Trailer Research Reports

Mike

Directory

Technical Report **2**

Data Introduction Report **9**

Literature Review of Trailer Research **15**

Proposed Plan of Trailer Research **18**

Weekly Reports **19**

Weekly Report_0828 19

Weekly Report_0912 20

Weekly Report_0922 24

Weekly Report_1002 26

Summary of Trailer Data Collection **28**

Technical Report

Mike Ma[1 *]

[1]SIRC. Ontario Tech University, ON, CA

We collected raw data of all movie information(81270 movies) from IMDb during 2013-2019; all trailer information(8616 trailers) from 9 channels of YouTube during 2013-2019; trailer reviews(864551 reviews) from 10 movie categories of the biggest YouTube channel "Movieclips Trailers"; as well as 3000 trailer videos from 10 movie categories of the biggest YouTube channel "Movieclips Trailers". For the research purpose of predicting movie gross income and popularity by trailer information, we matched the above all trailer information and all corresponding movies together, in order to match meta-data of trailer and movie. We used sentiment analysis tool TextBlob to calculate the sentiment and subjectivity of each trailer comment. By multiplying 'sentiment' and 'commentlike' count, we got 'preference' of each trailer. We also put forward a data normalization method to solve the time inconsistency problem. By co-efficiency analysis, We finally found that the trailer comments information has a much better performance and reliability than simple intuitive data 'like/dislike' to predict the movie gross income.

Motivation

The success or failure of a movie is often determined in its first weekend of play. In order to make this opening successful, movie producers must employ a number of promotional strategies including movie trailers, to publicize the movie for a significant period of time prior to its release. A movie trailer, as described in Wikipedia, is an advertisement or a commercial for a feature film that will be exhibited in the future. Of some ten billion videos watched online annually, film trailers rank the third, after news and user-created video. It is clear that businesses have a strong interest in tapping into this huge data source to extract information that might improve their decision making process. For example, predictive models derived from movie and its trailers may facilitate filmmakers making more profitable decisions.

In our research, by analyzing the correlation between the meta-data of trailer and movie, we aim to derive a predictive model from movie and its trailer meta-data, and focus on the prediction of movie gross income and movie popularity using trailer information as an effective pre-released meta-data.

RELATED WORK

For details, see 'Literature Review of Trailer Research'

Predicting the Gross income(including the stock price) of movie.

Research has been done to generate models for predicting revenues of movies. Most of them derived results from single data sources. Specifically, Joshi and others[1] use linear regression that joined meta-data with text features from pre-release critique to predict earnings for movies with a coefficient of determination R^2=0.671.

Mishne and Glance[2] correlate sentiments in blog posts with movie box-office scores. The correlations they observed for positive sentiments are fairly low and not sufficient to use for predictive purposes. They neither build prediction models or show the value of the correlation because they think the result is not good enough for accurate modeling.

Zhang and Skiena[4] have used a news aggregation model along with IMDb data to predict movie box-office numbers.

In a very interesting approach Asur and Huberman set up a prediction system for the revenue of movies based on the volume of Twitter mentions[5]. They build a linear regression model based on the chatter of Twitter and achieve an adjusted coefficient of determination of 0.97 on the night before the movie release for the first weekend revenue of a sample of 24 movies. In addition, they even tried to predict the Hollywood Stock Price given that social media can accurately predict box office results and the Hollyhood Stock Exchange adjusts the price for a movie stock to reflect the actual box office gross. They tested social media data efficacy compared with historical HSX prices at forecasting the stock prices of the HSX index and their tweet-rate proves to be significantly better at predicting the actual stock value than the historical HSX prices. It's a good inspiration of considering stock price because according to [6], Prices of movie stocks accurately predict real box office results, which demonstrates the strong correlation between the movie stock price and real box office gross income.

In a later work, however, Wong et al. show that Tweets do not necessarily represent the financial success of movies[7]. They consider a sample of 34 movies and compare the Tweets about the movies to evaluations written by users of the movie review websites. They argue that predictions based on social media could have high precision but low recall.

Predicting the rating(popularity) of movie.

Sharda and Delen[3] have treated the prediction problem as a classification problem and used neural networks to process pre-release data, such as quality and popularity variables, and classify movies into nine categories ranging from 'flop' to 'blockbuster'. Apart from the fact that they are predicting ranges over actual numbers, the best accuracy that their model can achieve is fairly low(36.9%).

Marton and Taha[8] have showed that the populartity of a movie can be predicted much before its release by measuring and analyzing the activity level of editors and viewers of the corresponding entry to the movie in Wikipedia. It's novel because it is the only research using data from wikipedia, but the data features are too simple and low to support their conclusion.

In a rather novel approach, Oghina et al. have made use of Twitter and YouTube activity streams to predict the ratings in the Internet Movie Database(IMDb), which is among the most popular online movie database[9].

Reference[10] described a movie rating approach based on data mining of 240 movies from IMDb where Weka and J48 were used to create the prototype model.

Reference[11] also applied Weka and J48 to generate three classes of movies: Hit, Neutral or Flop, in order to predict the movie box office performance. It's amazing that they also use trailer information and their data is quite similar to us and give us inspirations of more possible data features. They generate a model consisting of genre of successful movies ranked by user ratings of the IMDb, popularity of director, leading actor and leading actress represented by the number of comments and views of official movie trailers accessible by YouTube, and sentiment toward a movie derived from YouTube viewers' comments. However, their conclusion is not reliable considering their only 35 movies data.

In [12], methods to predict the popularity of movies were discussed to evolve as a guiding strategy for Content Distribution/ Delivery Network(CDD). Actor and director popularity were considered as base criteria for predicting the popularity of a movie.

Predicting the popularity of trailer.

The only research focusing on trailers is [14], they analyzed whether subjective multimedia features be developed to predict the viewer's preference presented by like or dislike during and after watching movie trailers. The results showed that the single low-level multimedia feature of shot length variance is highly predictive of a viewer's "like/dislike" for a large portion of movie trailers. However, their features are too narrow and data is from only 1375 trailers. There are still lots of features to demonstrate the popularity of trailers. Moreover, there has been substantial interest in the NLP community on using movie reviews as a domain to test sentiment analysis methods. e.g.,[13], et al. Basically speaking, they apply information retrieval or machine learning techniques to classify movie reviews into some categories and hope to produce better classification accuracy than human being. The classification categories are like "thumbs up" vs."thumbs down", "positive"vs."negative", or "like" vs."dislike".

Video shots analysis.

In the video shots analysis part, researches could be classified as two categories:
1. Relationship between Mediaand Audiences' Affecting State:
Related research focusedon the emotion recognition of videos or movies. In these articles, some electronic signals, such as electroencephalogram(EGG), facial images, color features, the combination of audio and visual features and so on, are employed as the fundamental data for registering the viewer's affecting state.
2. Visual Data Feature Extraction:
Shot segmentation and key-frame were extracted and the lighting key cues, motion,shot density, color energy cues, and other miscellaneous cues, including some audio data, were used as the features to predict the affecting potential of a film.[14]

Brief analysis of related works.

1. Most of the works are focusing on movie gross prediction. Different people work on movie gross prediction from different perspectives. Most previous work forecast movie grosses based on IMDb data with regression or stochastic models. However, their models either work poorly or need post-release data in order to make reasonable prediction, which are not acceptable in practice, because it is difficult to give shape estimation for either model parameters or gross if they don't have any early stage movie gross data. Although the post-release models are also useful in some situations, pre-release models are of more practical importance. Luckily, trailer is a good pre-release data source but seldom considered in related works.

2. While there has been research on predicting movie sales, almost all of them have used meta-data information on the movies themselves to perform the forecasting, such as the movies genre, MPAA rating, running time, release data, the number of screens on which the movie debuted, and the presence of particular actors or actresses in the cast. Trailer is seldom considered in related works.

3. Since predictions based on classic quality factors fail to reach a level of accuracy high enough for practical application usage of user-generated data to predict the success of a movie becomes a very temping approach. It indicates that sentiments analysis from reviewers' comments are worth to do.

4. Most predictions work on using movie data to predict movie performance. Seldom predictions work on using trailer data to predict movie performance. Only one prediction works on using trailer data to predict trailer performance, while their features are too simple(only like/dislike) and data size are too small.

5. Most predictions work on movies focuses on forecasting revenues, not ratings. Seldom predictions work on movies focuses on ratings. Only two predictions work on trailers.

However, trailers are good pre-release data but ignored by many.

6. Predicting movie stock price is worth to try because the strong correlation of movie performance and movie stock price. Most papers ignored this point.

7. Video shots analysis is not practical and not necessary in current stage considering its complexity and lots of simpler problems need to focus on. What also well worth mentioning is that almost all papers about video shots analysis get no more than 30 citations.

Analysis and Inspiration.

1. Improving Movie Gross Prediction Through News Analysis[4]: Their sentiment statistics are good. They derive several sentiment measures, including polarity, subjectivity, positive references per reference, and positive-negative differences per reference and give their definitions.

2. Early Prediction of Movie Box office Success Based on Wikipedia Activity Big Data[8]: They use features both individually and combined to repeat experiments, such as T,V,S,V,T,S,V,T,V,S. And they indicate that applicability of prediction model on movies with medium and low popularity levels remains an open question.

3. Predicting IMDB movie ratings using social media[9]: Their data is similar to us, including surface features from platform Youtube and Tweeter, as well as textual features from tweets as well as Youtube comments.

4. The real power of artificial markets[6]: It indicates that the prices of movie stocks could accurately predict real box office results.

