Summary of Trailer Data Collection v1.0

Mike Ma

Table of Contents

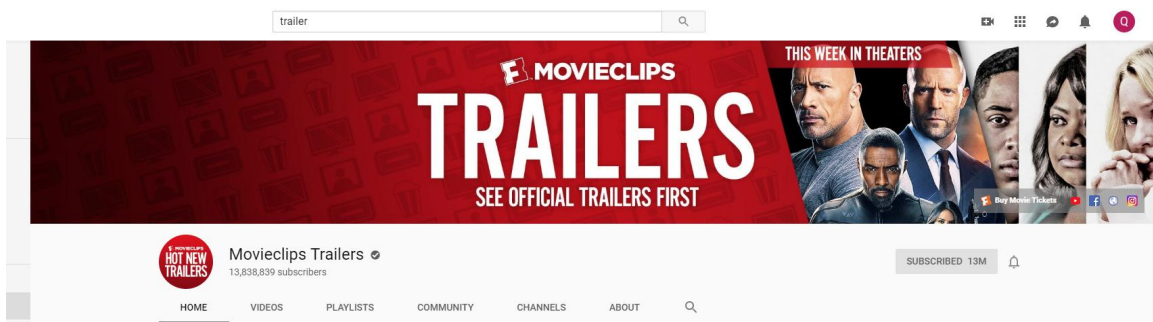# 1. Data Introduction

1.1 Data Source

- YouTube- Movieclips Trailers (8K+ official trailers & 14M+ subscribers)

https://www.youtube.com/user/movieclipsTRAILERS

- Twitter

https://twitter.com/BestTrailers(480 Tweets, 144K Follower)

The problem is—trailers from twitter is not directly shared from YouTube, twitter has a special channel for trailers, so although the film name is the same, if one film has multi trailers, the data from Twitter and YouTube is difficult to match.

Ex. The King's Man trailer:

Youtube:

However in Twitter:

Personally I don't think Twitter sharing analysis is necessary.(The multi types of data from Youtube are enough to judge viewer's preference and the popularity of a trailer)

But if necessary, I could get the number of comments, sharing, like. (I have already got them)

1.2 Data Introduction

- Data-table: youtube_info_all

(All trailers published from 2018.1.1-2019.8.1, 700+trailers)

1. **Published time**
2. **Title**
3. **Full description**
4. **Duration**
5. **ViewCount**
6. **LikeCount**
7. **DislikeCount**
8. **FavoriteCount(add this trailer as favorite video, hidden data)**
9. **commentCount(How many comments)**

Mysql database ex:

time	channelId	title	description	categoryId	duration	viewCount	likeCount	dislikeCount	favoriteCount	commentCount
2018-02-09 00:00	UCi8e0iOV	The Grinch TV Spot | 'Olympics' | (2018) | Movieclips Trailers	Check out the official The Grinch TV Spot starring Ber	1	PT30S	684500	6018	1596	0	1208
2018-02-08 00:00	UCi8e0iOV	Why Go. | Black Panther	Are you somehow still on the fence about BLACK PAN	1	PT1M8S	39782	533	191	0	207
2018-02-08 00:00	UCi8e0iOV	Now In Theaters: 50 Shades Freed, The 15:17 to Paris, Peter Rabbit | Wee	Which movie is right for you this weekend? Fifty Shad	1	PT1M5S	23892	229	32	0	24
2018-02-05 00:00	UCi8e0iOV	Skyscraper Trailer #1 | Movieclips Trailers	Check out the official Skyscraper trailer starring Dway	1	PT2M31S	291698	4130	340	0	439
2018-02-05 00:00	UCi8e0iOV	Mission: Impossible - Fallout Super Bowl TV Spot | Movieclips Trailers	Check out the official Mission: Impossible - Fallout Su	1	PT30S	89117	1069	33	0	60
2018-02-05 00:00	UCi8e0iOV	The Cloverfield Paradox Super Bowl TV Spot | Movieclips Trailers	Check out the official The Cloverfield Paradox Super E	1	PT31S	355952	3265	137	0	571
2018-02-05 00:00	UCi8e0iOV	Jurassic World: Fallen Kingdom Super Bowl Trailer | Movieclips Trailers	Check out the official Jurassic World: Fallen Kingdom	1	PT1M36S	1372947	16416	1130	0	1076
2018-02-03 00:00	UCi8e0iOV	Top New Trailers - January 2018	Here are the top 10 trailers from last month based on	1	PT23M32S	50103	590	57	0	30

