

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

Occupation includes

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

How to find?

● Mainly Chinese Websites

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



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 Data Library website. At the top, there's a navigation bar with tabs for 首页 (Home), 明星 (Star), 电影 (Movie), 电视 (TV), 音乐 (Music), and 排行榜 (Ranking). The '明星' tab is highlighted with a green border. Below the navigation is a search bar with placeholder text '按关键字查找' (Search by keyword) and a '搜索' (Search) button. To the right of the search bar is a '热门搜索' (Hot Search) section with terms like 李嘉欣, 伊能静, 烈焰奇侠, 黄金军团, and 珠光宝气. On the left, there's a sidebar with sections for '按地区检索' (Search by region) and '趣味查询' (Interesting queries). The '按地区检索' section has a dropdown menu with options: 全部明星, 内地明星 (highlighted with a green border), 港台明星, 亚洲明星, and 欧美明星. The '趣味查询' section includes '同生日查询' (Search by birthday) and '同星座查询' (Search by zodiac). The main content area displays a grid of star names categorized by letter (A-J, K-T) and number (0-9). Some stars listed include 安又琪, 安琥, 爱戴, 爱乐团, 艾敬, 艾尔肯, 艾梦萌, 阿朵, 阿穆隆, 阿斯茹, 阿修罗乐队, 阿丘, 阿兰, 阿里郎, AK-47乐队, A-OK乐队, 布衣乐队, 波澜童话乐队, 痴情乐队, 边原, 贝维, 北帝里兹乐队, 韩国安, 韩晶.

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. At the top, there's a logo for '乐酷明星网' (6Kstar) and the URL 'star.6k.com'. Below the logo is a subtext: '国内最大全球明星档案大宝库'. The top navigation bar includes links for 首页 (Home), 明星档案 (Star Archive), 明星新闻 (Star News), 明星图库 (Star Photo Collection), and 明星写真 (Star Photo). The page title is '中国内地明星' (Chinese Mainland Stars). The search bar at the top right contains fields for '造型' (Style), '剧照' (Photo), '型男' (Hunk), '帅气' (Handsome), '深V' (Deep V-neck), '性感' (Sexual), '妩媚' (Feminine), '优雅' (Elegant), '甜美' (Sweet), '清纯' (Pure), and '清新' (Fresh). There's also a search button and a field for '输入姓名或图片种类' (Enter name or photo type). Below the search bar are filters for '按职业' (By profession) and '按地区' (By region). The '按职业' filter includes 演员, 歌手, 全能艺人, 主持人, 导演, 模特, 音乐家, and 其它. The '按地区' filter includes 中国内地 (highlighted with a green border), 港台, 亚洲, 欧美, and 其它. The '按字母' filter shows a grid of letters from A to Z. The main content area displays a grid of star names under the heading '中国内地明星'. Some stars listed include 安又琪, 安琥, 爱乐团, 爱戴, 艾敬, 艾尔肯, 艾梦萌, 阿朵, 阿穆隆, 阿斯茹, 阿修罗乐队, 阿丘, 阿兰, 阿里郎, AK-47乐队, A-OK乐队, 布衣乐队, 波澜童话乐队, 痴情乐队, 边原, 贝维, 北帝里兹乐队, 韩国安, 韩晶.

7/30/2017

Summary of Data Collecting

8

(3) 天天排行榜 (ttpaihang)

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



7/30/2017

9

Summary of Data Collecting

(4) 百度百科人气榜 (star rank)

- <http://baike.baidu.com/starrank>

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	吃吃饭睡睡觉
6	黄子韬	77136	Devil丶心痛
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

10

Summary of Data Collecting

(5) 明星写真馆 (n63)

