视频下载与上传小软件

# 要求

快手（已经有小软件，可以暂时使用他们的）

西瓜视频

抖音

火山小视频

按id，类别，视频地址爬

视频，标题，长度，清晰度

去除短于15秒的视频，长于3分钟的视频。

**上传部分**

头条，

腾讯视频，

# 下载部分

发现有一个统一的下载视频的软件you-get，西瓜视频的部分虽然已经可以解析了，但是先测试you-get.

## You-get

安装python多版本支持。

安装与测试

## 一点资讯

策略，目前没有you-get的支持，可能需要一步一步的按照西瓜的逻辑来解析。

鉴于1，如果按照用户抓的话，可能其内容有的是非视频，鉴别会增加额外的工作量。2，按照用户抓的话目的也是形成一定的类别，而按照类别抓则很容易达到这个目的。

综上，决定按照类别爬。

获得某一类别下的视频类别，可能需要循环抓，类别下是https抓不到包。

看通过android查看是否可以强制转化为http来抓。

Paros竟然可以得到https的内容，感觉我们同样可以得到这个内容，事实证明python request可以使用https来爬取内容。

尝试类别抓，用户id抓，视频地址抓。

按类别抓，python post请求看结果，可以得到内容，但是需要登录，会话机制，需要先获得session\_id。

类别下的视频获取数量规则，采用循环刷新来获取视频的机制，已可行，每个目录可以下载1300条左右。

视频地址暂时如下：

<http://v1.go2yd.com/video/ea0b041c8bdbbf88a247f044e17f8be1.mp4_bd.mp4>

下载使用西瓜的下载方法即可。

界面，类别，考虑界面设计，融合西瓜、快手、一点资讯。

如果用网页版只能缓存，百度网盘，开会员；缓存到本地，得增大空间，且要通知机制。

而且两个都有网络的延迟，都有一个从中国到外国，然后再从外国到中国的过程，不完美。

因此对于界面决定还是用软件的形式来做，决定使用可以兼容python3的支持最多平台的kivy。

考虑python多个版本的问题，pyenv 还是 anaconda， 切换版本必须用pyenv， python3的环境，使用anaconda最高版本。

下载一部分视频先（萌宠, 搞笑）。

需要下载的类别：搞笑 萌宠 音乐 综艺 动物 生活 看欧美

Zsh

访问受限，应该是检查ip的问题，利用间歇性循环来解决这个问题，因为好像过一段时间又可以爬了。

看欧美下不了暂时的，需要调试，其接口不相同需要使用新的post接口, 其实只差一个appid，都添加即可， url爬下来发现不是视频的，而是其它的，有的时候按类别爬的搞笑的也是的，发现不能再用vieo\_url应该用下面视频列表中的视频，取高清的，建索引时更新用的名字，所以索引重新更新。

防重复, 类别每天重复的情况，以搞笑为例, 大概有1/3，通过为每个类别建立已经爬取的视频索引来解决，用的是服务器上视频的名字，应该是一个hash值，已完成。

Vim python 语法插件， 在vimrc里设置将语法高亮打开就可以, 还可以通过修改python.vim来修改语法高亮的具体设置。

开始使用kivy demo， 由于kivy无法用包来安装，只能用dmg的形式，相当于用kivy命令代替了python命令，这样就破坏了pyenv的执行环境，而且其额外包的安装并不清楚， 可以通过kivy –m pip install来安装，不过不知道为什么会出错，故而放弃，转用tkinter。

将一点资讯加入到界面逻辑中去。

下载时禁用frame再打开，已经通过检查下载来实现这个功能

日志记录下来以备忘

添加退出的问题，以关闭相应的资源，暂时绑定在窗口事件中。

存在一个视频有多个视频源的情况，暂不考虑。

Path的问题，输入，新建， 为了方便用默认的方法，默认用名字来代替，当前目录。

添加线程的监护线程，为了实现状态监督，最后的处理，成功时弹窗口，失败时有处理（至少有一个提示）。

新的http异常的处理，还是用循环等待的方法， connectionerror。

看一下日志中的错误跟上相同。

爬取指定类别的视频，用新程序。

添加时长的限制在用户面板上，放在了菜单里面，因为是通用的。

部署到机器上面去。

Json的编码有问题，先在本地windows测试一下，原来是python2与python3处理编码时的位置不同造成的。

熟悉了解anaconda， notebook.

开始使用起来，尤其是在mac上，就这样定了，但其不适合写大型的程序，还是适合写一些测试用的代码或者说交互式的数据分析文档用的，以后尝试着将所有的测试代码全部都移植到jupyter notebook上，所以编辑器上还是在pycharm与spyder上进行选择，鉴于自己在everonte上对于编辑器的核心要求，还是选择pycharm，其实spyder也是可以的。

对比与pycharm以确定自己的编辑器（见上）。

文件名保存前的处理，尤其是一些乱码字符？回车等的处理。

文件名乱码。

日志中加入时间戳

既然pythonw造成无法log\_main输入的事实（应该是因为dos窗口消失的问题，所有的到标准输出的重定向都没有打印出来），得想个办法将error给打印出来才行，日志或者是scrolledtext都可以。

**添加按用户id爬的逻辑，这种情况下先不考虑去重了。**

上传单独一个软件。

## 西瓜视频

提供好用户id, 构造url如下：  
<http://m.365yg.com/video/app/user/home/?to_user_id=95538078135&iid=29295669458&device_id=14231720144&format=html&app=video_article>