- Data-table: youtube_trailer_comments_thisweektrailer

  (All comments from trailers at This Week Trailer, 8400+comments from about 10 trailers in this channel, some trailers have more than 10K comments, so it need to be selected, this data table is whole comments from 10 trailers)

1. **Title of Trailer**
2. **Video Link()**
3. **Comments**
4. **CommentsLike(how many likes this comment get)**
5. **Comments Time**

- Data-table: youtube_trailer_comments_selected_thisweektrailer

(Selected comments from trailers at This Week Trailer, 800+comments from about 10 trailers in this channel, I only collect the comments which Commentslike>5, reduce more than 9/10 valueless comments)

IN THEATERS THIS WEEK ▶ PLAY ALL

Watch trailers from movies out in theaters this week. So many awesome films, so little time.

Hobbs & Shaw Trailer #2 (2019) | Movieclips Trailers	Luce Trailer #1 (2019) | Movieclips Trailers	Once Upon a Time in Hollywood Trailer #1 (2019)...	The Lion King Trailer #1 (2019) | Movieclips Trailers	The Art of Self-Defense Trailer #1 (2019) | Movieclip...
Movieclips Trailers ✓	Movieclips Trailers ✓	Movieclips Trailers ✓	Movieclips Trailers ✓	Movieclips Trailers ✓
1.1M views • 3 months ago CC	185K views • 1 month ago	1.1M views • 4 months ago CC	1.1M views • 3 months ago CC	140K views • 1 month ago

1. **Title of Trailer**
2. **Video Link()**
3. **Comments**
4. **CommentsLike(how many likes this comment get)**
5. **Comments Time**

● Data-table: youtube_trailer_comments_selected_Action (Still Running the program)

(Selected Comment from trailors at Action Category, ,55,000+Comments from 500+trailers, ==collect the comments which Commentslike>5==)

(If necessary, I could get all comments from trailers at all different categories, but it need too much time to run, personally I think use a typical category like Action Trailers is enough, because to run NLP to analysis comments attitude also need much more time than scrapping data).

Trailers by Genre

Family & Animation Trailers Spotlight
Movieclips Trailers ⊘ Updated 2 days ago

Cats Trailer #1 (2019) | Movieclips Trailers · 2:33
The Lion King Movie Clip - Circle of Life (2019) | Movieclips Trailers · 1:01
VIEW FULL PLAYLIST (301 VIDEOS)

Horror Trailers Spotlight
Movieclips Trailers ⊘ Updated yesterday

Little Monsters International Red Band Trailer #1 (2019) | Movieclips Trailers · 2:34
The Hunt Trailer #1 (2019) | Movieclips Trailers · 2:33
VIEW FULL PLAYLIST (187 VIDEOS)

Action Trailers Spotlight
Movieclips Trailers ⊘ Updated yesterday

In Theaters Now: Fast & Furious Presents: Hobbs & Shaw | Weekend Ticket · 2:18
The Hunt Trailer #1 (2019) | Movieclips Trailers · 2:33
VIEW FULL PLAYLIST (513 VIDEOS)

Comedy Trailers Spotlight
Movieclips Trailers ⊘ Updated yesterday

Little Monsters International Red Band Trailer #1 (2019) | Movieclips Trailers · 2:34
Zombieland: Double Tap Trailer #1 (2019) | Movieclips Trailers · 2:34
VIEW FULL PLAYLIST (328 VIDEOS)

SHOW MORE

https://www.youtube.com/playlist?list=PLScC8g4bqD45-Bue4BX7U2h4IkzEV3hcL

- Downloaded File: All trailers videos at all different categories.