5. Predicting movie sales from blogger sentiment[2]: They analyze both pre-release data and post-release data with good sentiment analysis methods. In addition, they focus on the opening weekend data rather than total sales since it normalizes the figure across movies that were released on different dates, preventing earlier movies from having a higher income just because they have been "out there" longer.

6. Prediction of Movies Box Office Performance Using Social Media[11]: Their data are almost the same to us. It is the only paper use both movie data and trailer data to predict the performance of movie. Their provides inspiration in new data, including genre frequency, the popularity(followers) of actor and director. They also differentiate the movie by if it is sequel movie or the first-version movie like us. In addition, their NLP methods toward Youtube comments are also worth to copy.

7. Predicting Movie Trailer Viewer's Like/Dislike via Learned Shot Editing Patterns[14]: It is the only paper use trailer data to predict trailer data, their experiment process and charts are worth to copy.

8. Predicting the Future With Social Media[9]: The whole paper is good and citations are above 8000. They analysis the relationship between tweets-rate, sentiments of tweets and movie outcomes as well as movie stock price. Their data is from Tweet. Their method could be repeated with our Youtube comments-rate and sentiments of Youtube trailers comments, and we could expect a good outcome.

DATASET CHARACTERISTICS

Raw Data.

We collected raw data of all movie information(81270 movies) from IMDb during 2013-2019; all trailer information(8616 trailers) from 9 channels of YouTube during 2013-2019; trailer reviews(864551 reviews) from 10 movie categories of the biggest youtube channel "Movieclips Trailers"; as well as 3000 trailer videos from 10 movie categories of the biggest youtube channel "Movieclips Trailers".

Matched Data.

For the research purpose of predicting movie gross income and popularity by trailer information, we matched the above all trailer information and all corresponding movies together, in order to use trailer information as pre-released meta-data to predict the movie gross income, rating, and popularity. We got merged data table consists of both trailer info and movie info.

Sentiment Preference.

We used sentiment analysis tool TextBlob to calculate the sentiment and subjectivity of each trailer comment. By multiplying 'sentiment' and 'commentlike' count, we got 'preference' of each trailer. By normalizing the comment time and movie publish time, we saw the reliability of 'preference' compared to non-normalized simple 'like/dislike'.

Coefficient analysis.

By coefficient analysis, we tested the correlation between 'movie rating' and meta-data of trailer, correlation between 'movie USA gross income' and meta-data of trailer, as well as 'movie Worldwide gross income' and meta-data of trailer.

For dataset details, see 'Data introduction report'.

TITLE MATCHING of TRAILERS MOVIES

After we collected the raw data of trailer info from YouTube and movie info from IMDb, it's necessary to solve the title matching problem, and merge the trailer info and movie info into one table.
The title matching rate is about 70%, which means about 70% trailer could match the correspondent movie, based on the dataset 8616 trailers 81270 movies.

Method of title matching.

1. Export the trailer data table and movie data table from

MySQL into Json.
2. Create list for each field.
3. Use re.match() to match.
4. Check the format of unmatched part, if it included subtitle or series number, the matched part would not be considered.
5. store all the fields into one database.

TITLE REVIEWS SENTIMENT ANALYSIS

Textblob Introduction.

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Sentiment Analysis Experiment.

TRAILER REVIEWS DATA NORMALIZATION

Major problem.

The only paper focusing on trailer prediction is "Predicting movie trailer viewer's "like/dislike" via learned shot editing patterns", However, they directly use "like/dislike" in YouTube to do the prediction, which has a problem. Think about this:

Movie 1 is released 3 days ago, whose trailer has 400 likes.
Movie 2 is released 3 years ago, whose trailer has 4000 likes.

However, we couldn't say movie 2 is better than movie 1, while in their paper they ignored this problem.
To solve the time problem(described in technical report) and normalization problem, we use sentiment analysis to calculate the preference from each trailer review to compare the performance and reliability of simple 'like dislike' for prediction in previous papers.

Method Introduction - Reviews Data Normalization.

1. Get "movie release time" and "trailer comments time" from the title-matched table and comments table.

2. Transfer "trailer comments time" ('1 days ago' transfered to '2 days until the data collecting time,2018-08-01').

0-24 hours ago — 2019-07-30 (1 days)
1 days ago — 2019-07-29 (2 days)
1 weeks ago - 2019-07-18(13 days)
1 months ago - 2019-06-02(59 days)
1 years ago - 2017-08-02(729 days)

3. Compare "movie release time" and "trailer comments time" to current data-collecting time(2019-08-01) individually. Calculate the differences:

Difference 1 = 2019-08-01 - "movie release time"
Difference 2 = 2019-08-01 - "trailer comment time"

4. Prerequisites for comments selection
(1)Guarantee the review of trailer before the movie released(guarantee the trailer reviews are pre-released data).

Difference 2 - Difference 1>0

(2)Set the comments selection period as 6 months(180 days) before the movie released time.

0 <Difference 2 - Difference 1 <180

(3)Delete trailer info which is released less than 6 months before data collecting time(2019-8-1).

Difference 2 >180

5. Comments selection
Using the value of "difference 2 - difference 1", we could normalize each Trailer review in a fixed period of time(6 months)

CORRELATION ANALYSIS

In this experiment, we calculated the co-efficiency between the movie rating and meta-data of trailer, co-efficiency between the movie USA gross income and meta-data of trailer, co-efficiency between the movie worldwide gross income and meta-data of trailer.

The meta-data of trailer includes 'trailer series', 'view count', 'like count', 'dislike count', 'comment count'.

Correlation Results.

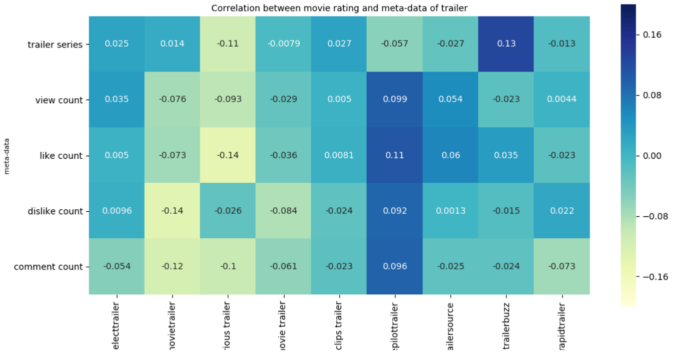


Fig. 1. Correlation between movie rating and meta-data of trailer

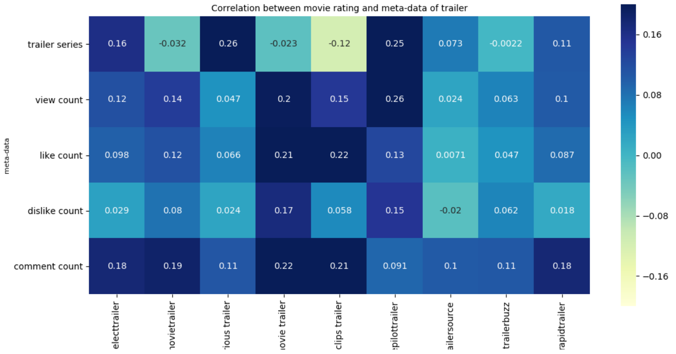


Fig. 2. Correlation between movie USA gross income and meta-data of trailer

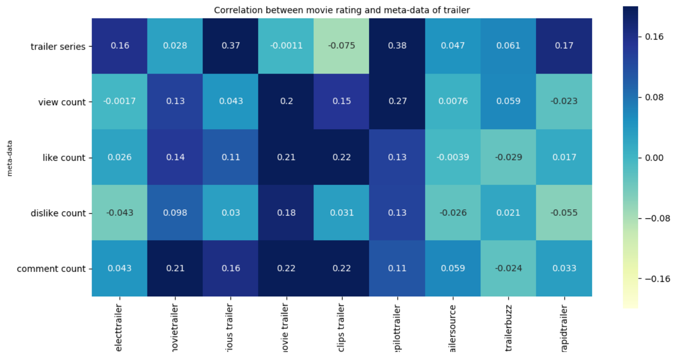


Fig. 3. Correlation between movie USA gross income and meta-data of trailer

Correlation Analysis.

Be aware that each channel has different trailer amount and subscribers, which means we should not compare the co-efficiency by columns. Rather, the below analysis based on the comparison between the rows.

According to the visualization above, we could not see any Characteristics from Fig.1. In Fig.2. and Fig.3., the color in the last row is much darker than the above 4 rows, which means "trailer comments count" has a much higher co-efficiency (more than 10 times) with "movie gross income" than simple "view count", "like count", and "dislike count".

From this result we could draw the conclusion that The comments preference value from trailer comments may be much better than simple "view count", "like count", "dislike count" to predict the movie gross income. And with the method to normalize the reviews time period suggested above, the comments preference value is much more reliable than simple meta-data.

By way of coefficient analysis, we preliminary found the results as above. From the coefficient analysis result above, we could draw the conclusion that 'comments count' has a much higher correlation with USA gross income and world-wide gross income than simple meta-data like 'view count', 'like count', and 'dislike count'.(about 10 times higher). Therefore we believe that the preference from trailer reviews have much higher performance than simple meta-data of trailer in YouTube. We would prove the assumption in further co-efficiency analysis.

