2025/04/24

微博关键词搜索工具

第一部分：微博搜索工具介绍

什么是微博关键词搜索工具?

这是一个基于Python开发的爬虫程序，可以:

- **自动搜索并获取**包含指定关键词的微博内容
- **批量下载**相关的微博文字、图片和视频
- **持续运行**不受50页微博限制
- **灵活筛选**多种类型的微博
- 对于研究、数据分析、舆情监测非常有价值!

第二部分：功能详解

核心功能：搜索范围

- **关键词搜索**：单个词、多个词、话题标签
  - 例如："迪丽热巴"、"疫情 口罩"、"#北京奥运#"
- **时间范围**：精确到日期
  - 例如：2020年3月1日至2020年3月16日
- **突破限制**：自动细分搜索条件
  - 当结果超过50页时，自动按小时、分钟细分搜索

核心功能：数据获取

对于热门关键词：

- 一天时间范围内可获取1000万+微博
- 扩大时间范围可获取更多数据
- 几乎能获取**全部**相关微博

例如搜索"新冠疫情"：

- 1天范围：可获取约1000万条微博
- 10天范围：可获取约1亿条微博

支持的保存方式

CSV文件（默认开启）

- 类似Excel的表格文件，方便查看和分析
- 可以直接用Excel、WPS等软件打开查看
- 数据以表格形式整齐排列，便于后续处理
- 无需额外软件即可使用

可以获取的微博信息

基础信息:
- 微博ID和bid（系统唯一标识符）
- 微博正文内容
- 发布时间和发布位置
- 发布工具（如iPhone客户端、华为手机等）

互动数据:
- 点赞数、转发数、评论数

多媒体内容:
- 图片URL和图片内容
- 视频URL和视频内容
- 头条文章URL

第三部分：安装步骤详解

前置条件

在开始前，请确保您的电脑已安装：

1.Python（3.6或更高版本）

1. 官网下载：python.org
2. 安装时勾选"Add Python to PATH"

2.Pycharm（之前已经安装了）

第三部分：安装步骤详解

前置条件

在开始前，请确保您的电脑已安装：

1.Python（3.6或更高版本）

1. 官网下载：python.org
2. 安装时勾选"Add Python to PATH"

2.Pycharm（之前已经安装了）

第四部分：配置详解

步骤1：找到配置文件

配置文件位于程序目录下的：

weibo/settings.py

您可以：使用pycharm等编程工具打开

Windows：

notepad weibo\settings.py

Mac/Linux：

nano weibo/settings.py

步骤2：设置Cookie（必须）

1.什么是Cookie?

1. Cookie是网站用来识别用户身份的数据
2. 微博搜索需要登录状态的Cookie才能获取完整结果

2.在配置文件中找到:

3.python

4.DEFAULT_REQUEST_HEADERS = { 'Accept': '...', 'Cookie': 'your cookie' }

5.将'your cookie'替换为您的真实Cookie值

如何获取Cookie（新版微博）

1.打开Chrome浏览器，访问微博首页
2.点击"立即登录"，完成验证进入新版微博
3.按F12键打开开发者工具
4.点击Network（网络）选项卡
5.在右侧Headers（标头）中找到Request Headers（请求标头）
6.找到"Cookie:"后面的一长串文本，这就是我们需要的Cookie值
7.全选并复制这段Cookie值

步骤3：设置搜索关键词（必须）

在settings.py中找到并修改KEYWORD_LIST：

搜索单个关键词：

python

```
KEYWORD_LIST = ['迪丽热巴']
```

搜索多个关键词（分别获取各自结果）：

python

```
KEYWORD_LIST = ['迪丽热巴', '杨幂']
```

搜索包含多个关键词的微博（同时包含）：

python

- `KEYWORD_LIST = ['迪丽热巴 杨幂']`

步骤3：设置搜索关键词（续）

搜索微博话题：

python

```
KEYWORD_LIST = ['#迪丽热巴#']
```

从文件读取关键词：

1.创建一个文本文件，如keywords.txt

2.每行写一个关键词

3.在settings.py中设置：

python

```
• KEYWORD_LIST = 'keywords.txt'
```

步骤4：设置搜索时间范围（必须）

在settings.py中找到并修改：

```python
# 开始日期，格式为yyyy-mm-dd
START_DATE = '2020-06-01'
# 结束日期，格式为yyyy-mm-dd
END_DATE = '2020-06-02'
```