(All trailers downloaded and sorted as categories, 40GB, 720P, MP4, name is same as "Title" at database, about 3000 trailers)

Categories:

1. **720P Biographical Trailer Spotlight**
2. **720P Comedy Trailers Spotlight**
3. **720P Drama Trailer Spotlight**
4. **720P Family & Animation Trailers Spotlight**
5. **720P Horror Trailers Spotlight**
6. **720P Hot Docs - New Documentary Trailers**
7. **720P NEW INDIE TRAILERS**
8. **720P Thriller Trailers Spotlight**
9. **720P Action Trailers Spotlight**

Supported format:

1080P FullHD

720P HD

480P VGA

360P SD

240P HQVGA

144P Low

128P Kb/s

I don't know the video analysis need how clear the video should be, so I choose the 720P, I don't have experience of video analysis, but according to previous experience of image analysis(image classification prediction using Alexnet) , I'm sure to run the program need huge time. High quality could get higher accuracy, however, I strongly recommend 144P if the result could be accepted. I use a software to download the videos, so download 144P you just need to press a button.(in the following pages I would introduce)

Besides, save videos into database is bad idea. If we would use python to do video analysis, we just need to open the file in python(like my previous image analysis).

Personally I think video analysis is still a immature technique. I don't know many algorithms about them.
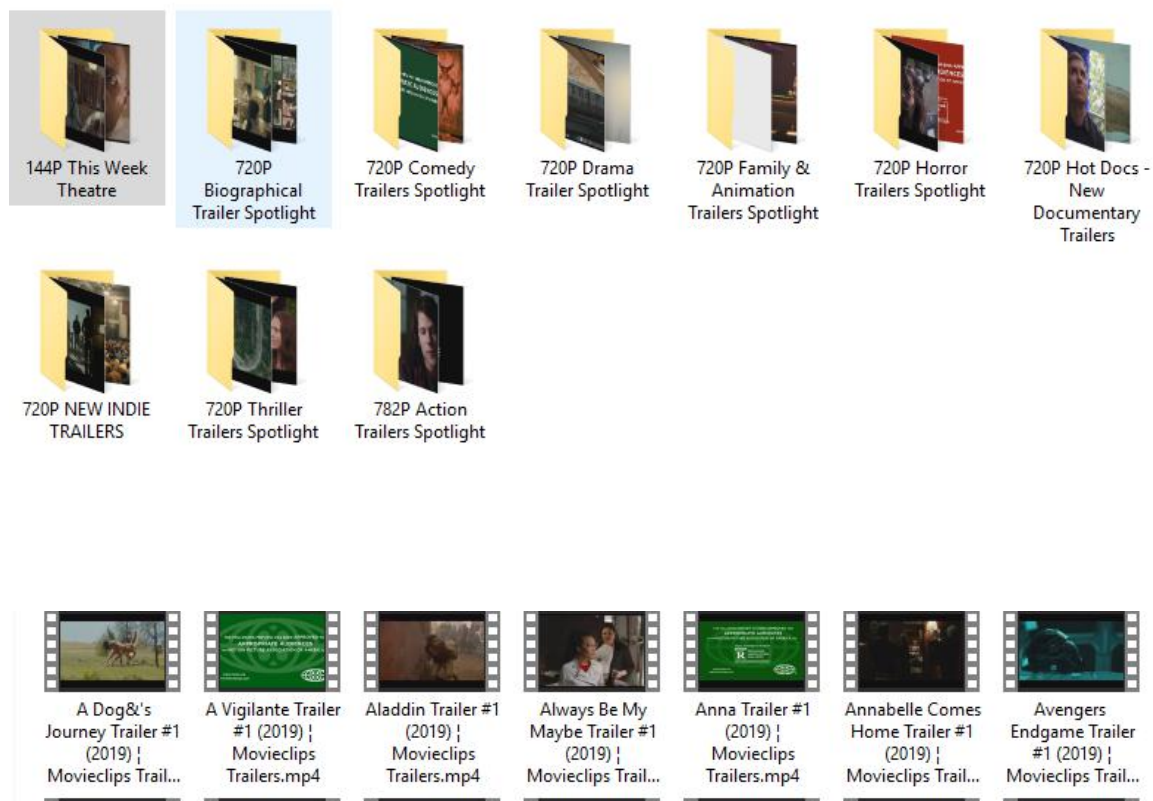
- Data-table: Twitter_trailers_info_all

