目录

[腾讯视频弹幕爬虫（多功能版：腾讯弹幕爬虫.py） 2](#_Toc39134166)

[腾讯视频弹幕爬虫（URL版：腾讯弹幕爬虫\_使用版.py） 5](#_Toc39134167)

[爱奇艺视频弹幕爬虫（多功能版：爱奇艺弹幕爬虫.py） 6](#_Toc39134168)

[爱奇艺视频弹幕爬虫（多功能版：爱奇艺弹幕爬虫.py） 8](#_Toc39134169)

[微博爬虫 9](#_Toc39134170)

[知乎爬虫 16](#_Toc39134171)

[百度贴吧爬虫 17](#_Toc39134172)

[大众点评爬虫 18](#_Toc39134173)

[Bilibili爬虫 19](#_Toc39134174)

# 腾讯视频弹幕爬虫（多功能版：腾讯弹幕爬虫.py）

1. 数据样例



1. 使用方法

在cmd或python ide运行程序，可以见到以下界面。



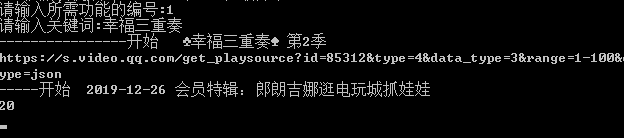
根据提示输入信息，完成弹幕爬取。爬取后的数据存储在MongoDB中。

1. 输入参数
2. 根据关键字爬取所有相关视频的弹幕

输入关键字后，程序会爬取搜索出来的所有系列里的所有视频。

如输入幸福三重奏后，会爬取第二季、加糖版第二季和第一季的所有视频的弹幕，但“郎朗凶吉娜”这种独立视频不会抓取。





1. 根据该系列的lid爬取所有相关视频的弹幕

该功能可以获取一个剧集的所有视频的弹幕。

所谓系列lid，即一个可以识别剧集系列的lid。通过该lid，程序可以把该系列的所有视频的弹幕抓下来。

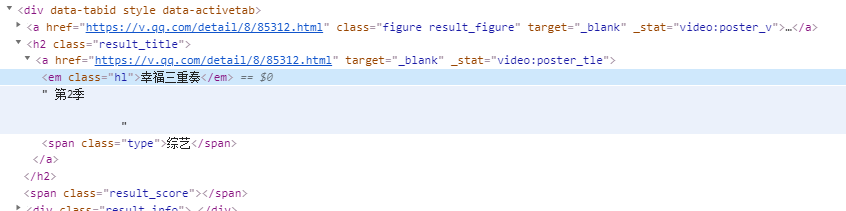
获取该lid的方法是，在网页搜索页面，点击键盘“F12”，会出现开发者工具，如下图。



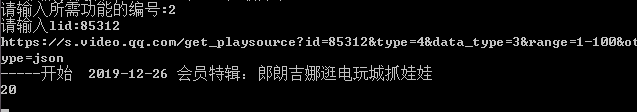
再点击系列标题，如下图。



此时，开发者工具会显示网页源代码，如下图。



其中，lid为<https://v.qq.com/detail/8/xxxxx.html>中的xxxxx。



1. 根据该视频的lid、cid和vid爬取弹幕

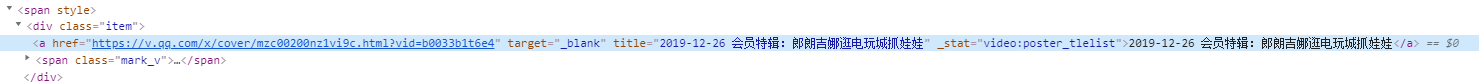
该功能可以获取单一视频的弹幕。

Lid的获取方式请参照上文2。

cid和vid的获取方式首先也是需要打开开发者工具（F12），再点击对应的视频标题。



此时，开发者工具显示如下。



cid和vid就在<https://v.qq.com/x/cover/xxxx.html?vid=yyyy>中，其中cid为xxxx，vid为yyyy。

# 腾讯视频弹幕爬虫（URL版：腾讯弹幕爬虫\_使用版.py）

1. 数据样例



1. 使用方法

输入name和url。



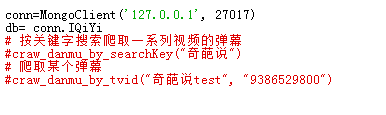
# 爱奇艺视频弹幕爬虫（多功能版：爱奇艺弹幕爬虫.py）

1. 数据样例



1. 使用方法

进入IDE，看到以下代码：



craw\_danmu\_by\_searchKey("奇葩说")是通过关键字爬取该系列所有视频的弹幕。

craw\_danmu\_by\_tvid("奇葩说test", "9386529800")是通过视频名字（用于命名弹幕数据库）和视频id来爬取该视频的弹幕。

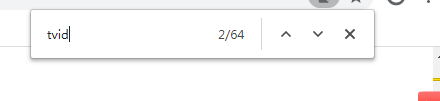
1. 输入参数

视频id的获取方式为：

进入视频页，点击鼠标右键，点击“查看网页源代码”



