闻 | 电影 电视剧 独家策划

**tfboys** 演员/歌手  
出生: 中国 星座: 处女座 身高: 186cm 体重: 75kg

个人主页 成名史 详细资料 最新消息 写真图片 影视剧

tfboys图片大全/写真图片/生活照/活动照片



TFBOYS王源晒踏青照 TFBOYS专辑宣传写真 TFBOYS王源生日会白 TFBOYS身穿西服帅气

## ❖ Name does not exist in this website

7/30/2017

Summary of Data Collecting

35

## ❖ Different words has same pinyin

- 杨紫 (actress) yangzi
- 杨梓 (host) yangzi



7/30/2017

Summary of Data Collecting

36

● 2,931 persons / 30,324 images

● Image quality: a few noise

- Wrong identity
- Big pose angle
- Irrelative images
- Blur
- Illumination



7/30/2017

Summary of Data Collecting

37

## (4) 明星写真馆: <http://www.n63.com/>

The screenshot displays the n63.com website interface. At the top, there is a navigation bar with the text 'n63.com 明星写真馆: 首页 /'. Below this, a horizontal menu lists various celebrities: 林心如, 李丽珍, 张韶涵, 汤加丽, 刘亦菲, 贾静雯, 舒淇, 张柏芝, 赵薇, 杨颖. The main content area shows two gallery pages. The first gallery is for '华人男星 陈楚河' (Chinese Male Star Chen Chuhe) with 64 photos and 6 pages. The second gallery is for '华人女星 陈慧娴' (Chinese Female Star Chen Huihuan) with 44 photos and 4 pages. Both galleries include navigation options like '设计图库' (Design Library) and '剧照' (Movie Stills). The browser address bar for the second gallery shows 'www.n63.com/n\_china/chenuixian'.

7/30/2017

Summary of Data Collecting

38

- Pattern of URL

- Female

- [http://www.n63.com/n\\_china/chenhuixian](http://www.n63.com/n_china/chenhuixian)

- Male

- [http://www.n63.com/n\\_chinam/chenchuhe](http://www.n63.com/n_chinam/chenchuhe)

- Also pinyin in URL

- Use pinyin name list

7/30/2017

Summary of Data Collecting

39

- Use similar python script and environment:



“imagescraper”

```
for i in range(0, len(text)):
 print '%d/%d\n' % (i+1, len(text))
 subject_no = text[i].split()[0].strip()
 subject_name = text[i].split()[1].strip()
 save_dir = 'F:/zn Chinese Celebrities/all images/' + str(subject_no)
 URL = 'http://www.n63.com/n_china/' + urllib.quote(subject_name)
 #URL = 'http://www.n63.com/n_chinam/' + urllib.quote(subject_name)

try:
 if not os.path.isdir(save_dir):
 os.makedirs(save_dir)
 call(["image-scraper", "-s", save_dir, URL])
except Exception:
 print subject_name + ' skipped.\n'
 continue
```