(All 480 tweets including Title, URL, Comments, Tweet Time, LikeCount, ShareCount, ReplyCount, From, @who)

Title

URL

Comments

Tweet Time

LikeCount

ShareCount

ReplyCount

From

@who

Title	URL	Comments	tweet-timestamp	Like	Share	Reply	From	@who
Queen & Slim	https://twitter.com/QueenA	#QueenAndSlimpic.twitter.c	2018.8.5	391	168	3	Queen & Slim	QueenAndSlim
The Irishman	https://twitter.com/TheIrish	THE IRISHMAN pic.twitter.c	2018.8.1	692	248	5	(Null)	
Zombieland: Double Tap	https://twitter.com/Zombie	#Zombieland2 pic.twitter.co	2018.7.29	1222	501	8	Zombieland: Double Tap	Zombieland2
Gemini Man	https://twitter.com/GeminiN	Will Smith takes on himself i	2018.7.26	1009	458	16	Gemini Man	GeminiMan
A Beautiful Day in the Neigh	https://twitter.com/ABeauti	#ABeautifulDayInTheNeight	2018.7.25	598	152	4	A Beautiful Day in the Neigh	ABeautifulDayInTheNeighbc
Regal	https://twitter.com/RegalMi	The incredible true story of f	2018.7.23	1101	365	9	Regal	HARRIET
		PART 2 pic.twitter.com/8dX	2018.7.21	1814	742	7		
		ALL OF THESE WERE JUST A	2018.7.21	4042	1953	18		
Top Gun	https://twitter.com/TopGun	TOPGUN IS BACK. 2020pic.t	2018.7.19	906	393	2		
The King's Man	https://twitter.com/Kingsm	The King's Man pic.twitter.c	2018.7.17	991	431	7	(Null)	
		The Disney upcoming movi	2018.7.13	720	263	9		
Walt Disney Studios	https://twitter.com/DisneyS	Go beyond the fairytale. #Di	2018.7.9	740	402	8	#Maleficent2	Disney,Maleficent2

1.3 Personal Opinion About Data

- youtube_trailer_info_all
    - Title
    - Full description
    - Duration
    - ViewCount
    - LikeCount
    - DislikeCount
    - FavoriteCount
    - commentCount

1. There are 5 Counts to decide the popularities of trailers and viewers' preferences. These 5 Counts are not proportional, that means through analysing these number, we could know how controversial the trailer is. And Later through Comments Analysis using NLP, we could know the reason(or attitude, it's easier) of controversy.

2. Full description could use NLP algorithms to analysis the features of trailor and help decide what valuable description for a trailer should like.

3. Duration is also a feature of trailer.(Some people, like me, never see trailers shorter than 2min)

- youtube_trailer_comments_selected
    - Title of Trailer
    - Video Link()
    - Comments
    - CommentsLike(how many likes this comment get)
    - Comments Time

1. Title of trailer in this table is same to the "Title" in table youtube_trailer_info_all, that means easy to inquire the video information at youtube_trailer_info_all after comments analyse .

2. Personally I think one specific category trailer analysis(55,000+comments) is enough for writing lots of papers, but if in the future we want to get trailers at all different categories, method is the same.