进入源代码页面，按ctrl+f，打开搜索功能，并输入tvid



找到tvid。



# 爱奇艺视频弹幕爬虫（多功能版：爱奇艺弹幕爬虫.py）

（1）数据样例



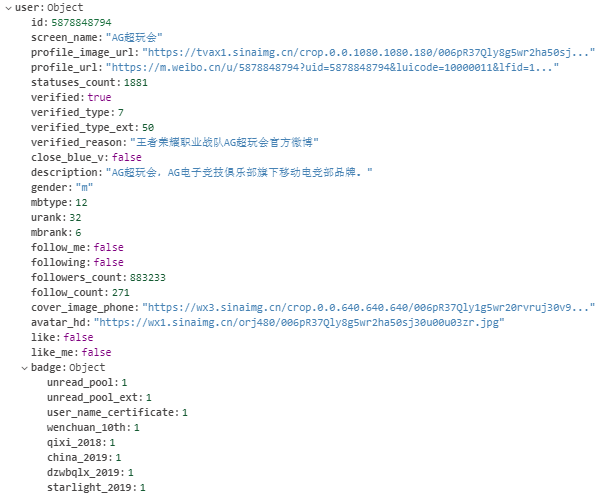
（2）使用方法



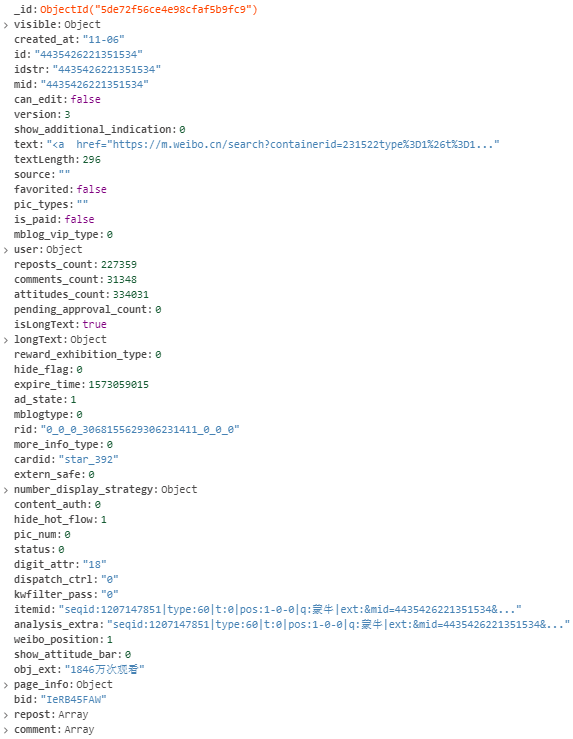
# 微博爬虫

1. 数据样例
2. 获取一个用户近n个月的所有微博





1. 根据关键字获取所有微博（包含评论、转发数据）



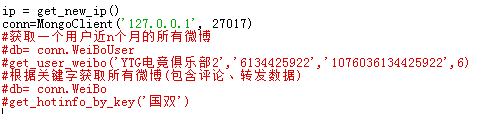






1. 使用方法

进入IDE，看到以下代码：



get\_user\_weibo('YTG电竞俱乐部2','6134425922','1076036134425922',6)是获取用户最近n个月的微博数据。

get\_hotinfo\_by\_key('国双')是根据搜索关键字获取关键字下的所有微博（包含转发和评论数据）。

get\_weibo\_by\_key\_time(key,beginDate,endDate,gap,cookie)是根据关键字获取指定时间内的所有微博（每天最多50页+只包含互动量数据）