CURRENT CONCLUTION

By coefficient analysis, "trailer comments count" has much higher coefficiency(more than 10 times) with "movie gross income" than simple "view count","like count", and "dislike count". according to our method to calculate the trailer comments 'preference' with Textblob, the comments preference value of trailer may have much better performance than simple "view count", "like count" "dislike count" to predict the movie gross income, and possibly the movie stock price. And with the method to normalize the reviews time period, comments preference value is also much more reliable than simple meta-data. We make up for the time problem of data in the past literature.

In our further research, we aim to derive a predictive model from movie and its trailer meta-data, and focus on the prediction of movie gross income using trailer comments information. Test the conclusion from our correlation analysis.

References

1. M. Joshi, D. Das, K. Grimpel, and N. A.Smith, "Movie revies and revenues: An experiment in text regression." *NAACL-HLT*, 2010.
2. G.Mishne and N.Glance, "Predicting movie sales from blogger sentiment." in *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
3. R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks." in *Expert Systems with Applications*, vol. 30, 2006, pp. 243–254.
4. W.Zhang and S.Skiena, "Improving movie gross prediction through news analysis." in *In Web Intelligence*, 2009, p. 301304.
5. H. B. Asur S, "Predicting the future with social media." in *In Processings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010, pp. 492–499.
6. D. M. Pennock and S. Lawrence, "The real power of artificial markets." in *Science: American Association for the Advancement of Science.*
7. W. FMF, S. S, and C. M, "Why watching movie tweets won't tell the whole story?" in *In Processings of the 2012 ACM workshop on Workshop on online social networks. New York, NY, USA: ACM, WOSN'12*, 2012, pp. 61–66.
8. M. Mestyan and T. Yasseri, "Early prediction of movie box office success based on wikipedia activity big data," in *PLoS ONE 8(8): e71226. doi:10.1371/journal.pone.0071226*, 2013.
9. O. A, B. M, T. E, and de Rijke M, "Predicting imdb movie ratings using social media." in *In: ECIR 2012: 34th European Conference on Information Retrieval. Springer-Verlag, Barcelona, Spain: Springer-Verlag*, 2012, pp. 503–507.
10. S. Kabinsingha, S. Chindasorn, and C. Chantrapornchai, "Movie rating approach and application based on data mining," in *International Journal of Engineering and Innovative Technology(IJEIT)*, 2012.
11. K. R. Apala and M. Jose, "Prediction of movies box office performance using social media," in *2013 IEEE/ACM International Conference on Adances in Social Networks Analysis and Mining*, 20123,.
12. L. Zhu, M. Zhu, and S. Yao, "The popularity of movies predict system based on data mining technology for cdn." in *IEEE International Conference on the 3rd Computer Science and Information Technology*, 2010.
13. P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches." in *In Proceedings of the Hawaii International Conference on System Sciences(HICSS)*, 2005.
14. Y. Hou, T. Xiao, and S. Zhang, "Predicting movie trailer viewer's like/dislike via learned shot editing patterns," in *IEEE Transactions on Affective Computing*, no. 1, 2016.

Table 1. Correlation between movie rating and meta-data of trailer

Rating correlation	trailer series	view count	like count	dislike count	comment count	sentiment	preference
filmselecttrailer	0.02533283	0.03526385	0.0050197	0.00958571	-0.053747		
freshmovietrailer	-0.01440181	-0.07580638	-0.07292159	-0.13561284	-0.11801474		
furious trailer	-0.10815836	-0.093043	-0.14001656	-0.02585848	-0.10327457		
joblomovietrailer	-0.00786402	-0.0292394	-0.03607939	-0.08427883	-0.06071927		
movieclips trailer	0.02700788	0.00504539	0.00806546	-0.02381409	-0.02317481		
moviepilottrailer	-0.05652897	0.09932129	0.1084559	0.09204856	0.09605606		
movietrailersource	-0.0271737	0.05370674	0.06031079	0.00127742	-0.02493521		
newtrailerbuzz	0.12772609	-0.022946	0.03457465	-0.01498928	-0.02350908		
rapidtrailer	-0.01301163	0.00444308	-0.0230677	0.02161095	-0.07269746		

Table 2. Correlation between movie USA gross income and meta-data of trailer

USA Gross Income	trailer series	view count	like count	dislike count	comment count	sentiment	preference
filmselecttrailer	0.16238215	0.11563359	0.09771129	0.02928777	0.17571305		
freshmovietrailer	-0.03247056	0.14004506	0.11536722	0.07975358	0.1853188		
furious trailer	0.2574477	0.04731082	0.06555064	0.0238418	0.10809361		
joblomovietrailer	-0.02304071	0.19566287	0.20924704	0.17024959	0.22012332		
movieclips trailer	-0.11687523	0.14675415	0.21936088	0.05769043	0.2123239		
moviepilottrailer	0.24708663	0.2620017	0.12796143	0.14920888	0.09145186		
movietrailersource	0.07302531	0.0243072	0.00707041	-0.02030733	0.10357889		
newtrailerbuzz	-0.00220898	0.06282479	0.0468655	0.0620744	0.10764305		
rapidtrailer	0.10672939	0.10335676	0.08667873	0.01767974	0.17640523		

Table 3. Correlation between Worldwide USA gross income and meta-data of trailer

Worldwide Gross	trailer series	view count	like count	dislike count	comment count	sentiment	preference
filmselecttrailer	0.1573731	-0.00174282	0.02645849	-0.04269658	0.0427057		
freshmovietrailer	0.02815839	0.13289882	0.14492246	0.0978303	0.21468858		
furious trailer	0.37442343	0.04289278	0.10565868	0.0302329	0.16334648		
joblomovietrailer	-0.0011454	0.20321018	0.21359551	0.18066378	0.22128298		
movieclips trailer	-0.07532938	0.14924679	0.22110975	0.03148821	0.22409317		
moviepilottrailer	0.37532633	0.26685664	0.13342402	0.13186306	0.11057562		
movietrailersource	0.04660466	0.00761057	-0.00388638	-0.02634109	0.05929984		
newtrailerbuzz	0.06052892	0.0593351	-0.02875914	0.0214751	-0.02366493		
rapidtrailer	0.17006521	-0.02297426	0.01676616	-0.05546407	0.03273731		

Table 4. Subscriber and data amount of Rating experiment channels in YouTube

Rating Analysis	Data amount	Subscriber
filmselecttrailer	225	3.08M
freshmovietrailer	233	4.16M
furious trailer	183	3.23M
joblomovietrailer	394	2.08M
movieclips trailer	1092	14M
moviepilottrailer	266	818K
movietrailersource	232	1.3M
newtrailerbuzz	168	1.14M
rapidtrailer	253	794K

Table 5. Subscriber and data amount of USA gross income experiment channels in YouTube

USA gross income Analysis	Data amount	Subscriber
filmselecttrailer	66	3.08M
freshmovietrailer	139	4.16M
furious trailer	87	3.23M
joblomovietrailer	233	2.08M
movieclips trailer	952	14M
moviepilottrailer	116	818K
movietrailersource	97	1.3M
newtrailerbuzz	85	1.14M
rapidtrailer	57	794K

Table 6. Subscriber and data amount of Worldwide gross income experiment channels in YouTube

Worldwide gross income Analysis	Data amount	Subscriber
filmselecttrailer	225	3.08M
freshmovietrailer	233	4.16M
furious trailer	183	3.23M
joblomovietrailer	394	2.08M
movieclips trailer	1092	14M
moviepilottrailer	266	818K
movietrailersource	232	1.3M
newtrailerbuzz	168	1.14M
rapidtrailer	253	794K

Data Introduction Report

Mike Ma[1] *

[1]SIRC. Ontario Tech University, ON, CA

A specific introduction about the research data. Data in youtube.sql is 4163071 in total, includes movie trailer information, movie information, movie professional pre-released reviews, trailer reviews, as well as trailer videos. The data includes original data, cleaned data, and results of reviews sentiment analysis(using textblob) and coefficient analysis.

This report only provides the data description. For research purpose, experiment method and result analysis, please read the technical report.

Brief Introduction.

Movie trailers information is meta-data of all trailers(8616 trailers) from 2013 to 2019 in 9 channels in YouTube. Movie information is meta-data of all movies(81270 movies) from 2013 to 2019 in IMDb. Movie professional pre-released reviews(257 reviews in the first page) are from New York Times. Trailer reviews are all reviews for trailers(864551 reviews) in 10 movie categories from Youtube Movieclips Trailers. By sentiment analysis, I get preference value for all trailers in 10 movie categories from Youtube Movieclips Trailers. By coefficient analysis, "trailer comments count" is has much higher coefficiency(more than 10 times) with "movie gross income" than simple "view count","like count", and "dislike count". Which means the comments preference value from trailer comments may be much better than simple "view count", "like count" "dislike count" to predict the movie gross income. And with the method to nomalize the reviews time period suggested in Technical Report, comments preference value is definitely also much more reliable than simple meta-data.

Group your tables in mysql.

After running youtube.sql file, please create groups and move tables into groups as below. This introduction is based on below format.