- 程序会搜索在这个时间范围内发布的微博，包括开始日期和结束日期当天。

步骤5：CSV文件设置

保存在哪里？

- CSV文件会保存在"结果文件"文件夹下
- 每个关键词会有一个独立的文件夹
- 文件名格式为：关键词_起始日期_结束日期.csv

查看CSV文件

- 可以用Excel、WPS等表格软件打开
- 每行代表一条微博
- 每列代表微博的不同属性（如内容、时间等）
- 建议初学者只使用CSV功能，简单易用且无需额外配置。

步骤6：其他常用设置

除了必须的Cookie、关键词和时间范围外，还有一些有用的设置：

设置等待时间：

python

- *# 每次请求间隔时间（秒）* DOWNLOAD_DELAY = 10

步骤6：其他常用设置

地区筛选

python

- *# 筛选发布微博的省份，不筛选请设置为"全部"* REGION = ['全部']
  # 例如：['北京', '上海', '广东']

步骤6：其他常用设置

步骤7：设置高级选项（续）

搜索细分阈值设置

python

- *# 是否进一步细分搜索的阈值* FURTHER_THRESHOLD = 46

第五部分：运行程序

运行程序（基本方式）

在命令行中，确保您在程序目录下，输入：

scrapy crawl search

程序将开始运行，并在控制台显示进度信息：

2025-04-22 12:34:56 [scrapy.utils.log] INFO: Scrapy 2.5.1 started ...
正在爬取:迪丽热巴 2020-06-01 00:00 2020-06-01 01:00 第1页 ...

第五部分：运行程序

程序运行过程中或结束后，会在当前目录下生成"结果文件"文件夹，内含：

1.关键词命名的文件夹：

1. 每个关键词一个独立文件夹

2.CSV文件：

1. 可用Excel或WPS打开
2. 文件名格式为：关键词_起始日期_结束日期.csv
3. 包含所有获取到的微博数据
4. 每行代表一条微博，每列代表不同的微博属性

表格数据已导入且可调整。

id	bid	user_id	用户昵称	微博正文
4877757841148450	MwCPS4OLo	5953744101	粉巷财经	【#丫丫回国后的房间准备好了#吃喝自由还有同伴一起玩】#丫丫回国后有伙伴一起
4877757710600414	MwCPF2wc6	1859620007	ALBEE_莫莫芬_	大熊猫美香添添超话为动物发声超话孟菲斯的丫丫乐乐超话只能说两国都低估了大
4877757533914812	MwCPngqq8	1793676294	xinxin1955	求求留一点人间的美好吧#丫丫#拒绝大熊猫旅美，拒绝大熊猫租借给美国，拯救美
4877757168749289	MwCOMAI5z	1891257890	Chill寒	#在美华人拍丫丫现状#谢谢在美华人的支持和关注🙏🙏🙏🙏，丫丫的状态越来越
4877757111077856	MwCOH4woM	7367840589	女朋友的吐槽箱	【#游客称重庆动物园的熊猫太会营业#共同呼吁丫丫早日回家！】3月5日（拍摄时
4877757106885314	MwCOGsTbs	7720549455	野性小仙女们	3月9日，汇源集团回应想认养大熊猫丫丫：为避免营销之嫌，正很小心做这件事。
4877756917092571	MwCOntL6j	1929862885	成都彭州公安	【#丫丫回国后将住在北京动物园#实现吃喝自由】3月8日，中国动物园协会称，北
4877742824754493	MwCrEjWRn	5521175085	大霞Fm	#丫丫回国将实现吃喝自由#希望大家也关注一下“美香”一家，美香也是旅美大熊猫
4877742800111493	MwCrC0t0h	2463270200	莱莉妈妈	孟菲斯的丫丫乐乐超话#丫丫#如果你累了可以休息，我做好自我调节，我休息几小
4877742430488361	MwCr1232N	1249729383	新浪女性	【#为什么大熊猫今年扎堆回国#】旅日大熊猫香香、永明、樱浜、桃浜，旅美大熊
4877742389331838	MwCqWD9Dw	1625064501	冰城网	哈尔滨一商场投屏“关注大熊猫丫丫”@哈尔滨的今天L冰城网的微博视频
4877742354989145	MwCqTkVU5	6017283028	一见读书BooksAlive	【功夫熊猫！#大熊猫舜舜几分钟内毁掉一条船#】3月6日，海南海口。大熊猫舜舜
4877742351320041	MwCqT5xoZ	5617775353	全球拍视频	【功夫熊猫！#大熊猫舜舜几分钟内毁掉一条船#】3月6日，海南海口。大熊猫舜舜
4877742316978695	MwCqPthtB	7509958552	夏语倾耳	#丫丫回国将实现吃喝自由#孟菲斯的丫丫乐乐超话今天看到丫丫缩成一团的样子，
4877742296011792	MwCqNpdWo	5521175085	大霞Fm	#丫丫回国将实现吃喝自由##美香添添小崽要一起回家##救救美香#请大家多多关注
4877727700353451	MwC3g1tWP	2035162954	韩辉在北京	#丫丫回国将实现吃喝自由#各国养熊猫为什么差距这么大？#丫丫#L韩辉在北京的
4877727259431900	MwC2xDzFq	5025809344	乐乐菜菜的小辫儿哥哥	#丫丫回国将实现吃喝自由#快接回来吧！焦急等待中#关注中国旅美大熊猫#还有美
4877727205164405	MwC2slFuR	3496691321	视觉银川	【#在美华人拍丫丫现状#：不光吃竹子还吃水果了！】当地时间3月8日，在美华人