7/30/2017

Summary of Data Collecting

40



## ● Problems: similar with(3)

- ❖ names in URL are not always in strict conformity with pinyin



Pinyin: guanzhibin



Pinyin: wujianhao

7/30/2017

Summary of Data Collecting

41

- ❖ pinyin names are not correct

- Polyphone

- ❖ For bands, fail to download by individual names

- Twins



7/30/2017

Summary of Data Collecting

42

● 1,849 persons, 14,112 images

● Image quality: little noise

- Big pose angle
- Blur
- Illumination
- Two pictures appear in one single image

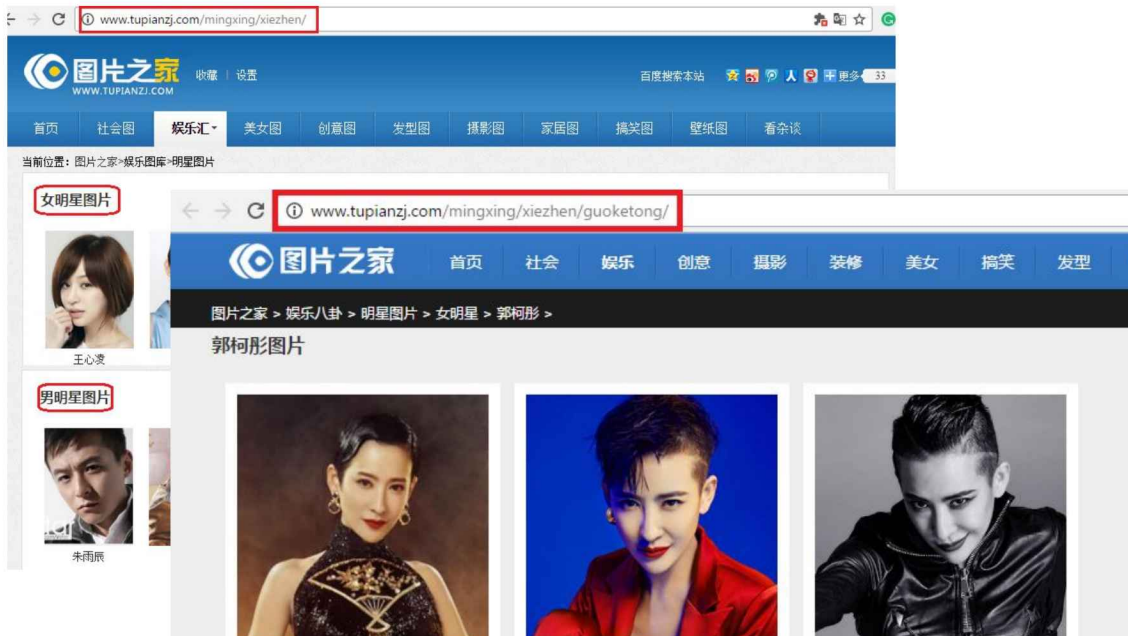


7/30/2017

Summary of Data Collecting

43

(5) 图片之家: <http://www.tupianzj.com/mingxing/xiezhen/>



7/30/2017

Summary of Data Collecting

44

- Pattern of URL
  - <http://www.tupianzj.com/mingxing/xiezheng/guoketong/>
- Also pinyin in URL
- Use pinyin name list

7/30/2017

Summary of Data Collecting

45

- Use similar python script and environment

✓ “imagescraper”

```

for i in range(0, len(text)-1):
 print '%d/%d\n' % (i+1, len(text))
 subject_no = text[i].split()[0].strip()
 subject_name = text[i].split()[1].strip()
 save_dir = 'F:/zn Chinese Celebrities/bd images/' + str(subject_no)
 URL = 'http://www.tupianzj.com/mingxing/xiezheng/' + subject_name + '/'

 try:
 if not os.path.isdir(save_dir):
 os.makedirs(save_dir)
 call(["image-scraper", "-s", save_dir, URL])
 except Exception:
 print subject_name + ' skipped.\n'
 continue

```

7/30/2017

Summary of Data Collecting

46

● Problems: similar with(3,4)