1. 输入参数

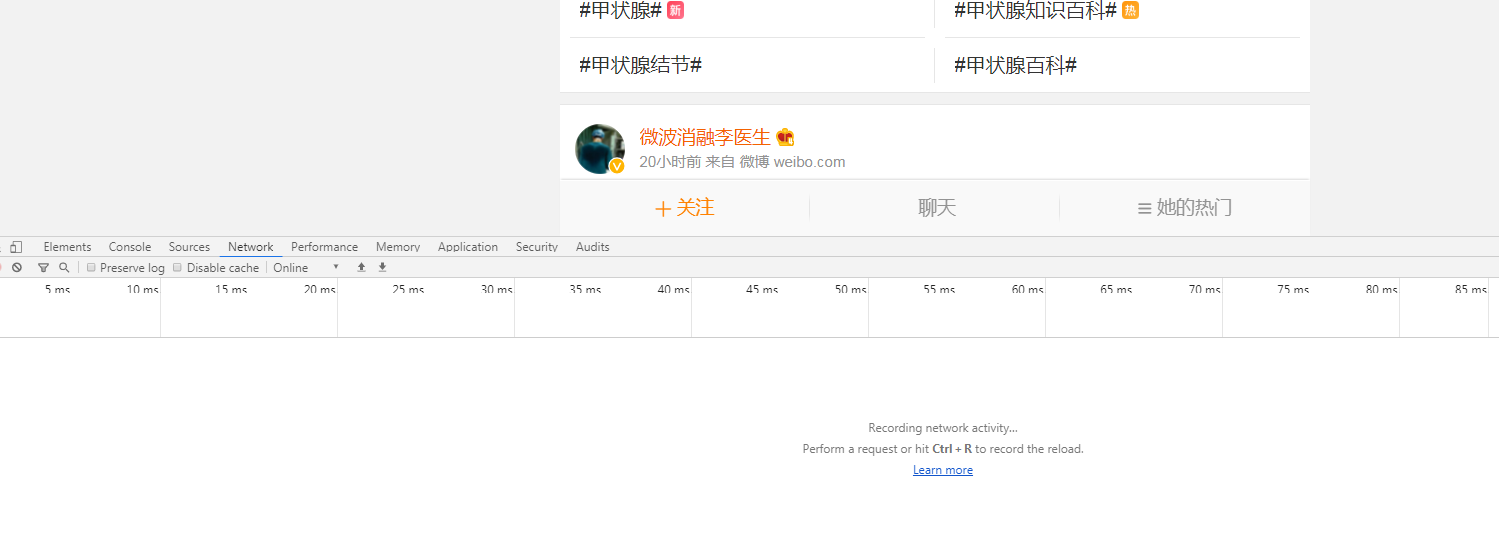
①get\_user\_weibo(用户名称，uid，containerid，近n个月)：

用户名称：用于保存数据库，准确与否没有关系。

Uid：该用户的id，可以从网址上看出来。如下的uid即为5780790015。注意，必须确认用户主页网址为m.weibo.cn/u/xxxxxx。



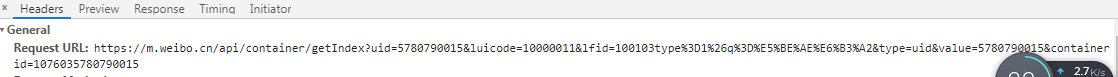
Containerid：账户自身的id，需要按f12打开开发者工具。



刷新一下微博页面，下方会出现一大堆东西。点击Network，寻找getindex开头的xhr文件。



点击，查看headers，此时可找到containerid。



②get\_weibo\_by\_key\_time(key,beginDate,endDate,gap,cookie)

Key为关键词。

BeginDate和endDate为开始和结束时间，输入格式为datetime.datetime(年,月,日)。

Gap为单位采集时间间隔。建议：搜索结果多的关键词可选gap小一点，搜索结果少（例如一个月才几条的），可选择大一点的gap，加快爬取速度。举例：若gap为1，则按日为单位采集，每日最大页数为50。若gap为5，则按10号-14号五天为单位采集，这五天的总最大页数也为50。

# 知乎爬虫

1. 数据样例



1. 使用方法

进入IDE，输入关键字，开始爬虫。



# 百度贴吧爬虫

1. 数据样例



1. 使用方法

打开IDE，输入搜索关键字。



# 大众点评爬虫

1. 数据样例



1. 使用方法

打开IDE，输入输出文件名、餐厅url和cookie，完成爬取。



1. 注意事项

若报错，一般为两个原因。

首先，看看文件夹里面有没有多出一个xxx.css.txt。



若没有txt，打开网页源代码，检查39行代码正则表达式中的url是否有变动，特别是划着的部分。如果有修改，将上下两句代码的url部分进行修改。



若有txt，将该txt删掉，并修改161行的值，一般为0、1、2，具体不确定，要尝试，一般每天都会变化。



# Bilibili爬虫

1. 数据样例







1. 使用方法

进入IDE，输入搜索关键字进行爬虫。