3. I think only comments which commentsLike>5 are valuable. Some trailer have more than 10K comments, but not all comments are valuable.(But in the data-table I also get all comments for some videos, but get all comments really need lots of time.)

4. CommentsLike is the most direct way to decide how valuable a comment is. CommentsTime could also help decide how valuable.

5. Video Link is the URL of the video in the website.

● Downloaded Videos

720P Biographical Trailer Spotlight

720P Comedy Trailers Spotlight

720P Drama Trailer Spotlight

720P Family & Animation Trailers Spotlight

720P Horror Trailers Spotlight

720P Hot Docs - New Documentary Trailers

720P NEW INDIE TRAILERS

720P Thriller Trailers Spotlight

720P Action Trailers Spotlight

1. All trailers are downloaded as 720P into different categories. I believe different movies have different features, and each category should be analysed individually. (Horror Trailers have horror clips which are valuable, Action Trailers have Action clips which are valuable.)

2. I could download 7 different quality of trailer, I choose 720P to guarantee both the accuracy of future analysis and running speed.(It could be adjusted)

● Twitter_trailers_info_all

    Title

    URL

    Comments

    Tweet Time

    LikeCount

    ShareCount

    ReplyCount

    From

1. I don't think these data are quite necessary in fact, because the function of LikeCount, ShareCount and ReplyCount are similar to ViewCount, LikeCount, DislikeCount, CommentsCount of Youtube, however, there are only 480 official trailers compared to huge 8K+ Trailers of Youtube.
2. There are no distinguish between different trailers of the same film, as I wrote at the first page.

2. Tools and Methods

- Tools:

Mysql Database + Navicat

Python3.7+Pycharm(only youtube_info_all needed)

Ummy Video Downloader(software, to download videos of different qualities)

ScrapeStorm(software, to scrapy data from web framework)

Google API Key — YouTube Data API v3

Mongodb Database(if one day you are interested to get all data from the whole YouTube, about 500GB, using the program of "Extended Program")

- Methods:

Youtube API + ScrapeStorm(Scrape Web framework)