749_movie_info_imdb_by_year_original

imdb_2013movielist_info

imdb_2014movielist_info

imdb_2015movielist_info

imdb_2016movielist_info

imdb_2017movielist_info

imdb_2018movielist_info

imdb_2019movielist_info

749_trailer_movie_merged_info

new_filmselecttrailer_movie_trailer_merged_info

new_freshmovietrailer_movie_trailer_merged_info

new_furioustrailer_movie_trailer_merged_info

new_joblomovietrailer_movie_trailer_merged_info

new_movie_trailer_merged_info

new_moviepilottrailer_movie_trailer_merged_info

new_movietrailersource_movie_trailer_merged_info

new_newtrailerbuzz_movie_trailer_merged_info

new_rapidtrailer_movie_trailer_merged_info

all_movie_info_imdb_by_year_original

2013_all_imdb_movielist_info

2014_all_imdb_movielist_info

2015_all_imdb_movielist_info

2016_all_imdb_movielist_info

2017_all_imdb_movielist_info

2018_all_imdb_movielist_info

2019_all_imdb_movielist_info

all_trailer_info_original

youtube_filmselecttrailer_info_all

youtube_freshmovietrailer_info_all

youtube_furioustrailer_info_all

youtube_joblomovietrailer_info_all

youtube_moviepilottrailer_info_all

youtube_movietrailerssource_info_all

youtube_new_freshmovietrailer_info

youtube_newtrailerbuzz_info_all

youtube_rapidtrailer_info_all

youtube_trailer_info_all

all_trailer_info_original_for_merge

filmselecttrailer_movie_trailer_merged_info

freshmovietrailer_movie_trailer_merged_info

furioustrailer_movie_trailer_merged_info

joblomovietrailer_movie_trailer_merged_info

movie_trailer_merged_info

moviepilottrailer_movie_trailer_merged_info

movietrailersource_movie_trailer_merged_info

newtrailerbuzz_movie_trailer_merged_info

rapidtrailer_movie_trailer_merged_info

all_trailer_movie_merged

all_new_filmselecttrailer_movie_trailer_merged_info

all_new_freshmovietrailer_movie_trailer_merged_info

all_new_furioustrailer_movie_trailer_merged_info

all_new_joblomovietrailer_movie_trailer_merged_info

all_new_movie_trailer_merged_info

all_new_moviepilottrailer_movie_trailer_merged_info

all_new_movietrailersource_movie_trailer_merged_info

all_new_newtrailerbuzz_movie_trailer_merged_info

all_new_rapidtrailer_movie_trailer_merged_info

forCoefficientAnalysis_rating_cleaned

rating_new_filmselecttrailer_movie_trailer_merged_info

rating_new_freshmovietrailer_movie_trailer_merged_info

rating_new_furioustrailer_movie_trailer_merged_info

rating_new_joblomovietrailer_movie_trailer_merged_info

rating_new_movie_trailer_merged_info rat-

ing_new_moviepilottrailer_movie_trailer_merged_info
- rating_new_movietrailersource_movie_trailer_merged_info
- rating_new_newtrailerbuzz_movie_trailer_merged_info
- rating_new_rapidtrailer_movie_trailer_merged_info

forCoefficientAnalysis_usa_cleaned
- usa_new_filmselecttrailer_movie_trailer_merged_info
- usa_new_freshmovietrailer_movie_trailer_merged_info
- usa_new_furioustrailer_movie_trailer_merged_info
- usa_new_joblomovietrailer_movie_trailer_merged_info
- usa_new_movie_trailer_merged_info
- usa_new_moviepilottrailer_movie_trailer_merged_info
- usa_new_movietrailersource_movie_trailer_merged_info
- usa_new_newtrailerbuzz_movie_trailer_merged_info
- usa_new_rapidtrailer_movie_trailer_merged_info

forCoefficientAnalysis_worldgross_cleaned
- world_new_filmselecttrailer_movie_trailer_merged_info
- world_new_freshmovietrailer_movie_trailer_merged_info
- world_new_furioustrailer_movie_trailer_merged_info
- world_new_joblomovietrailer_movie_trailer_merged_info
- world_new_movie_trailer_merged_info
- world_new_moviepilottrailer_movie_trailer_merged_info
- world_new_movietrailersource_movie_trailer_merged_info
- world_new_newtrailerbuzz_movie_trailer_merged_info
- world_new_rapidtrailer_movie_trailer_merged_info

movie_merged_info
- 749_imdb_merged_movielist_info
- all_imdb_merged_movielist_info

trailer_reviews_original
- youtube_comments_biographicaltrailer
- youtube_comments_comedytrailersspotlight
- youtube_comments_dramatrailer
- youtube_comments_familyanimationtrailers
- youtube_comments_horrortrailers
- youtube_comments_hotdocs
- youtube_comments_newindie
- youtube_comments_selected_actiontrailersspotlight
- youtube_comments_thisweektrailer
- youtube_comments_thriller

movie_merged_info
- 749_imdb_merged_movielist_info
- all_imdb_merged_movielist_info

trailer_reviews_original_TimeTransfered
- biographicaltrailer
- comedytrailersspotlight
- dramatrailer
- familyanimationtrailers
- horrortrailers
- hotdocs
- newindie
- selected_actiontrailersspotlight
- thisweektrailer
- thriller

trailer_reviews_SentimentAnalysis_textblob
- biographicaltrailer_textblob
- comedytrailersspotlight_textblob
- dramatrailer_textblob
- familyanimationtrailers_textblob
- horrortrailers_textblob
- hotdocs_textblob
- newindie_textblob
- selected_actiontrailersspotlight_textblob
- thisweektrailer_textblob
- thriller_textblob

trailer_reviews_SentimentAnalysis_textblob_TimeTransfered
- new_biographicaltrailer
- new_comedytrailersspotlight
- new_dramatrailer
- new_familyanimationtrailers
- new_horrortrailers
- new_hotdocs
- new_newindie
- new_selected_actiontrailersspotlight
- new_thisweektrailer
- new_thriller

all_new_movie_trailer_merged_info_copy1
new_familyanimationtrailers_copy1
sheet1
twitter_trailers_info_all

Data Features.

Movie information in IMDb
- Movie title
- Movie link
- certificate
- Synopsis
- rating
- runtime
- Director
- Movie genres
- Metascores
- USA Gross income
- Headline
- Writers
- Stars
- Plot keywords
- Official sites
- Country
- Language
- Filming location
- Cumulative Worldwide gross income
- Production Co.
- Sound Mix
- Color
- Aspect ratio, Description
- Trailer runtime
- Movie release time
- Budget
- US information
- Writers 2
- Stars 2
- Stars 3
- Storyline
- Plot Keywords 2
- Plot Keywords 3
- Plot Keywords 4

Plot Keywords 5

Trailer information in Youtube

- trailer publish time
- channelId
- trailer title
- description
- categoryId
- trailer duration
- viewCount
- likeCount
- dislikeCount
- favoriteCount

Trailer Reviews in Youtube

- Trailer title
- Trailer title link
- Trailer comments
- Trailer comments time
- Trailer comments like
  (how many people like this comment)
- Trailer transfered time
  (days between trailer publish date and 2019/08/01)
- Movie release time
- Movie transfered time
- Transfered time difference
  (trailer transfered time - Movie transfered time)
- Trailer comments sentiment(using textblob NLP,-1 1)
- Trailer comments subjectivity(using textblob NLP,-1 1)
- Trailer comments preference.
  (See technical report for functions.)

Twitter trailer information(479 tweets)

- Trailer title
- Trailer link
- Tweet comments
- Tweet time
- Tweet like
- Tweet share
- Tweet reply
- Tweet source
- Tweet whom

Trailer videos
(All trailers downloaded and sorted by categories, 40GB, 720P, MP4, name is same as 'Trailer title' in trailer information tables, about 3000 trailers in total.)

- 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

Data Table Introduction.
The above lists are group name mentioned in 'Group tables in mysql'.

749_movie_info_imdb_by_year_original
Movie information from popular movies(749 popular movies) lists in 2013-2019 from IMDb.

749_trailer_movie_merged_info
Merged trailer information (8616 trailers from 9 channels in YouTube) and matched popular movie information(from 749 popular movies, about 20% matching rate).

all_movie_info_imdb_by_year_original
Movie information from all movies(81270 movies) lists in 2013-2019 from IMDb.

all_trailer_info_original &
all_trailer_info_original_for_merge
Trailer information of all trailers(8616 trailers) from 2013 to 2019 in 9 channels in YouTube.

all_trailer_movie_merged_info
Merged trailer information (8616 trailers from 9 channels in YouTube) and matched all movie information(from 81270 movies, about 70% matching rate).

forCoefficientAnalysis_rating_cleaned &
forCoefficientAnalysis_usagross_cleaned &
forCoefficientAnalysis_worldgross_cleaned
Cleaned merged trailer information and matched all movie information by rating/usa gross income/worldwide gross income, these 3 tables are data for coefficient analysis.

movie_merged_info
Movie information of popular movies(749 movies) during 2013-2019 in IMDb. &
Movie information of all movies(81270 movies) during 2013-2019 in IMDb.

trailer_reviews_original
All reviews of trailers(864551 reviews) in 10 movie categories from Youtube Movieclips Trailers. Original scraped data.

trailer_reviews_original_TimeTransfered
All reviews of trailers(864551 reviews) in 10 movie categories from Youtube Movieclips Trailers. Transfered time is between days between trailer publish date and 2019/08/01.

trailer_reviews_SentimentAnalysis_textblob
All reviews of trailers(864551 reviews) in 10 movie categories from Youtube Movieclips Trailers. Using textblob for sentiment analysis and get 'sentiment -1 1' and 'subjectivity -1 1' for each review. Using preference as final value to replace 'like&dislike'.
(preference = sentiment*comment_like)

trailer_reviews_SentimentAnalysis_textblob_TimeTransfered

All reviews of trailers(864551 reviews) in 10 movie categories from Youtube Movieclips Trailers with 'preference' from sentiment analysis. By calculating time difference between movie_transfered_time and trailer_transfered_time to filter pre-released reviews and normalize reviews by time. (See technique report for time normalization details)

twitter_trailers_info_all

trailer information in twitter trailer channel(479 tweets)

Cleaned data.