❖ names in URL are not always in strict conformity with pinyin



章泽天  
zhangzetian

7/30/2017

Summary of Data Collecting

47

❖ pinyin names are not correct

- Polyphone

❖ For bands, fail to download by individual names

- Tfboys

❖ There is no such person in this website

❖ Different words has same pinyin

7/30/2017

Summary of Data Collecting

48

● 598 persons / 4,146 images

● Image quality: little noise

Big pose angle

Blur

Illumination

occlusion

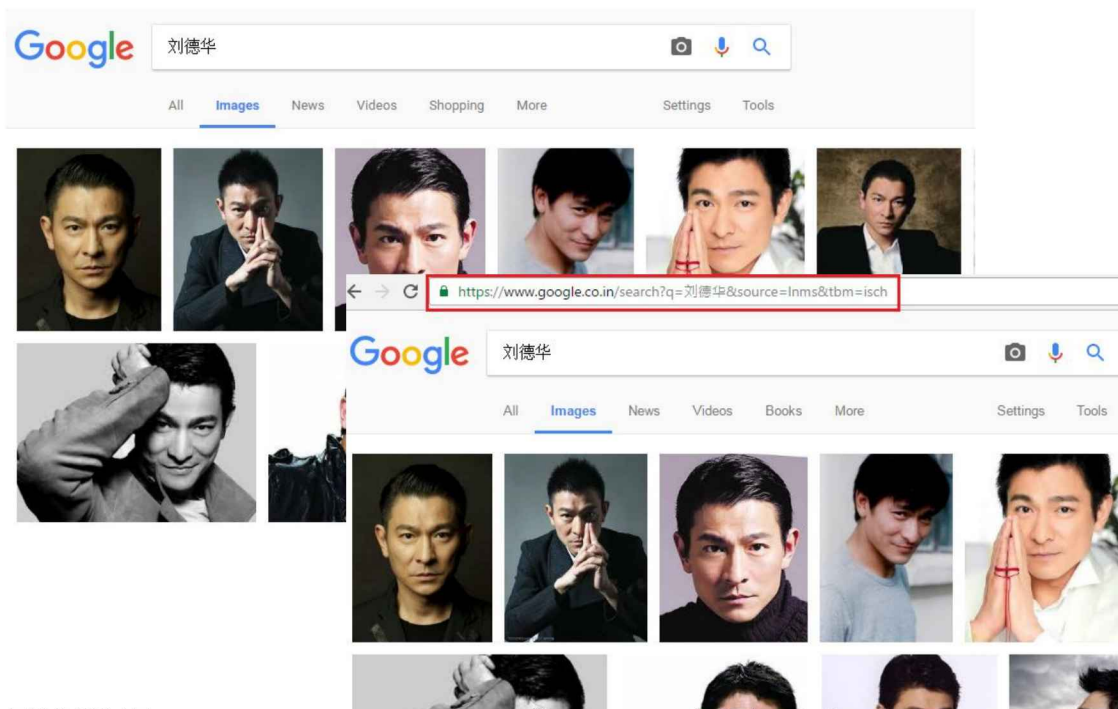


7/30/2017

Summary of Data Collecting

49

(6) Google Images: <https://images.google.com/>



7/30/2017

Summary of Data Collecting

50



- Pattern of URL :
  - <https://www.google.co.in/search?q=刘德华&source=lnms&tbm=isch>
- Use Chinese words in URL
- Not use 'ImageScrapper' package to download
  - Fail to download from Google
- Write new script

7/30/2017

Summary of Data Collecting

51

## ● Download the first 25<sup>th</sup> images

```

browser = webdriver.Firefox()
for i in range(0, len(text) - 1):
 print '%d/%d\n' % (i + 1, len(text))
 subject_no = text[i].split()[0].strip()
 subject_name = text[i].split()[1].strip()
 save_dir = 'F:/zn Chinese Celebrities/all images/' + str(subject_no)
 url = "https://www.google.co.in/search?q="+subject_name+"&source=lnms&tbm=isch"

 browser.get(url)
 header={'User-Agent':"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)"}
 counter = 0
 succounter = 0

 for _ in range(50):
 browser.execute_script("window.scrollTo(0,1000)")

 for x in browser.find_elements_by_xpath("//div[@class='rg_meta']"):
 counter = counter + 1
 if succounter > 25:
 break
 img = json.loads(x.get_attribute('innerHTML'))["ou"]
 imgtype = json.loads(x.get_attribute('innerHTML'))["ity"]
 try:
 req = urllib2.Request(img, headers={'User-Agent': header})
 raw_img = urllib2.urlopen(req).read()
 File = open(os.path.join(save_dir , save_dir + "_" + str(counter) + "." + imgtype), "wb")
 File.write(raw_img)
 File.close()
 succounter = succounter + 1
 except:
 print "can't get img"
 browser.close()