- <http://www.n63.com/>

华人女星		华人男星		韩国女星		韩国男星		日本女星		日本男星		欧美女星		欧美男星		体育明星		素材图库									
明星全列表，可按拼音首字母筛选。提示：按 CTRL+F 快速查询页内明星，详细搜索请移步右上角。																											
ALL		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
华人女星	2R	3T	4 In Love	ASOS	大S(徐熙媛)		小S(徐熙娣)		Cookies		Danielle																
Lili Tien	Maggie Q	One Fifth	S.H.E	Sweety	Twins		阿娇(钟欣潼)		阿Sa(蔡卓妍)		爱戴																
艾敬	徐立	阿朵(符莹)	艾莉丝	艾梦萌	艾尚真		艾雪		Amanda		安达熙																
安妮	安琥	安琪儿	安心	安心亚	安雅		安以轩		安又琪		阿茹娜																
张含韵	阿斯茹	ayawawa	阿幼朵	BabyQ	白百何		白冰		白静		白灵																
佰灵	白庆琳	白珊	白歆惠	白雪	柏雪		白延琴		白雨		班嘉佳																
鲍蕾	鲍逸琳	贝心渝	本多RuRu	边潇潇	Bibi		贝安琪		冰清玉洁		柏嘉莹																
柳妍宏	董伊莎	慈飞雨	慈慧琳	慈纬瑶	慈蔚琳		慈琴		慈培臻		慈斯筠																

7/30/2017

Summary of Data Collecting

11

(6) 百度知道 (zhidao)

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

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 search results for the query '百度风云榜' (Baidu Top Buzz). The main content is the 'Personality Buzz Ranking' for hosts. The left sidebar lists categories like 'Entertainment Icons', 'Female Actors', 'Male Actors', etc., with 'Host' being the selected category. The right side displays a table of hosts ranked by search volume, with detailed profiles for each.

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	董卿	简介 贴吧 视频	4357 ↓
3	撒贝宁	简介 贴吧 视频	3015 ↓
4	沈梦辰	简介 贴吧 视频	2846 ↓
5	何炅	简介 贴吧 视频	2749 ↓
6	谢娜	简介 贴吧 视频	2435 ↓

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> 展开 ▼

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		www.xuexila.com/news/1243489.html	7)
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, 李忠谊, 么正杰, 李雪刚, 崔蓉荣, 金炜, 徐国义, 朱志根, 韩冰岩, 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: 程洁, 王强	
		女运动员:	
		熊丽琴, 陈萍, 康立, 陈洁, 陈欣怡, 中婧婧, 叶连文, 干婉佳, 姜鑫, 董洁, 汤梦, 曹柏, 田翌雪, 孙琳	



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>

刘德华

刘德华 | 39 | 1351

刘德华写真

刘德华主演的电影

刘德华电视剧剧照

https://baike.baidu.com/pic/%E5%88%A9%EF%BC%9E%EF%BC%9A114923?fr=lemma

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/ImageScraper/>

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

The screenshot shows the Baidu Baike search results for the name '宋佳'. The search bar at the top has '宋佳' entered. Below the search bar, there are several red boxes highlighting different search results:

- 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
- 青年女作家

Below these results, there is a list of URLs:

- ① baike.baidu.com/item/宋佳/7175785#viewPageContent
- ① baike.baidu.com/item/宋佳/5812#viewPageContent
- ① baike.baidu.com/item/宋佳/10484754#viewPageContent
- ① baike.baidu.com/item/宋佳/17648669#viewPageContent

A blue speech bubble on the right contains the text "not follow the pattern".

7/30/2017

Summary of Data Collecting

23

❖ No album exist

The screenshot shows a profile page for a person. At the top, there is a table with basic information:

中文名	张经伟	职 业	电影编剧、导演
国 籍	中国	主 演	励志《上流才华》获得优胜奖
生 出 地	香港	代 表 作 品	《大话西游之爱你一万年》
性 别	男	相 片	广东深圳电视剧

Below this, there is a '剧本' (Script) section with two items:

1. 剧本
2. 导演作品

Under the '剧本' heading, there is a thumbnail image of a group of people. Below the '导演作品' heading, there is a thumbnail image of a man in a suit holding an award. A small note at the bottom left says '调用标题： 编剧， 导演， 演员人物， 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-scrapers", "-s", save_dir, URL])
    except Exception:
        print subject_name + ' skipped.\n'
        continue
    