The cleaned final data are groups as below

'749_trailer_movie_merged_info'
'all_trailer_movie_merged_info'
'forCoefficientAnalysis_rating_cleaned'
'forCoefficientAnalysis_usagross_cleaned'
'forCoefficientAnalysis_worldgross_cleaned'
'movie_merged_info'
'trailer_reviews_SentimentAnalysis_textblob_TimeTransfered'

Original Data.

The original data(directly from website using the method in 'Data Collection Intruction.pdf') are groups as below

'all_trailer_info_original'
'749_movie_info_imdb_by_year_original'
'all_movie_info_imdb_by_year_original'
'trailer_reviews_original'
'twitter_trailers_info_all'

Data Usage and Remarks.

In youtube.sql we provide raw data of all movie information(81270 movies) from IMDb during 2013-2019, all trailers information(8616 trailers) from 2013 to 2019 in 9 channels in YouTube, trailer reviews (864551 reviews) in 10 movie categories from Youtube Movieclips Trailers, as well as 3000 trailer videos in 10 movie categories in Youtube.

For current research, we match all trailers and all corresponding movies together, in order to use trailers information as pre-released meta-data to predict the movie gross income, rating, and popularity.

To solve the time problem(described in technical report) and normalization problem, we use sentiment analysis to calculate the preference from each trailer review to compare the performance and reliability of simple 'like&dislike' for prediction in previous papers.

By way of coefficient analysis, we preliminary find the result as below

From the coefficient analysis result, we could draw the conclusion that 'comments count' has much higher correlation with USA gross income and worldwide gross income than simple meta-date like 'view count','like count', and 'dislike count'.(about 10 times higher). Therefore we believe that the preference from trailer reviews have much higher performance than simple meta-data of trailer in YouTube. We would prove the assumption in further co-efficiency analysis.

Be aware that each channel has different trailer amount and subscribers. The data amount is also shown in the below tables.

Table 1. Correlation between movie rating and meta-data of trailer

Rating correlation	trailer series	view count	like count	dislike count	comment count	sentiment	preference
filmselecttrailer	0.02533283	0.03526385	0.0050197	0.00958571	-0.053747		
freshmovietrailer	-0.01440181	-0.07580638	-0.07292159	-0.13561284	-0.11801474		
furious trailer	-0.10815836	-0.093043	-0.14001656	-0.02585848	-0.10327457		
joblomovietrailer	-0.00786402	-0.0292394	-0.03607939	-0.08427883	-0.06071927		
movieclips trailer	0.02700788	0.00504539	0.00806546	-0.02381409	-0.02317481		
moviepilottrailer	-0.05652897	0.09932129	0.1084559	0.09204856	0.09605606		
movietrailersource	-0.0271737	0.05370674	0.06031079	0.00127742	-0.02493521		
newtrailerbuzz	0.12772609	-0.022946	0.03457465	-0.01498928	-0.02350908		
rapidtrailer	-0.01301163	0.00444308	-0.0230677	0.02161095	-0.07269746		

Table 2. Correlation between movie USA gross income and meta-data of trailer

USA Gross Income	trailer series	view count	like count	dislike count	comment count	sentiment	preference
filmselecttrailer	0.16238215	0.11563359	0.09771129	0.02928777	0.17571305		
freshmovietrailer	-0.03247056	0.14004506	0.11536722	0.07975358	0.1853188		
furious trailer	0.2574477	0.04731082	0.06555064	0.0238418	0.10809361		
joblomovietrailer	-0.02304071	0.19566287	0.20924704	0.17024959	0.22012332		
movieclips trailer	-0.11687523	0.14675415	0.21936088	0.05769043	0.2123239		
moviepilottrailer	0.24708663	0.2620017	0.12796143	0.14920888	0.09145186		
movietrailersource	0.07302531	0.0243072	0.00707041	-0.02030733	0.10357889		
newtrailerbuzz	-0.00220898	0.06282479	0.0468655	0.0620744	0.10764305		
rapidtrailer	0.10672939	0.10335676	0.08667873	0.01767974	0.17640523		

Table 3. Correlation between Worldwide USA gross income and meta-data of trailer

Worldwide Gross	trailer series	view count	like count	dislike count	comment count	sentiment	preference
filmselecttrailer	0.1573731	-0.00174282	0.02645849	-0.04269658	0.0427057		
freshmovietrailer	0.02815839	0.13289882	0.14492246	0.0978303	0.21468858		
furious trailer	0.37442343	0.04289278	0.10565868	0.0302329	0.16334648		
joblomovietrailer	-0.0011454	0.20321018	0.21359551	0.18066378	0.22128298		
movieclips trailer	-0.07532938	0.14924679	0.22110975	0.03148821	0.22409317		
moviepilottrailer	0.37532633	0.26685664	0.13342402	0.13186306	0.11057562		
movietrailersource	0.04660466	0.00761057	-0.00388638	-0.02634109	0.05929984		
newtrailerbuzz	0.06052892	0.0593351	-0.02875914	0.0214751	-0.02366493		
rapidtrailer	0.17006521	-0.02297426	0.01676616	-0.05546407	0.03273731		

Table 4. Subscriber and data amount of Rating experiment channels in YouTube

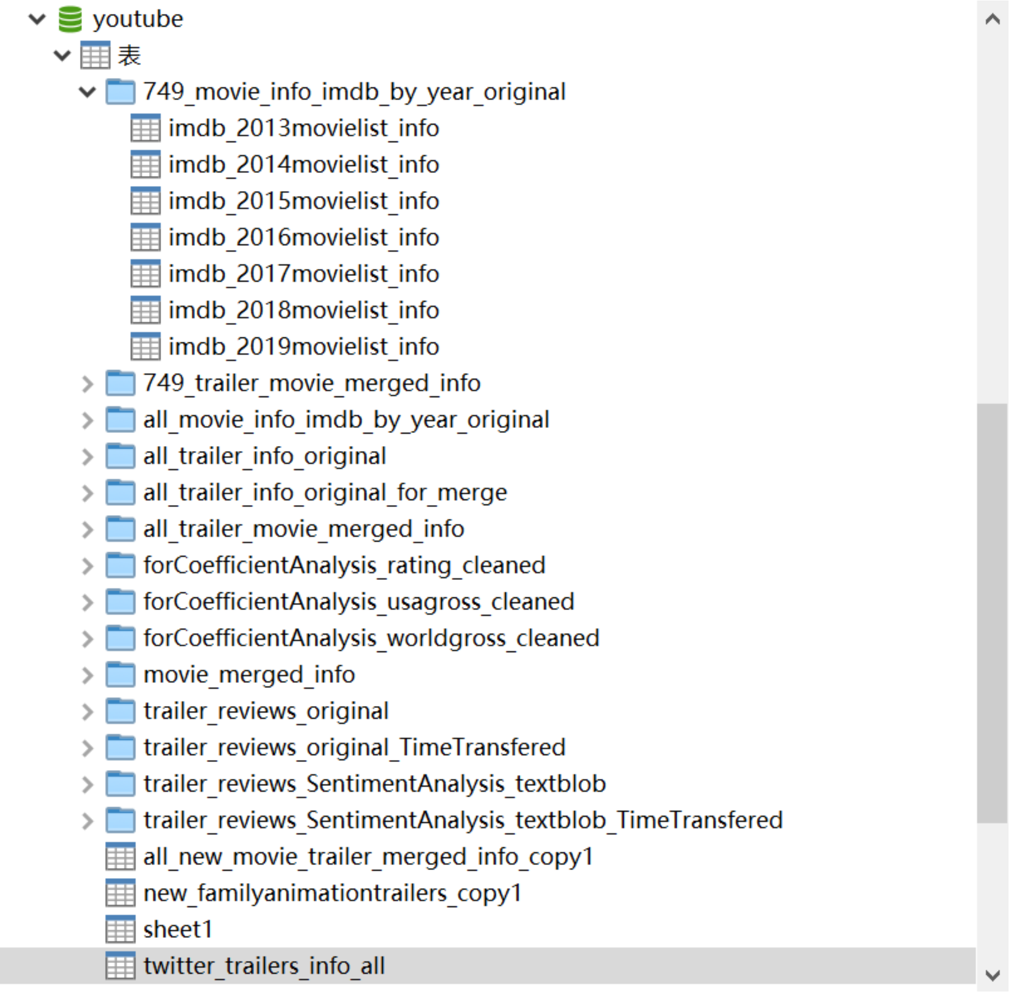
Rating Analysis	Data amount	Subscriber
filmselecttrailer	225	3.08M
freshmovietrailer	233	4.16M
furious trailer	183	3.23M
joblomovietrailer	394	2.08M
movieclips trailer	1092	14M
moviepilottrailer	266	818K
movietrailersource	232	1.3M
newtrailerbuzz	168	1.14M
rapidtrailer	253	794K

Table 5. Subscriber and data amount of USA gross income experiment channels in YouTube

USA gross income Analysis	Data amount	Subscriber
filmselecttrailer	66	3.08M
freshmovietrailer	139	4.16M
furious trailer	87	3.23M
joblomovietrailer	233	2.08M
movieclips trailer	952	14M
moviepilottrailer	116	818K
movietrailersource	97	1.3M
newtrailerbuzz	85	1.14M
rapidtrailer	57	794K

Table 6. Subscriber and data amount of Worldwide gross income experiment channels in YouTube