问题1：Cookie设置后仍无法运行

可能原因：

- Cookie已过期（一般2-3天后会过期）
- Cookie复制不完整或有多余字符
- 微博账号状态异常

解决方法：

- 重新获取Cookie
- 确保完整复制（从第一个字符到最后一个）
- 尝试使用其他微博账号

问题2：程序运行速度很慢

可能原因:

- 避免被微博限制，程序设置了访问间隔
- 数据量大，处理需要时间
- 网络状况不佳

解决方法:

- 这是正常现象，请耐心等待
- 可以考虑缩小时间范围，分批次获取
- 确保网络连接稳定

问题3：没有搜索到任何结果

可能原因:

- 关键词拼写错误
- 时间范围内无相关微博
- Cookie无效或已过期
- 微博反爬虫机制触发

解决方法:

- 检查关键词拼写
- 尝试扩大时间范围
- 重新获取Cookie
- 增加DOWNLOAD_DELAY值

第七部分：实际应用示例

示例1：追踪品牌舆情

目标：获取某品牌一周内的微博讨论

设置：

python

- KEYWORD_LIST = ['品牌名'] START_DATE = '2025-04-15' END_DATE = '2025-04-22' WEIBO_TYPE = 0 *# 获取全部微博*

第七部分：实际应用示例

示例2：收集行业热点

目标：获取AI行业最近的热门讨论

设置：

python

- KEYWORD_LIST = [‘deepseek', 'AI', '机器学习'] START_DATE = '2025-03-22' END_DATE = '2025-04-22' WEIBO_TYPE = 2 *# 获取热门微博*

第七部分：实际应用示例

示例3：学术研究数据收集

- **目标**：收集某社会事件的公众反应

设置：

python

- KEYWORD_LIST = ['事件名称', '#相关话题#'] START_DATE = '事件发生日期' END_DATE = '当前日期' CONTAIN_TYPE = 0 *# 不限内容类型* REGION = ['全部'] *# 不限地区*

现在已经可以独立使用
这个工具进行微博数据采集了!

推特（X）搜索工具

2025/04/24

1. 工具介绍

什么是Twitter文本爬虫工具?