3. Operate Steps(Run yourself or get more data)

- youtube_trailer_info_all

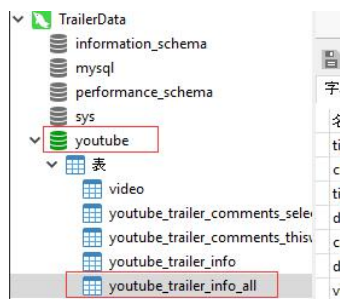 Title

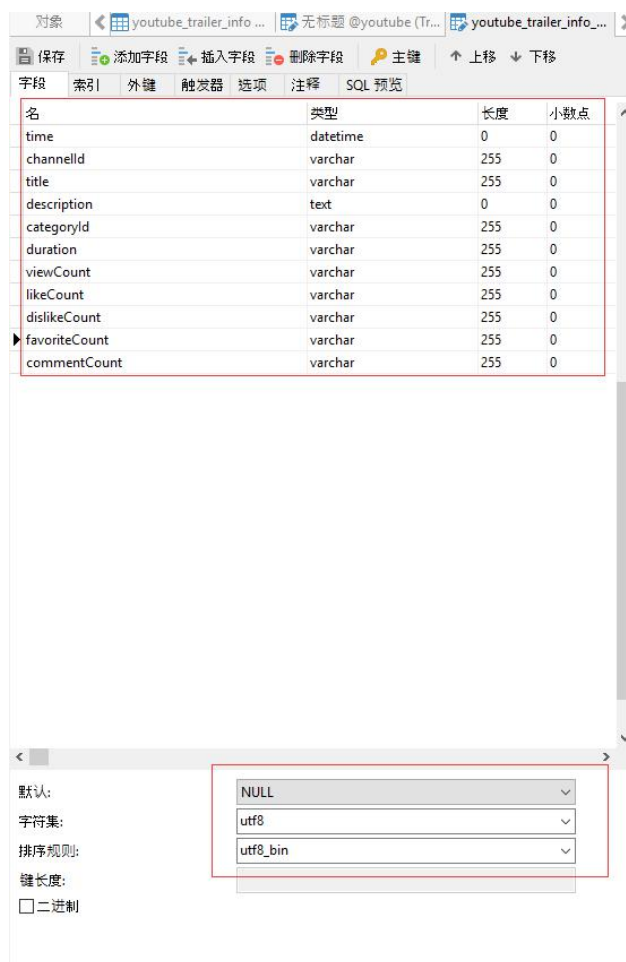 Full description

  Duration

  ViewCount

LikeCount

DislikeCount

FavoriteCount

commentCount

Steps:

1. Download Navicat, and Mysql, Build Connection in the Navicat.

2. Build a Database named youtube, build a datatable named youtube_trailer_info_all

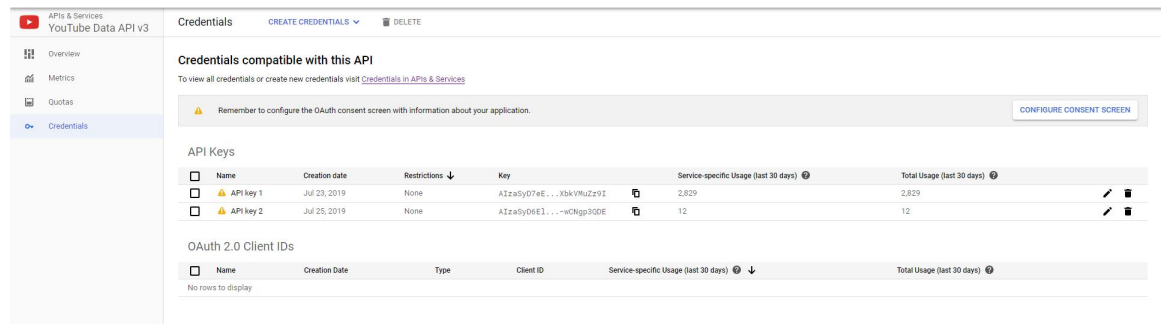

3. Set field name like this:(English version is the same)

4. Open Python Project youtube_trailerinfo_all, see the notes in code or see below.
   a) Set Configuration Parameters at settings.py
   b) Create api key at   https://console.developers.google.com/
      Choose Youtube Data API v3

   c) Set api key at youtube_scrapy_all.py   line 40
   d) Set Publish duration of data   line 46, 55, 131, 159
   e) Run

● youtube_trailer_comments_selected
      Title of Trailer
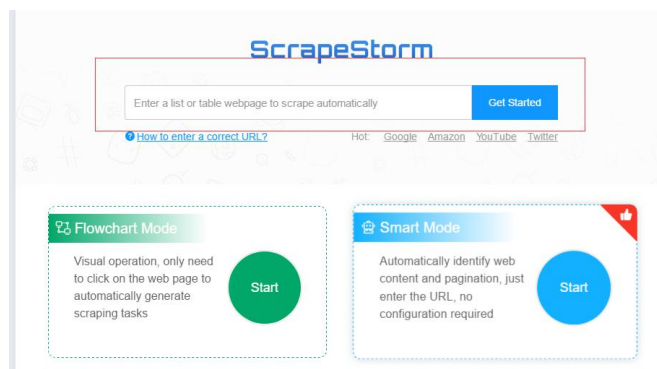   Video Link()
      Comments
      CommentsLike(how many likes this comment get)
      Comments Time

Steps: (I don't use Python because the data is at different layers and the web using rolling to show, it's difficult to achieve. However, the below method could even be used by people without coding experience, and could collect data from all kinds of websites without the risk of IP blocked)