Worldwide gross income Analysis	Data amount	Subscriber
filmselecttrailer	225	3.08M
freshmovietrailer	233	4.16M
furious trailer	183	3.23M
joblomovietrailer	394	2.08M
movieclips trailer	1092	14M
moviepilottrailer	266	818K
movietrailersource	232	1.3M
newtrailerbuzz	168	1.14M
rapidtrailer	253	794K



- youtube
  - 表
    - 749_movie_info_imdb_by_year_original
      - imdb_2013movielist_info
      - imdb_2014movielist_info
      - imdb_2015movielist_info
      - imdb_2016movielist_info
      - imdb_2017movielist_info
      - imdb_2018movielist_info
      - imdb_2019movielist_info
    - 749_trailer_movie_merged_info
    - all_movie_info_imdb_by_year_original
    - all_trailer_info_original
    - all_trailer_info_original_for_merge
    - all_trailer_movie_merged_info
    - forCoefficientAnalysis_rating_cleaned
    - forCoefficientAnalysis_usagross_cleaned
    - forCoefficientAnalysis_worldgross_cleaned
    - movie_merged_info
    - trailer_reviews_original
    - trailer_reviews_original_TimeTransfered
    - trailer_reviews_SentimentAnalysis_textblob
    - trailer_reviews_SentimentAnalysis_textblob_TimeTransfered
    - all_new_movie_trailer_merged_info_copy1
    - new_familyanimationtrailers_copy1
    - sheet1
    - twitter_trailers_info_all

Fig. 1. Group your tables in mysql

Literature Review of Trailer Research

Mike Ma[1] *

[1]SIRC. Ontario Tech University, ON, CA

We have read more than 20 papers and select 10 of related ones to do analysis in our Literature Review. The first part is the related works on movie and trailer prediction. The second part is a brief analysis of related works. The third part is an analysis of worthy papers, focusing on inspiration to our research. The final part is potential research points for our research and updated proposed plan.

Related Works.

Movie and trailer prediction researches could be classified as three categories:
1.Predicting the Gross income(including stock price) of movie.(90%)
2.Predicting the rating(popularity) of movie.(15%)
3.Predicting the popularity of trailers(5%).

The directions could also be classified as three categories:
1.Using direct data(related pre-release and post- release data of trailers and movies) to build machine learning models and do correlation analysis.
2.Sentiment Analysis of comments of movies(not comments of trailers) using shallow NLP and machine learning.
3.Video shots analysis.

Research has been done to generate models for predicting revenues of movies. Most of them derived results from single data sources. Specifically, Joshi and others[1] use linear regression that joined meta-data with text features from pre-release critique to predict earnings for movies with a coefficient of determination $R\hat{2}$=0.671.
Mishne and Glance[2] correlate sentiments in blog posts with movie box-office scores. The correlations they observed for positive sentiments are fairly low and not sufficient to use for predictive purposes. They neither build prediction models or show the value of the correlation because they think the result is not good enough for accurate modeling.
Sharda and Delen[3] have treated the prediction problem as a classification problem and used neural networks to process pre-release data, such as quality and popularity variables, and classify movies into nine categories ranging from 'flop' to 'blockbuster'. Apart from the fact that they are predicting ranges over actual numbers, the best accuracy that their model can achieve is fairly low(36.9%).
Zhang and Skiena[4] have used a news aggregation model along with IMDb data to predict movie box-office numbers.
In a very interesting approach Asur and Huberman set up a prediction system for the revenue of movies based on the volume of Twitter mentions[5]. They build a linear regression model based on the chatter of Twitter and achieve an adjusted coefficient of determination of 0.97 on the night before the movie release for the first weekend revenue of a sample of 24 movies. In addition, they even tried to predict the Hollywood Stock Price given that social media can accurately predict box office results and the Hollyhood Stock Exchange adjusts the price for a movie stock to reflect the actual box office gross. They tested social media data efficacy compared with historical HSX prices at forecasting the stock prices of the HSX index and their tweet-rate proves to be significantly better at predicting the actual stock value than the historical HSX prices. It's a good inspiration of considering stock price because according to [6], Prices of movie stocks accurately predict real box office results, which demonstrates the strong correlation between the movie stock price and real box office gross income.
In a later work, however, Wong et al. show that Tweets do not necessarily represent the financial success of movies[7]. They consider a sample of 34 movies and compare the Tweets about the movies to evaluations written by users of the movie review websites. They argue that predictions based on social media could have high precision but low recall.
Marton and Taha[8] have showed that the populartity of a movie can be predicted much before its release by measuring and analyzing the activity level of editors and viewers of the corresponding entry to the movie in Wikipedia. It's novel because it is the only research using data from wikipedia, but the data features are too simple and low to support their conclusion.
In a rather novel approach, Oghina et al. have made use of Twitter and YouTube activity streams to predict the ratings in the Internet Movie Database(IMDb), which is among the most popular online movie database[9].
Reference[10] describes a movie rating approach based on data mining of 240 movies from IMDb where Weka and J48 were used to create the prototype model.
Reference[11] also applied Weka and J48 to generate three classes of movies: Hit, Neutral or Flop, in order to predict the movie box office performance. It's amazing that they also use trailer information and their data is quite similar to us and give us inspirations of more possible data features. They generate a model consisting of genre of successful movies ranked by user ratings of the IMDb, popularity of director, leading actor and leading actress represented by the number of comments and views of official movie trailers accessible by Youtube, and sentiment toward a movie derived from Youtube viewers' comments. However, their conclusion is not reliable considering their only 35 movies data.
In [12], methods to predict the popularity of movies were discussed to evolve as a guiding strategy for Content Distribution/ Delivery Network(CDD). Actor and director

popularity were considered as base criteria for predicting the popularity of a movie.

The only research focusing on trailers is [14], they analyzed whether subjective multimedia features be developed to predict the viewer's preference presented by like or dislike during and after watching movie trailers. The results showed that the single low-level multimedia feature of shot length variance is highly predictive of a viewer's "like/dislike" for a large portion of movie trailers. However, their features are too narrow and data is from only 1375 trailers. There are still lots of features to demonstrate the popularity of trailers.

Moreover, there has been substantial interest in the NLP community on using movie reviews as a domain to test sentiment analysis methods. e.g.,[13], et al. Basically speaking, they apply information retrieval or machine learning techniques to classify movie reviews into some categories and hope to produce better classification accuracy than human being. The classification categories are like "thumbs up" vs."thumbs down", "positive"vs."negative", or "like" vs."dislike".

In the video shots analysis part, researches could be classified as two categories: 1. Relationship between Media and Audiences' Affecting State: Related research focused on the emotion recognition of videos or movies. In these articles, some electronic signals, such as electroencephalogram(EGG), facial images, color features, the combination of audio and visual features and so on, are employed as the fundamental data for registering the viewer's affecting state.

2. Visual Data Feature Extraction: Shot segmentation and key-frame were extracted and the lighting key cues, motion, shot density, color energy cues, and other miscellaneous cues, including some audio data, were used as the features to predict the affecting potential of a film.[14]

Brief analysis of related works.

1. Most of the works on focusing on movie gross prediction. Different people work on movie gross prediction from different perspectives. Most previous work forecast movie grosses based on IMDb data with regression or stochastic models. However, their models either work poorly or need post-release data in order to make reasonable prediction, which are not acceptable in practice, because it is difficult to give shape estimation for either model parameters or gross if they don't have any early stage movie gross data. Although the post-release models are also useful in some situations, pre-release models are of more practical importance. Luckily, trailer is a good pre-release data source but seldom considered in related works.

2. While there has been research on predicting movie sales, almost all of them have used meta-data information on the movies themselves to perform the forecasting, such as the movies genre, MPAA rating, running time, release data, the number of screens on which the movie debuted, and the presence of particular actors or actresses in the cast. Trailer is seldom considered in related works.

3. Since predictions based on classic quality factors fail to reach a level of accuracy high enough for practical application usage of user-generated data to predict the success of a movie becomes a very temping approach. It indicates that sentiments analysis from reviewers' comments are worth to do.

4. Most predictions work on using movie data to predict movie performance. Seldom predictions work on using trailer data to predict movie performance. Only one prediction works on using trailer data to predict trailer performance, while their features are too simple(only like/dislike) and data size are too small.

5. Most predictions work on movies focuses on forecasting revenues, not ratings. Seldom predictions work on movies focuses on ratings. Only two predictions work on trailers. However, trailers are good pre-release data but ignored by many.

6. Predict movie stock price is worth to try because the strong correlation of movie performance and movie stock price. Most papers ignored this point.

7. Video shots analysis is not practical and not necessary in current stage considering its complexity and lots of simpler problems need to focus on. What also well worth mentioning is that almost all papers about video shots analysis get no more than 30 citations.

Analysis and Inspiration.

1. Improving Movie Gross Prediction Through News Analysis[4]: Their sentiment statistics are good. They derive several sentiment measures, including polarity, subjectivity, positive references per reference, and positive-negative differences per reference and give their definitions.

2. Early Prediction of Movie Box office Success Based on Wikipedia Activity Big Data[8]: They use features both individually and combined to repeat experiments, such as T,V,S,V,T,S,V,T,V,S. And they indicate that applicability of prediction model on movies with medium and low popularity levels remains an open question.

3. Predicting IMDB movie ratings using social media[9]: Their data is similar to us, including surface features from platform Youtube and Tweeter, as well as textual features from tweets as well as Youtube comments. Therefore, their experiment process is worth to study, although in deep learning we don't need correlations analysis like them. And although in their paper there is nothing about prediction as they said, all are correlations.