- 一个基于Python开发的推特文本数据采集工具
- 可以下载指定用户的所有推文（文本内容）
- 支持时间范围筛选
- 可以选择是否包含转发内容
- 保存为CSV格式，方便查看和分析

2. 功能概述

2.1 核心功能

- **多用户采集**：支持同时获取多个Twitter用户的推文
- **时间筛选**：可以指定开始和结束日期范围
- **格式化保存**：将推文内容和相关信息保存为CSV表格
- **转推筛选**：可以选择是否包含用户转发的内容

2.2 获取的数据内容

- 用户基本信息（显示名称、用户名）
- 推文发布时间
- 推文URL链接
- 推文文本内容
- 互动数据（点赞数、转发数、回复数）

3. 安装准备

3.1 安装Python环境

1.访问 python.org 下载Python（3.6或更高版本）

2.安装时勾选"Add Python to PATH"选项

3.完成后打开命令提示符，输入python --version验证安装成功

3.2 安装必要的库

在命令提示符或终端中执行：

pip install httpx

- httpx是一个现代化的HTTP客户端，用于发送网络请求

3. 安装准备

设置Cookie（必须）

python

- cookie = 'auth_token=xxxxxxx; ct0=xxxxxxx;' *# 需要填入真实的 Twitter cookie（auth_token与 ct0字段）*

不赘述，和微博一样

```python
# 填入 cookie (auth_token与ct0字段) //重要:替换掉其中的x即可, 注意不要删掉分号

user_lst = ['jeleechandayo','yorukura_anime']
# 填入要下载的用户名(@后面的字符),支持多用户下载,在列表里添加即可

time_range = "2024-04-21:2030-01-01"
# 时间范围限制,格式如 1990-01-01:2030-01-01

has_retweet = False
# 是否包含转推

##########配置区域##########



def time2stamp(timestr:str) -> int:  2 用法
    datetime_obj = datetime.strptime(timestr,  format: "%Y-%m-%d")
    msecs_stamp = int(time.mktime(datetime_obj.timetuple()) * 1000.0 + datetime_obj.microse
    return msecs_stamp

start_time,end_time = time_range.split(':')
```

4. 运行

程序报错"API次数已超限"

可能原因:

- Twitter API有访问频率限制
- 短时间内请求过多

解决方法:

- 等待一段时间（通常几小时）再尝试
- 减少一次爬取的用户数量
- 使用不同的Twitter账号获取新Cookie

4. 运行

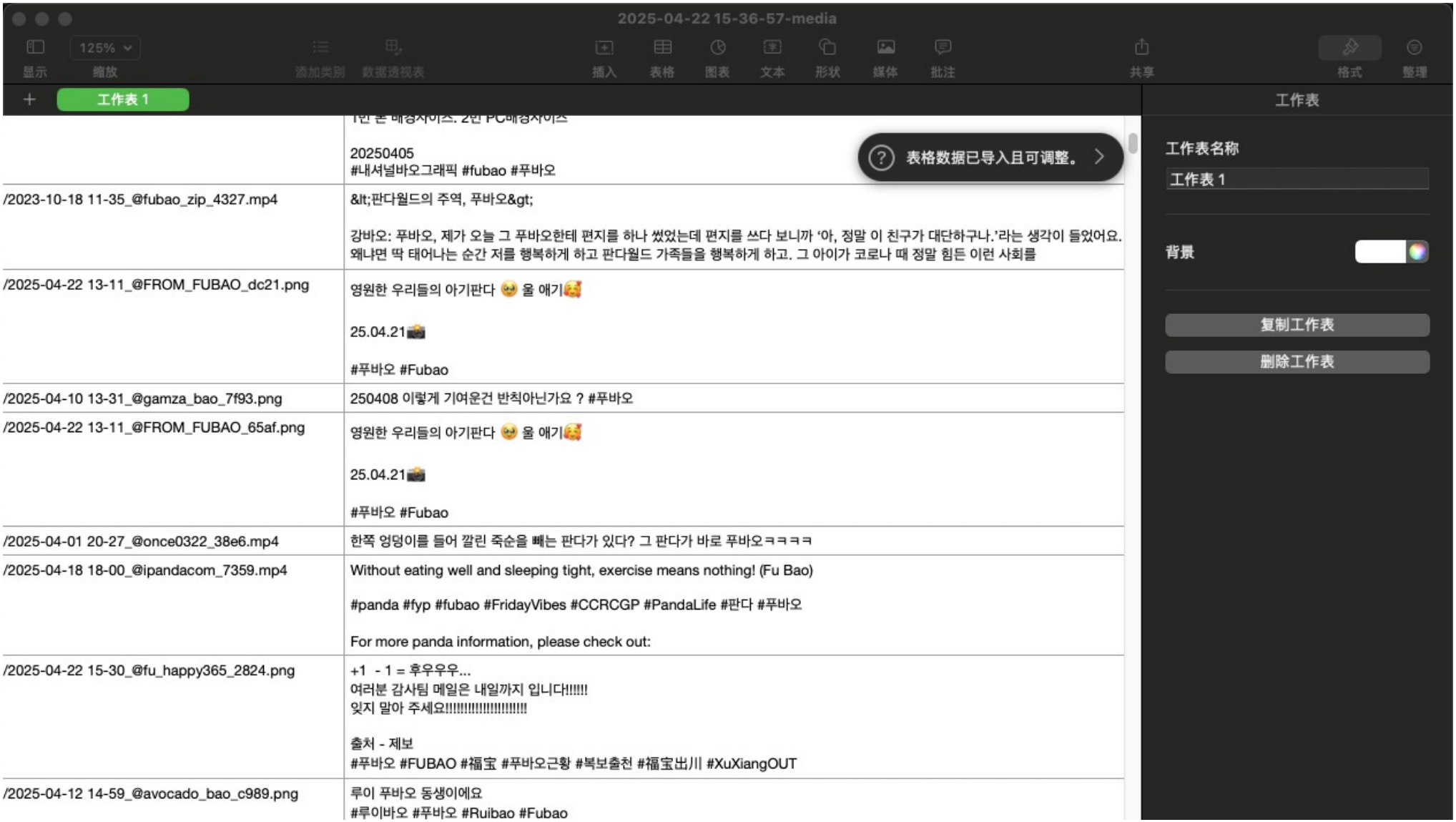
程序报错"API次数已超限"