```

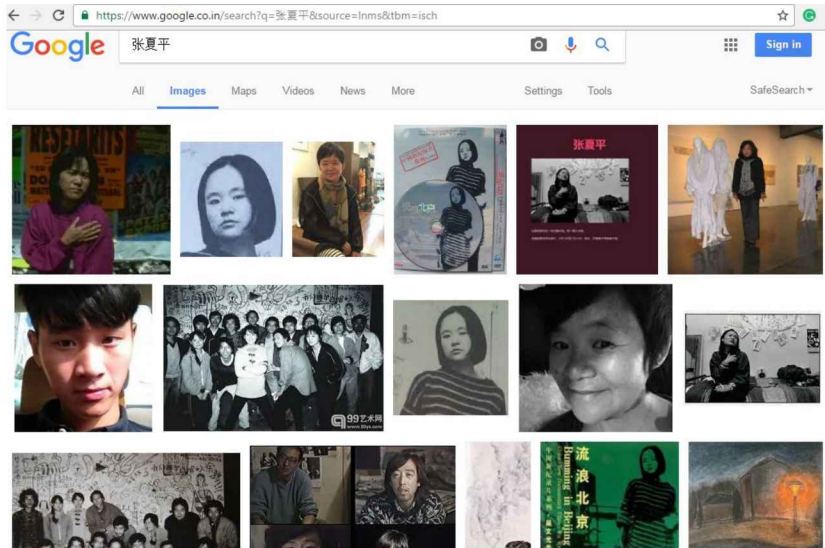
7/30/2017

Summary of Data Collecting

52

- Problems

- ❖ Unfamous names, get wrong & irrelative pictures



7/30/2017

Summary of Data Collecting

53

- 7,191 persons / 223,540 images



- Image quality: big noise in certain subjects

- Big pose angle
- Blur
- Illumination
- Occlusion
- Complex background



Data Collecting

54

# Summary

- Table of each image source

| Image source    | #subjects    | #images        | quality      | Clean? |
|-----------------|--------------|----------------|--------------|--------|
| Baike           | 6,979        | 184,580        | Big noise    | Yes    |
| Tieba           | 5,330        | 70,233         | Big noise    | Yes    |
| Mingxing        | 2,931        | 30,324         | a few noise  | Yes    |
| n63             | 1,849        | 14,112         | Little noise | No     |
| tupianzj        | 598          | 4,146          | Little noise | No     |
| Google          | <b>7,190</b> | <b>234,490</b> | Big noise    | Yes    |
| <b>In Total</b> | <b>9,128</b> | <b>537,885</b> |              |        |

7/30/2017

Summary of Data Collecting

55

# Data Cleaning



- Six subsets from six image sources
  - Baike
  - Tieba
  - Mingxing
  - N63
  - Tupianzj
  - Google

## Dataset Details

| Image source    | #subjects    | #images        | quality      | Clean?   |
|-----------------|--------------|----------------|--------------|----------|
| Baike           | 6,979        | 184,580        | Big noise    | Yes      |
| Tieba           | 5,330        | 70,233         | Big noise    | Yes      |
| Mingxing        | 2,931        | 30,324         | a few noise  | Yes      |
| N63             | 1,849        | 14,112         | Little noise | Yes,easy |
| Tupianzj        | 598          | 4,146          | Little noise | Yes,easy |
| Google          | <b>7,190</b> | <b>234,490</b> | Big noise    | Yes      |
| <b>In Total</b> | <b>9,128</b> | <b>537,885</b> |              |          |

- Noise distributions in them are **not even**
- For Baike and Tieba, each subject has **high noisy** images
- For Google, some subjects are **pretty clear** if the persons are very famous, and some others have **too much noise** if not famous
- Mingxing, N63 and Tupianzj have much **less noise** than the three subsets above

## Method: Clean manually

- All data will be cleaned by hand

- Our rule:

**If you can easily detect the target face in the image, please keep it no matter what quality it is.**

# Problems (Consider to remove)

- 1. Occlusion - common problem

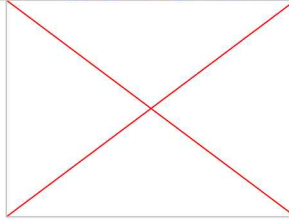
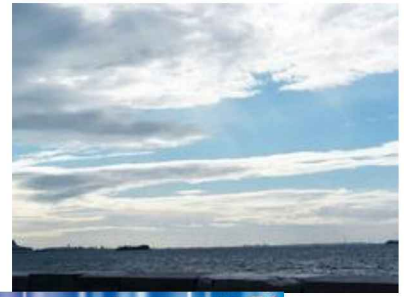
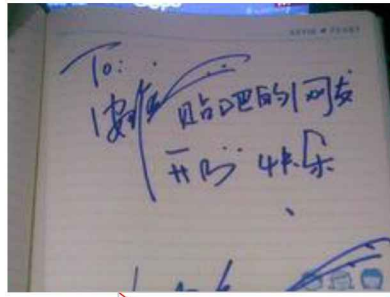
- Heavily occluded, hardly recognize whether the face belongs to the target person (Remove)



- Although occluded, it is not so hard to recognize that the face belongs to the target person (Keep)



● 2. No faces in the image



● 3. Too many faces in one image – common problem

- So many persons, really hard for you to recognize the target face (Remove)



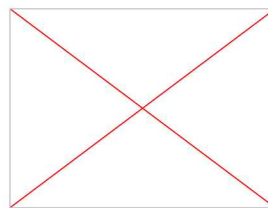


– Although many faces, easy for you to recognize the target face (Keep)



● 4. Crowded & Noisy Background

○ Background is so noisy that hard to find the target face (Remove)



– Although background is noisy, the target face is easy for you to detect (Keep)



● 5. Multi-face image

○ Multiple faces of the target subject, clear to recognize (Keep)



- Multiple faces of different persons, the target face is in the image (Keep)

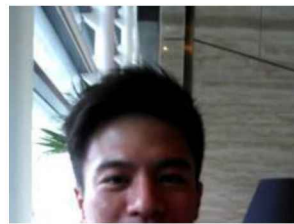
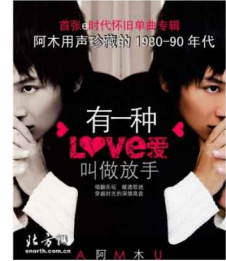