```

- 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

- 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/>

The screenshot shows the homepage of mingxing.com. The top navigation bar includes links for 明星网 (Mingxing), 资料 (Information), 图库 (Photo Library), 剧照 (Stage Photos), 星闻 (Star News), 电影 (Movies), 电视剧 (TV Shows), and 独家策划 (Exclusive Planning). Below the navigation, there are tabs for 明星资料 (Star Information), 内地明星 (Mainland Stars), 港台明星 (Hong Kong/Taiwan Stars), 日韩明星 (Japanese/Korean Stars), and 欧美明星 (Western Stars). The main content area features a section for 内地明星 (Mainland Stars) with thumbnail images of several celebrities. A specific star profile for 吴千语 (Wu Qianyu) is shown, featuring her photo, a bio, and links to her personal page, career history, details, latest news, and photo library. Another profile for 陈龙 (Chen Long) is also displayed, showing his photo, birth information, and links to his personal page, career history, details, latest news, photo library, stage photos, movie works, and TV works. At the bottom, there is a grid of various star photos.

① 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

吴千语 主页 图片 星闻 个人资料

2010年，拍摄Tempo纸巾电视广告，获黄百鸣赏识成为天马电影出品有限公司合约艺人。

① 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 电视作品

陈龙图片大全/写真图片/生活照/活动照片

陈龙 (Chen Long)

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.

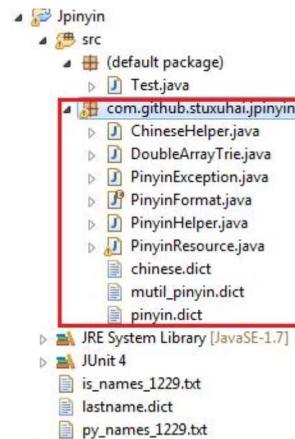
- Method: Java + Eclipse + Jpinyin

- Jpinyin: Java Class Library



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

tfboys图片大全/写真图片/生活照/活动照片



❖ 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 shows the homepage of n63.com. At the top, there's a search bar with the URL 'www.n63.com'. Below it, a banner says 'n63.com 明星写真馆：首页 /'. A message at the top of the page reads: '欢迎经纪公司/艺人提供高清大图。本站共享的图片均经过严格审核筛选。内容健康，欢迎监督。' Below this are several buttons for different 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 has three sections: 1. '华人男星 陈楚河' with a link 'www.n63.com/n_china/chenchuhe'. 2. '华人女星 陈慧娴' with a link 'www.n63.com/n_china/chenhuixian'. 3. A grid of four thumbnail images of a female celebrity. Each section has a '设计图库' (Design Library) and '剧照' (Stills) button.

7/30/2017

Summary of Data Collecting

38

● Pattern of URL

- Female

- http://www.n63.com/n_china/chenhuixian

- Male

- 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/xiezen/>

7/30/2017

Summary of Data Collecting

44

- Pattern of URL
 - <http://www.tupianzj.com/mingxing/xiezen/quoketong/>
- 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/xiezen/' + 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



7/30/2017

Summary of Data Collecting

47

- ❖ pinyin names are not correct
 - Polysyllabic
- ❖ For bands, fail to download by individual names
 - Tfboys
- ❖ There is no such person in this website
- ❖ Different words have 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

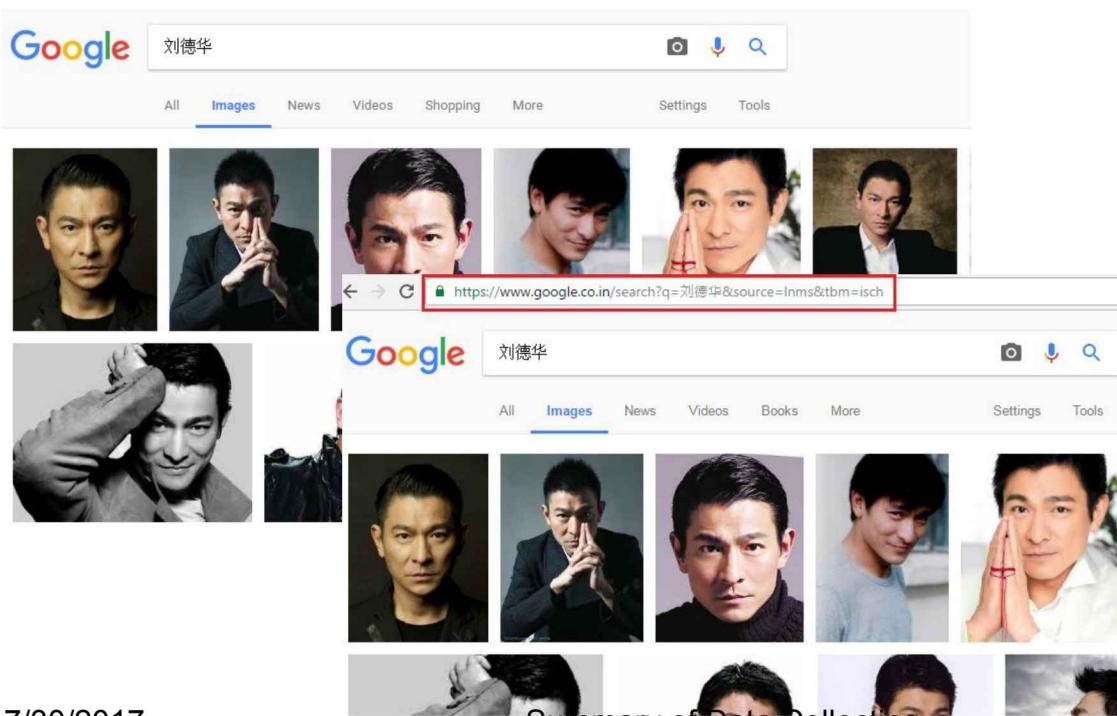


7/30/2017

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&tbo=isch>
- Use Chinese words in URL
- Not use ‘ImageScraper’ package to download
 - Fail to download from Google
- Write new script

7/30/2017

Summary of Data Collecting

51

● Download the first 25th 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&tbo=isch"

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

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

    for x in browser.find_elements_by_xpath("//div[@class='rg_meta']"):
        counter = counter + 1
        if succcounter > 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()
            succcounter = succcounter + 1
        except:
            print "can't get img"
    browser.close()
  