4. The real power of artificial markets[6]: It indicates that the prices of movie stocks accurately predict real box office results.

5. Predicting movie sales from blogger sentiment[2]: They analysis both pre-release data and post-release data. Their sentiment analysis methods are worthy. In addition, they focuse on the opening weekend data rather than total sales since this normalizes the figure across movies that were released on different dates, preventing earlier movies from having a higher income just because they have been "out there" longer.

6. Prediction of Movies Box Office Performance Using Social Media[11]: Their data are almost the same to us. It is

the only paper use both movie data and trailer data to predict the performance of movie. I believe it is the third most second most related paper among all to us.They give me inspiration of new data, including genre frequency, the popularity(followers) of actor and director. They also differentiate the movie by if it is sequel movie or the first-version movie. I strongly recommend to use the same way to differentiate our trailer, and give each trailer a sequel number. I would do this in data-cleaning process. Besides, their NLP methods toward Youtube comments are also worth copy.

7. Predicting Movie Trailer Viewer's Like/Dislike via Learned Shot Editing Patterns[14]: It is the only paper use trailer data to predict trailer data, their experiment process and charts are worth to copy. I believe it is also the second most related paper among all to us.

8. Predicting the Future With Social Media[9]: To be honest, this is my favourite paper among all those. The whole paper is good and citations are above 8000. They analysis the relationship between tweets-rate, sentiments of tweets and movie outcomes as well as movie stock price. I even want to copy all his methods, because their data could be replaced by my Youtube comments-rate and sentiments of Youtube trailers comments. I could even add more pre-release data from trailers to further the prediction. I am highly expect the potential results.

Potential Points.

1. Given meta-data about a movie and trailer, predict the popularity of the trailer after a fixed period of time.

2. Given meta-data about a movie and trailer, predict the popularity(ratings) of movie after a fixed period of time.

3. Given meta-data about a movie and trailer, predict the gross income of movies and the stock price of movie after a fixed period of time.

4. Test whether the sentiment of user-generated comments correlates with the movie's box office information better than simple counts of like/dislike from Youtube. Make predictions on movie popularity and stock price in terms of sentiments of user-generated comments. Dataset would be processed using a set of positive and negative words and then classified as positive, negative or neutral. Use sentiments data as new predictors.

Proposed Plan.

1. Collect Related Data(July 23 - Sep 1)

   - Description, duration, viewCount, likeCount, dislikeCount, favoriteCount, commentCount, Stars, Directors, ReleaseTime, Synopsis from about 10K trailers from 10 channels in youtube, including different language trailers.

   - About 100K Comments of all 8 categories trailers from the biggest channel in youtube.

   - About 3K all 8 categories trailers videos from the biggest channel in youtube, 720K, 40GB.

   - About 1000 trailers Sharing info from the trailer channel

   - Title, Link, Certificate, Content, rating, runtime, Director, Stars, Writers, Genre, PublishYear, Metascore, Gross revenue, Headline, Plot Keywords, certificate, Official Sites, Country, Language, Filming Location, Gross USA, Cumulative World Gross, Production Co, Sound Mix, Color, Aspect Ratio, Discription, TrailerRuntime, UK Release Time, Budget, Storyline of about 1K Movies from 2011-2020 movie lists in IMDb

2. Data Cleaning and save data to Cloud(Sep 1-Sep 8 if possible)
3. Read Papers and finish literature review.(Aug 19-Sep 1)
4. Learn deep learning, NLP and build models(Sep 1-Sep 10)
5. Begin Experiment.(After data cleaning)
6. Write a formal paper.(Maybe after I go back, I'm not ambiguous about this)

Bibliography.

1. M. Joshi, D. Das, K. Grimpel, and N. A.Smith, "Movie revies and revenues: An experiment in text regression." *NAACL-HLT*, 2010.
2. G.Mishne and N.Glance, "Predicting movie sales from blogger sentiment." in *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
3. R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks." in *Expert Systems with Applications*, vol. 30, 2006, pp. 243–254.
4. W.Zhang and S.Skiena, "Improving movie gross prediction through news analysis." in *In Web Intelligence*, 2009, p. 301304.
5. H. B. Asur S, "Predicting the future with social media." in *In Processings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010, pp. 492–499.
6. D. M. Pennock and S. Lawrence, "The real power of artificial markets." in *Science: American Association for the Advancement of Science*.
7. W. FMF, S. S, and C. M, "Why watching movie tweets won't tell the whole story?" in *In Processings of the 2012 ACM workshop on Workshop on online social networks. New York, NY, USA: ACM, WOSN'12*, 2012, pp. 61–66.
8. M. Mestyan and T. Yasseri, "Early prediction of movie box office success based on wikipedia activity big data," in *PLoS ONE 8(8): e71226. doi:10.1371/journal.pone.0071226*, 2013.
9. O. A, B. M, T. E, and de Rijke M, "Predicting imdb movie ratings using social media." in *In: ECIR 2012: 34th European Conference on Information Retrieval. Springer-Verlag, Barcelona, Spain: Springer-Verlag*, 2012, pp. 503–507.
10. S. Kabinsingha, S. Chindasorn, and C. Chantrapornchai, "Movie rating approach and application based on data mining," in *International Journal of Engineering and Innovative Technology(IJEIT)*, 2012.
11. K. R. Apala and M. Jose, "Prediction of movies box office performance using social media," in *2013 IEEE/ACM International Conference on Adances in Social Networks Analysis and Mining*, 20123,.
12. L. Zhu, M. Zhu, and S. Yao, "The popularity of movies predict system based on data mining technology for cdn." in *IEEE International Conference on the 3rd Computer Science and Information Technology*, 2010.
13. P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches." in *In Proceedings of the Hawaii International Conference on System Sciences(HICSS)*, 2005.
14. Y. Hou, T. Xiao, and S. Zhang, "Predicting movie trailer viewer's like/dislike via learned shot editing patterns," in *IEEE Transactions on Affective Computing*, no. 1, 2016.

Proposed Plan of Trailer Research

Mike Ma[1 *]

[1]SIRC. Ontario Tech University, ON, CA

In this proposed plan I described the motivation of choosing data and research direction in the first part. In the second part I discussed the possible problems before literature reviews. In the third part I discussed my proposed plan with time-frame, including a short summary of data collected.

Motivation.

The success of failure of a movie is often determined in its first weekend of play. In order to make this opening successful, movie producers must employ a number of promotional strategies including movie trailers, to publicize the movie for a significant period of time prior to its release. A movie trailer, as described in Wikipedia, is an advertisement or a commercial for a feature film that will be exhibited in the future. Of some ten billion videos watched online annually, film trailers rank the third, after news and user-created video.It is clear that businesses have a strong interest in tapping into this huge data source to extract information that might improve their decision making process. For example, predictive models derived from movie and its trailers may facilitate filmmakers making more profitable decisions.

Possible Problems.

1. Given meta-data about a movie and trailer, predict the number of views the trailer will receive after a fixed period of time.
2. Given meta-data about a movie and trailer, create a method to measure the preference(like, dislike, tweet, share, and preference of comments) of audience toward trailer, and predict the preference after a fixed period of time.
3. Analyze the relationship between preference of trailer and rating of movie, and predict the rating of a movie before it published.
4. Analyze the relationship between preference of trailer and movie grosses, and predict the grosses and movie stock price of the movie.
6. Test whether the sentiment of user-generated comments correlates with the movie's box office information better than simple counts of like/dislike from Youtube. Make predictions on movie popularity and stock price in terms of sentiments of user-generated comments. Dataset would be processed using a set of positive and negative words and then classified as positive, negative or neutral. Use sentiments data as new predictors.

Proposed Plan.

1. Collect Related Data(July 23 - Sep 1)

- Description, duration, viewCount, likeCount, dislikeCount, favoriteCount, commentCount, Stars, Directors, ReleaseTime, Synopsis from about 10K trailers from 10 channels in youtube, including different language trailers.
- About 100K Comments of all 8 categories trailers from the biggest channel in youtube.
- About 3K all 8 categories trailers videos from the biggest channel in youtube, 720K, 40GB.
- About 1000 trailers Sharing info from the trailer channel
- Title, Link, Certificate, Content, rating, runtime, Director, Stars, Writers, Genre, PublishYear, Metascore, Gross revenue, Headline, Plot Keywords, certificate, Official Sites, Country, Language, Filming Location, Gross USA, Cumulative World Gross, Production Co, Sound Mix, Color, Aspect Ratio, Discription, TrailerRuntime, UK Release Time, Budget, Storyline of about 1K Movies from 2011-2020 movie lists in IMDb

2. Data Cleaning and save data to Cloud(Sep 1-Sep 8 if possible)
3. Read Papers and finish literature review.(Aug 19-Sep 1)
4. Learn deep learning, NLP and build models(Aug 25- Sep 10)
5. Begin Experiment.(Sep 5-Sep 15)
6. Write a formal paper.(Sep 15-Oct X)

Weekly Report 20190828

Current progress:

Currently I have selected 4 excellent papers which are very inspirational for our projects(even the same purpose) from about 15 papers , I did notes very carefully for each of these papers. I plan to read their reference papers in these 3 days. I would begin writing my formal literature on Friday and I would send it to you before our online meeting.

Today I went to library because I forget to take my access card. This morning I have read a very good paper which use tweets-rate and sentiment analysis to predict the movie revenue and even stock price. It is even same to my idea and gives me lots of inspiration, I would write my plans details in literature review. Please be assured that I would guarantee the efficiency and work time.