- Multiple faces of different persons, the target face is hard to detect in the image (Remove)



● 6. Partial faces - common problem

- The face in the image is cut by edge, hard to obtain facial landmarks (Remove)



- Although the face is partial, more than half of the face is shown and easy to recognize it (Keep)





## ● 7. Cartoon & sketch

○ Some images are easy to be viewed as cartoon or sketch (Remove)

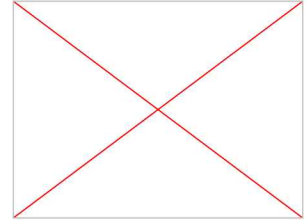


– Although it seems that the face in the image is like a cartoon face, if it is close to a real target face (Keep)

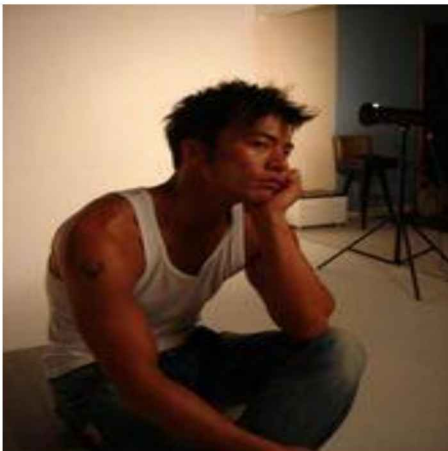


● 8. light problem

- Too dark or too bright, hard to recognize the face (Remove)

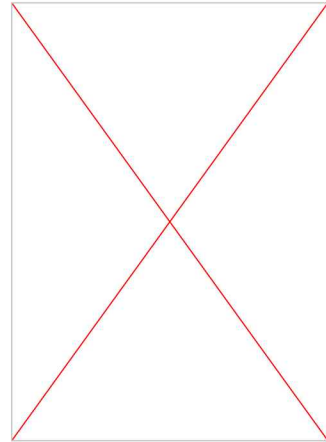


– Although there exists illumination problem, it is still easy to identify the target face (keep)



## ● 9. Poster

- Same rule; Keep it if target face is easy for you to detect, or remove it.



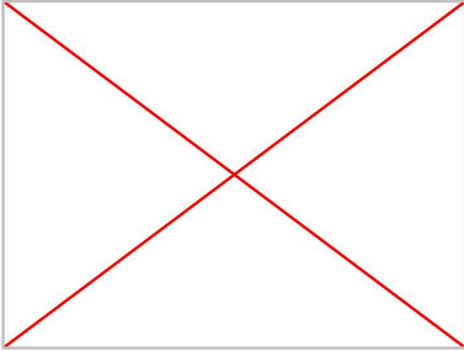
**Keep**

**Remove**

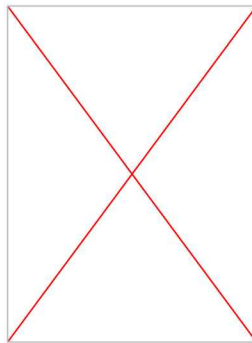


● 10. Large pose

- Face contains heavy pose problem, hard to recognize it (Remove)



– The pose is not so large, it is still easy for you to verify whether the face is target person (Keep)





● 11. Too small/vague faces

○ Hard to detect faces (Remove)



– For you, can figure out the target face (Keep)





- 12. Faces in image are totally different with the target faces shown (Remove)

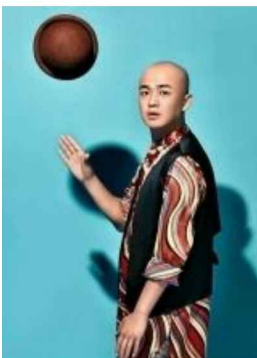


**Target identity**



**Wrong identity, Remove**

- 13. Wrong Gender



**Target Identity**



**Wrong face, Remove**

● 14. Damaged image

○ Fail to display (Remove)



Consider to keep

✓ One complete face of right person





More than one person

1. The correct person has full face
2. Others' faces are partial, small or vague



More faces in one image

1. The right person with complete face



# Cleaning Result

- A total of 356,381 images of 7,676 subjects

| name         | count  |
|--------------|--------|
| 0_New_Images | 14352  |
| 1_baike      | 140261 |
| 2_tieba      | 40264  |
| 3_mingxing   | 19616  |
| 4_n63        | 10216  |
| 5_tupianzj   | 3489   |
| 6_google     | 127426 |
| 7_new        | 757    |
|              | 356381 |