```

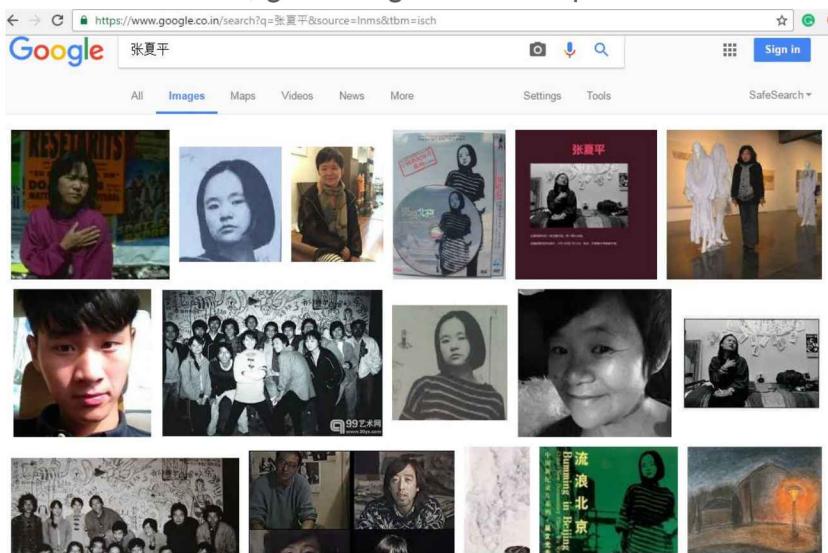
7/30/2017

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



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	7,190	234,490	Big noise	Yes
In Total	9,128	537,885		

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	7,190	234,490	Big noise	Yes
In Total	9,128	537,885		

- Noise distributions in them are **not even**
- For Baidu 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