1. Install ScrapeStorm

2. Copy the url(https://www.youtube.com/playlist?list=PLScC8g4bqD45-Bue4BX7U2h4IkzEV3hcL) and get started

3. Click Auto Detect and select the field name you want

4. Click "scrape in" in the right side of the above red box

5. Click Select in Page and click 2 comments randomly, and select the field name you want

6. Click Start and it would start scrape.(If you can't get data at first, don't worry, just wait, the loading time depends on viewers amount)

7. After getting all data, click export and choose export ways.(Usually I would choose export to Excel, and then insert data from excel to mysql. Please be aware that the format of mysql datatable shoud be utfmb4 and utf8mb4_bin, because the some comments have Emoticon)
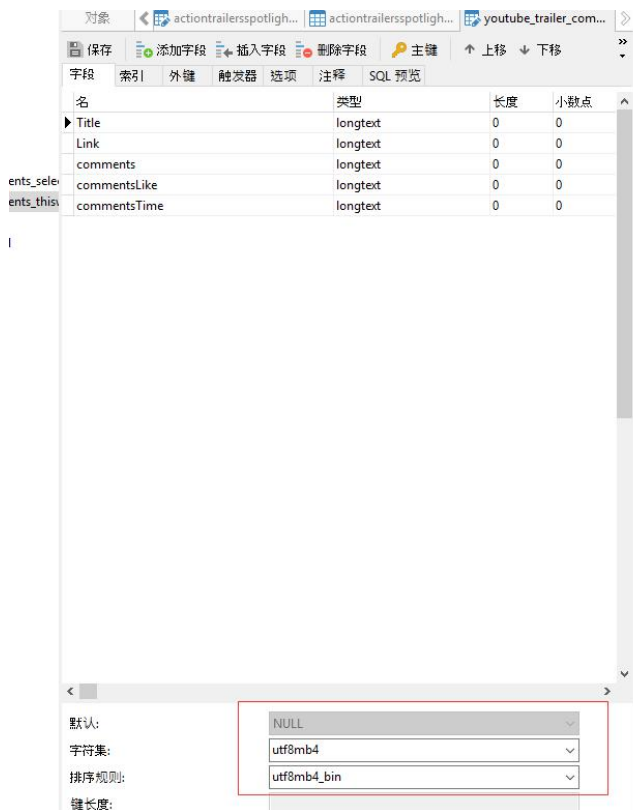
- Downloaded Videos

720P Biographical Trailer Spotlight

720P Comedy Trailers Spotlight

720P Drama Trailer Spotlight

720P Family & Animation Trailers Spotlight

720P Horror Trailers Spotlight

720P Hot Docs - New Documentary Trailers

720P NEW INDIE TRAILERS

720P Thriller Trailers Spotlight

720P Action Trailers Spotlight

Steps:

1. Install Ummy Video Downloader

2. Copy URL()
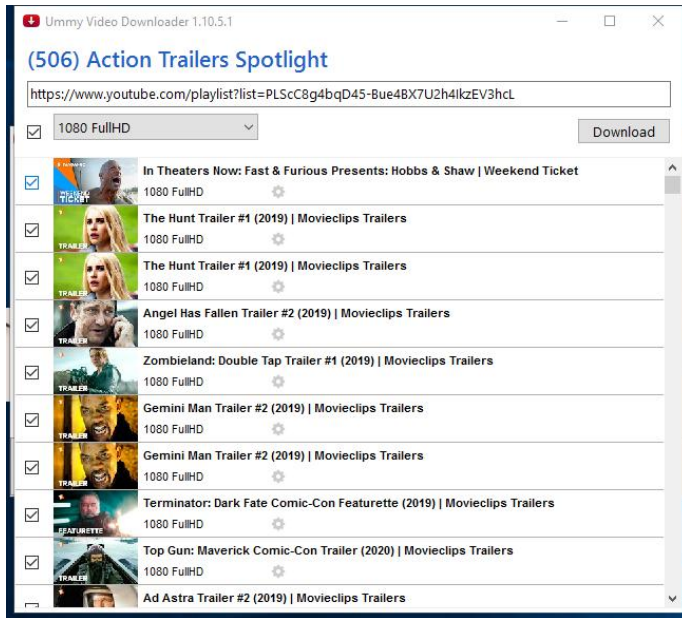
3. Choose Quality and Download.

4. Open these videos in python when you need to do analyse(File->Open->Videos)
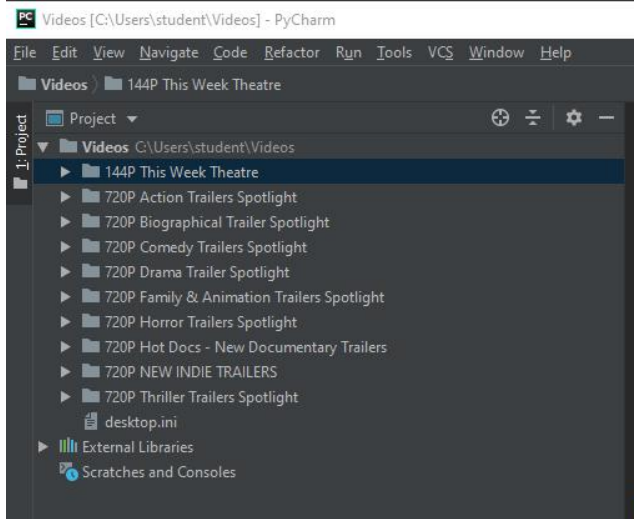
- Twitter_trailers_info_all (If Twitter is necessary, I would add this part in details)

Steps:

Same as Youtube_comments_selected, using ScrapyStorm, just use the url of Twitter.

4. Extended Program

The python project of "youtube" is a project which could get all data of all channels from youtube.(about 500 GB). This is a previous project of my friends and I.

You also need to change API key at youtubespyder.py.

You need to install Mongodb if you want to use this project. All setting would be completed automatically after you installed Mongodb. All you need to do then is wait...(500GB)

5. Summary

Four parts of my task:

1. All data from Youtube Trailer Channel √

   Published time, Title, Full description, Duration, ViewCount, LikeCount, DislikeCount, FavoriteCount, commentCount

   (2018.1.1-2019.8.1)

   700+ rows in total

2. Comments from Youtube Trailer Channel √

   Title of Trailer, Video Link, Comments, CommentsLike, Comments Time

   ("This Week Theatre" 一 All comments of about 10 trailers)   8,500+ rows

   ("This Week Theatre" 一 Selected comments of about 10 trailers)   800+ rows

   (One category "Action Trailers Spotlight" 一 Selected comments of all 500+ trailers)

   55,000+ rows

   65,000 rows in total

3. Download All Videos from Youtube Trailer Channel √

   720P Biographical Trailer Spotlight

   720P Comedy Trailers Spotlight

   720P Drama Trailer Spotlight

   720P Family & Animation Trailers Spotlight

720P Horror Trailers Spotlight

720P Hot Docs - New Documentary Trailers

720P NEW INDIE TRAILERS

720P Thriller Trailers Spotlight

720P Action Trailers Spotlight

(all 3000 trailers divided into 9 categories, 40GB, 7 kinds of quality could be chosen)

4. All data from Twitter like sharing of trailers(Don't recommend) √

(All 480 tweets including Title, URL, Comments, Tweet Time, LikeCount, ShareCount, ReplyCount, From, @who)

6. Problems and Improvements

Problems	Improvements
If one film has multi trailers, the data from Twitter and YouTube is difficult to match	Don't use data from Twitter
Limits of API quota	Use different methods to get data. Only youtube_info_all need api, and 10,000 quota is enough to get data from a whole year
The English version of ScrapeStorm is not free to export data. 95 dollars/month to export without limits, 49 dollars/month to export with 10000 rows/day The Ummy Video Downloader is not free, 5 dollars/month	However we don't need to be limited by api quota, and API could only get the most popular comments. In the short term it is worthy.