可能原因:

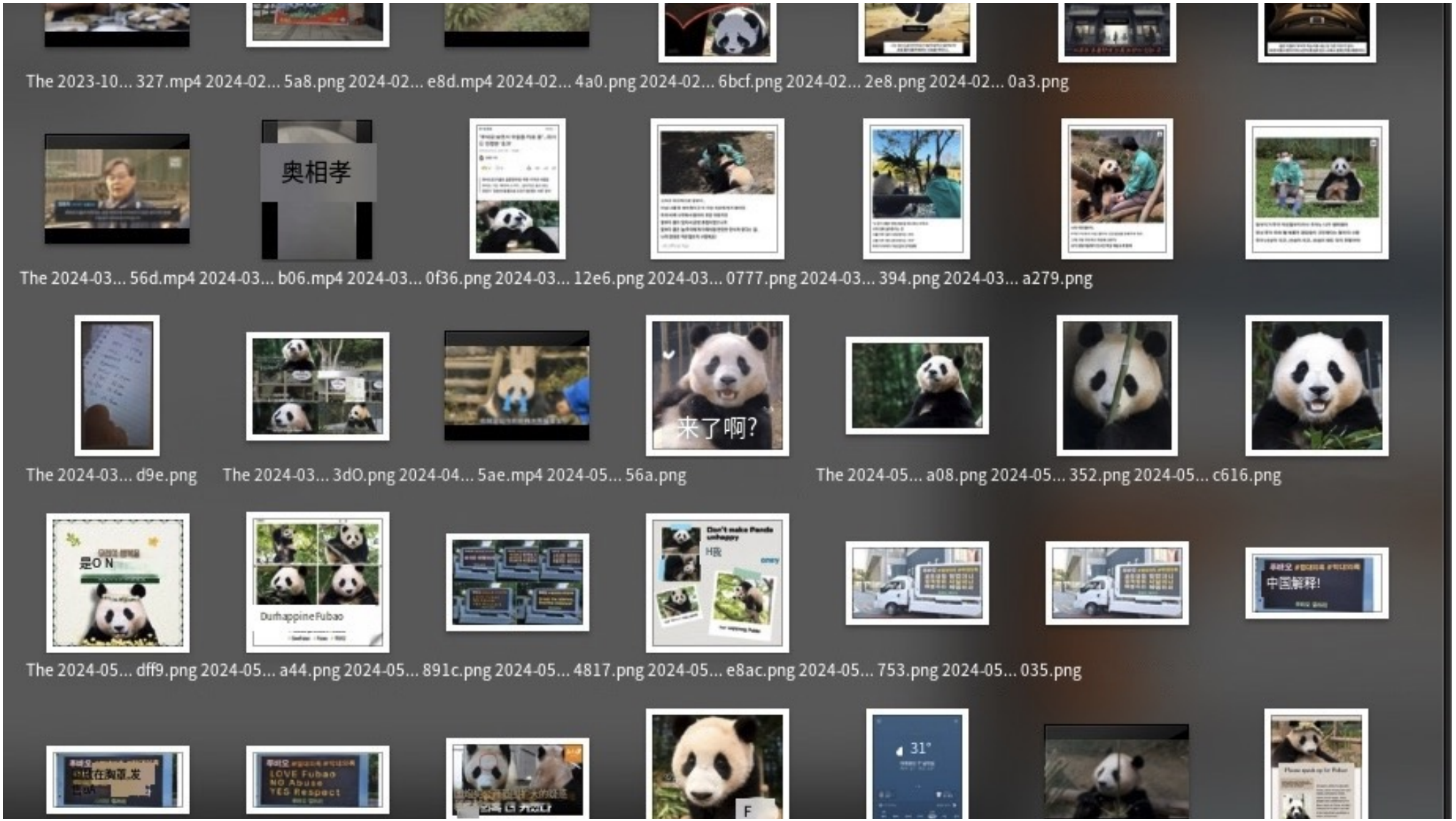
- Twitter API有访问频率限制
- 短时间内请求过多

解决方法:

- 等待一段时间（通常几小时）再尝试
- 减少一次爬取的用户数量
- 使用不同的Twitter账号获取新Cookie



	1만 본 배경자이즈, 2만 PC배경자이즈 20250405 #내셔널바오그래픽 #fubao #푸바오
/2023-10-18 11-35_@fubao_zip_4327.mp4	<판다월드의 주역, 푸바오> 강바오: 푸바오, 제가 오늘 그 푸바오한테 편지를 하나 썼었는데 편지를 쓰다 보니까 '아, 정말 이 친구가 대단하구나.'라는 생각이 들었어요. 왜냐면 딱 태어나는 순간 저를 행복하게 하고 판다월드 가족들을 행복하게 하고. 그 아이가 코로나 때 정말 힘든 이런 사회를
/2025-04-22 13-11_@FROM_FUBAO_dc21.png	영원한 우리들의 아기판다 🥹 올 애기🥰 25.04.21📸 #푸바오 #Fubao
/2025-04-10 13-31_@gamza_bao_7f93.png	250408 이렇게 기여운건 반칙아닌가요 ? #푸바오
/2025-04-22 13-11_@FROM_FUBAO_65af.png	영원한 우리들의 아기판다 🥹 올 애기🥰 25.04.21📸 #푸바오 #Fubao
/2025-04-01 20-27_@once0322_38e6.mp4	한쪽 엉덩이를 들어 깔린 죽순을 빼는 판다가 있다? 그 판다가 바로 푸바오ㅋㅋㅋㅋ
/2025-04-18 18-00_@ipandacom_7359.mp4	Without eating well and sleeping tight, exercise means nothing! (Fu Bao) #panda #fyp #fubao #FridayVibes #CCRCGP #PandaLife #판다 #푸바오 For more panda information, please check out:
/2025-04-22 15-30_@fu_happy365_2824.png	+1 - 1 = 후우우우... 여러분 감사팀 메일은 내일까지 입니다!!!!!! 잊지 말아 주세요!!!!!!!!!!!!!!!!!!!!!! 출처 - 제보 #푸바오 #FUBAO #福宝 #푸바오근황 #복보출천 #福宝出川 #XuXiangOUT
/2025-04-12 14-59_@avocado_bao_c989.png	루이 푸바오 동생이에요 #루이바오 #푸바오 #Ruibao #Fubao


奥相孝
来了啊?
Don't make Panda unhappy
DurhappineFubao
LOVE Fubao
NO Abuse
YES Respect
中国解释!
31°

现在已经可以独立使用
这个工具进行推特数据采集了!