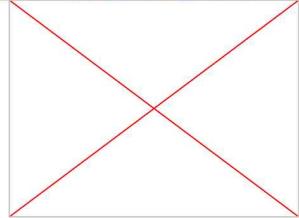
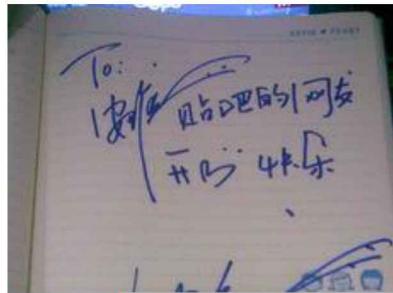
- Heavily occluded, hardly recognize whether the face belongs to the target person (Remove)



- 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)

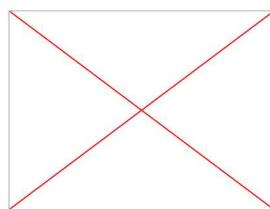


- Although many faces, easy for you to recognize the target face (Keep)



● 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)

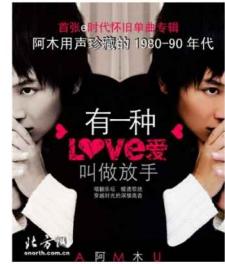


- Multiple faces of different persons, the target face is hard to detect in the image (Remove)



● 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)

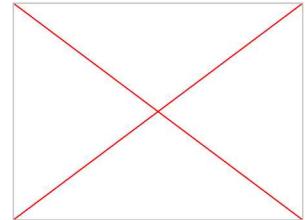


- Although it seems that the face in the image is like a cartoon face, if it is close to a real target face (Keep)

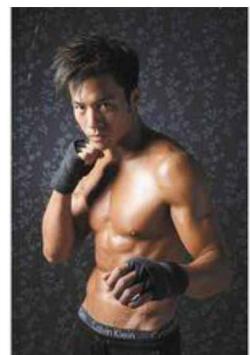


● 8. light problem

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

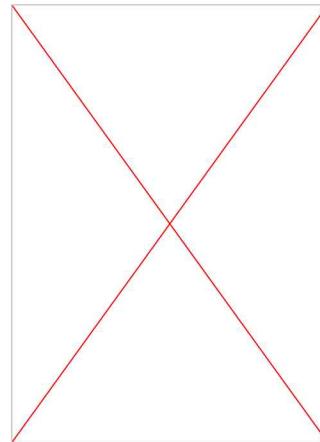


- Although there exists illumination problem, it is still easy to identify the target face (keep)



● 9. Poster

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



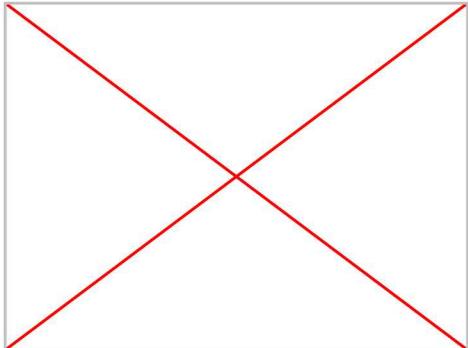
Keep

Remove

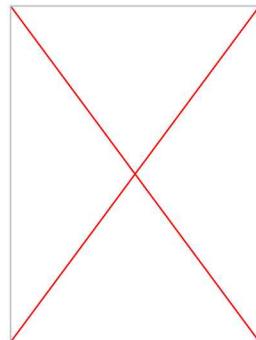


- 10. Large pose

- 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)

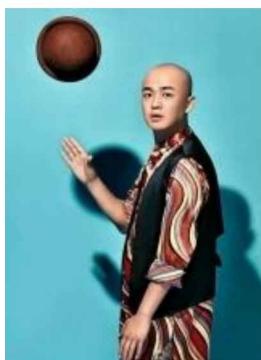


Target identity



Wrong identity, Remove

- 13. Wrong Gender



Target Identity



Wrong face, Remove

● 14. Damaged image

- Fail to display (Remove)



0-aTt3RC.jpg



32857.jpg

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