获得视频列表，其中元素如下：

<http://toutiao.com/group/6540255318931145223/?app=video_article&amp;scheme=snssdk141%3A%2F%2Fdetail%3Fgroupid%3D6540255318931145223>

点开后的最终播放地址是：

<http://www.365yg.com/a6540255318931145223>

页面中的视频地址是通过构造如下的request后得到的json反馈，

<http://ib.365yg.com/video/urls/v/1/toutiao/mp4/a519dca5dd6442839c23a73889e56395?r=08633517709597993&s=3290301971&action=GetVideoPlayInfo_YG365A&callback=axiosJsonpCallback1&_=1523763335035>

包含在<http://s0.pstatp.com/cdn/expire-1-M/byted-player-service/1.0.6/byted-player-service.min.js>

其中：

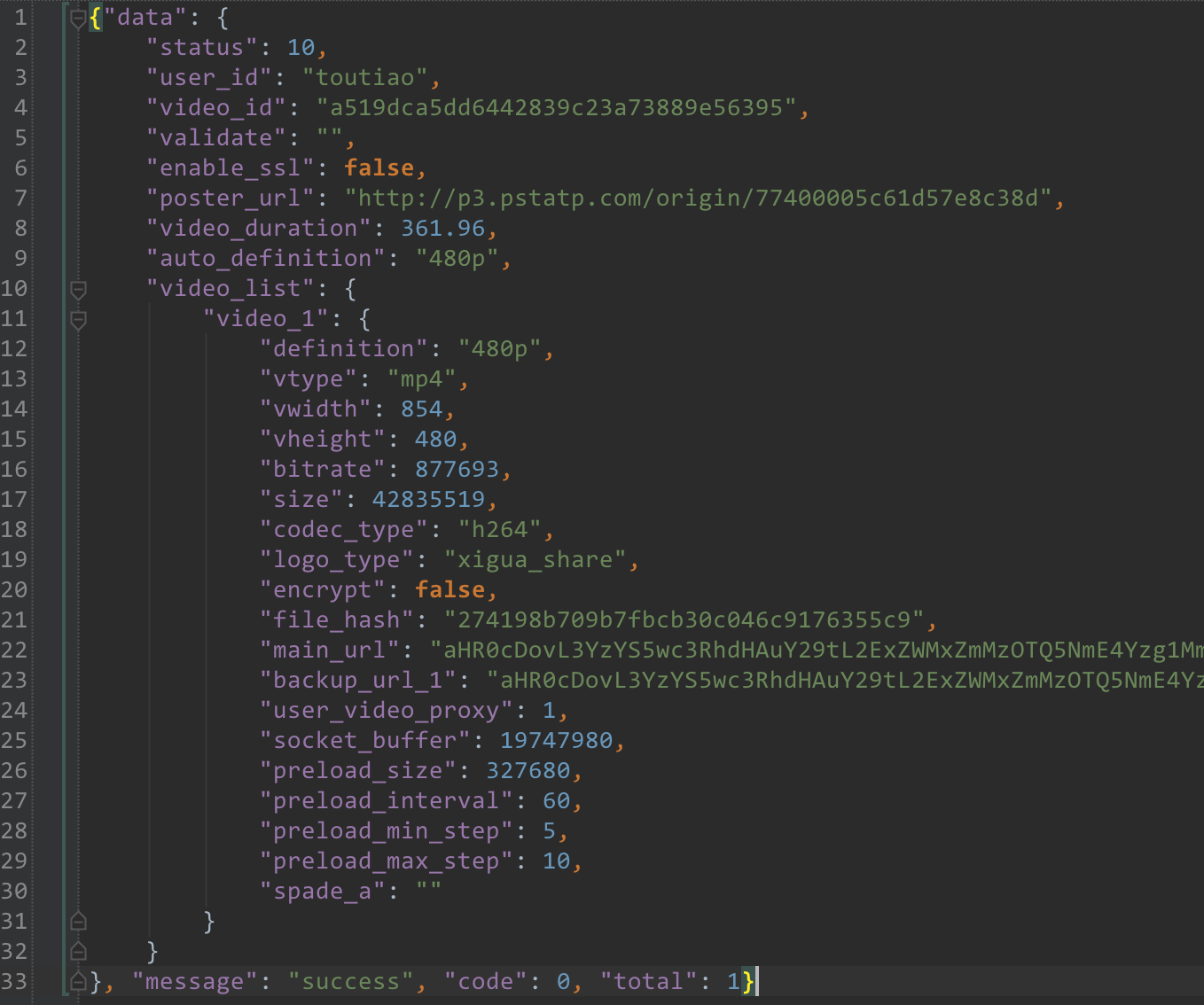
a519dca5dd6442839c23a73889e56395 为视频id,

r参数为一个随机数，Math.random().toString(10).substring(2)

s参数为这是前面部分地址的CRC32校验值无符号右移0位

其它的参数都可以省略，照样可以得到解析结果，只不过播放地址与带这些参数的有不同。

可以得到response.



其中main\_url就是视频的源地址，base64反编码以后就可以得到视频的播放地址。

Python中下载视频的方法，参见python\_learning project，其中还可以得到进度条。

# 上传部分