Current problems:

1. The biggest problem is matching of movies in IMDb and trailers in Youtube. In fact currently I don't have good ideas for matching the titles.(There are two data tables for IMDb and YouTube and we need to combines them into one table).
2. Could you please recommend me some papers of "deep learning for prediction" to read. Zach gives me a pdf book of Deep Learning by Ian Goodfellow, Yoshua Bengio and Aaron Courville and a good website. But I can't focuses on them in short term. Bahara suggests me to read papers firstly and find the potential specific methods and then learn it. Could you please recommend me some papers or a introduction ppt. Thank you very much.

Current plan:

1. Read papers and finish literature reviews this week.
2. Data cleaning and solving matching problems next week(and start learning necessary methods).

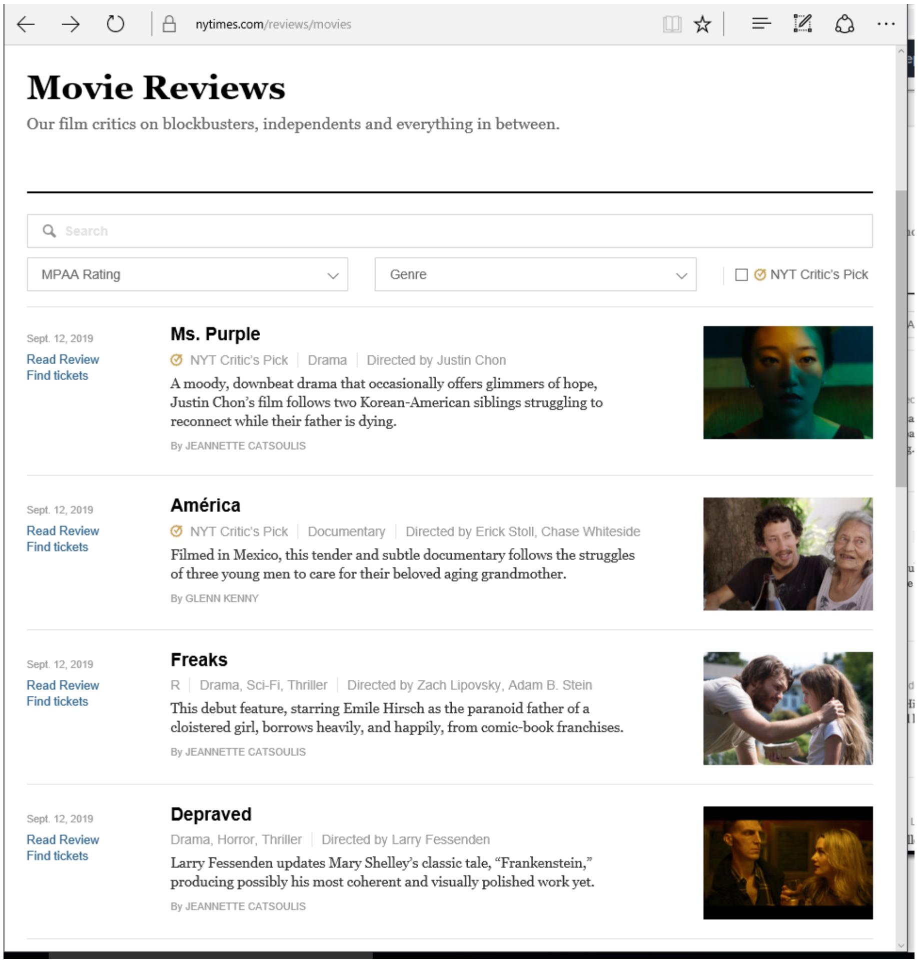
Weekly Report

Mike Ma

September 12

1 Brief Summary

There is a Brief Summary of what I have done recently.
1. Having solved the title matching problem, and merge the trailer data from Youtube and movie data from IMDb into one table.
(This means we correspond specific movie details for each trailer and we could use trailer data to predict movie now.)
2. There are new problems after title matching, the movie data are incomplete, which means some trailer couldn't find corresponded movie data. Therefore, I am collecting the whole movies information from IMDb from 2011-2019, about 9000 movies per years. The software is still running, but I have already written the title-matching codes for this new movie data table. By tomorrow, I could achieve a much more complete merged table.
(I could collect everything you see in the first page, but I couldn't get access to the second page.)



Movie Reviews

Our film critics on blockbusters, independents and everything in between.

Sept. 12, 2019
Read Review
Find tickets

Ms. Purple

NYT Critic's Pick | Drama | Directed by Justin Chon

A moody, downbeat drama that occasionally offers glimmers of hope, Justin Chon's film follows two Korean-American siblings struggling to reconnect while their father is dying.

By JEANNETTE CATSOULIS

Sept. 12, 2019
Read Review
Find tickets

América

NYT Critic's Pick | Documentary | Directed by Erick Stoll, Chase Whiteside

Filmed in Mexico, this tender and subtle documentary follows the struggles of three young men to care for their beloved aging grandmother.

By GLENN KENNY

Sept. 12, 2019
Read Review
Find tickets

Freaks

R | Drama, Sci-Fi, Thriller | Directed by Zach Lipovsky, Adam B. Stein

This debut feature, starring Emile Hirsch as the paranoid father of a cloistered girl, borrows heavily, and happily, from comic-book franchises.

By JEANNETTE CATSOULIS

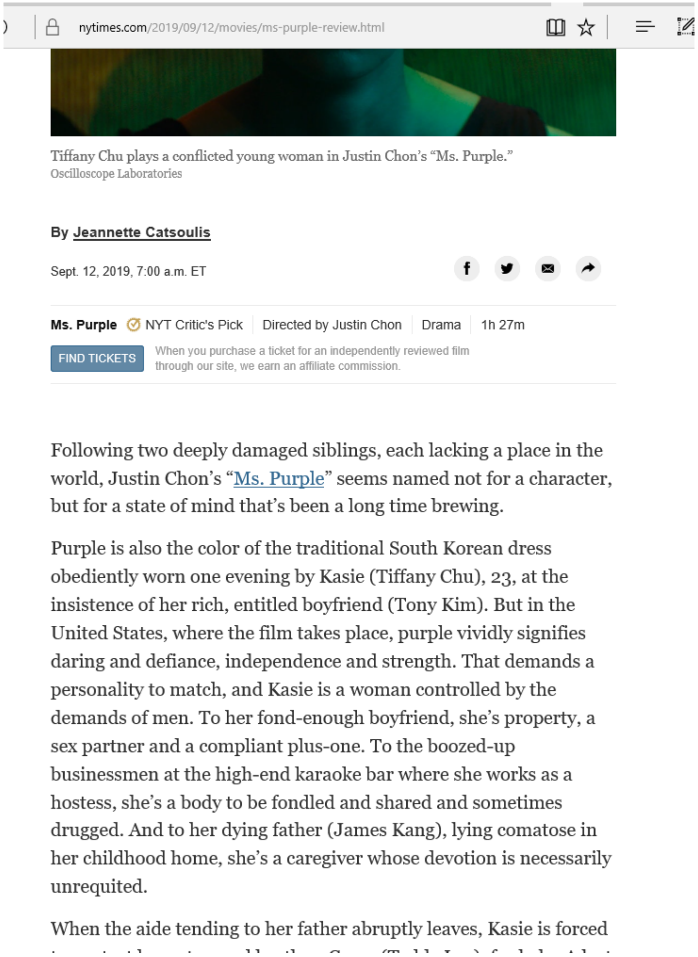
Sept. 12, 2019
Read Review
Find tickets

Depraved

Drama, Horror, Thriller | Directed by Larry Fessenden

Larry Fessenden updates Mary Shelley's classic tale, "Frankenstein," producing possibly his most coherent and visually polished work yet.

By JEANNETTE CATSOULIS



nytimes.com/2019/09/12/movies/ms-purple-review.html

Tiffany Chu plays a conflicted young woman in Justin Chon's "Ms. Purple."
Oscilloscope Laboratories

By Jeannette Catsoulis

Sept. 12, 2019, 7:00 a.m. ET

Ms. Purple | NYT Critic's Pick | Directed by Justin Chon | Drama | 1h 27m

FIND TICKETS When you purchase a ticket for an independently reviewed film through our site, we earn an affiliate commission.

Following two deeply damaged siblings, each lacking a place in the world, Justin Chon's "Ms. Purple" seems named not for a character, but for a state of mind that's been a long time brewing.

Purple is also the color of the traditional South Korean dress obediently worn one evening by Kasie (Tiffany Chu), 23, at the insistence of her rich, entitled boyfriend (Tony Kim). But in the United States, where the film takes place, purple vividly signifies daring and defiance, independence and strength. That demands a personality to match, and Kasie is a woman controlled by the demands of men. To her fond-enough boyfriend, she's property, a sex partner and a compliant plus-one. To the boozed-up businessmen at the high-end karaoke bar where she works as a hostess, she's a body to be fondled and shared and sometimes drugged. And to her dying father (James Kang), lying comatose in her childhood home, she's a caregiver whose devotion is necessarily unrequited.

When the aide tending to her father abruptly leaves, Kasie is forced

3. Having eliminated the invalid data.
(for example, there are some videos "New Best Movies Trailers top 10" rather than trailer for specific movie.)
4. Having extracted the series number for each trailer.
5. Having organized the data and stored them into google cloud.
6. Having tried collect movie reviews from New York Time. I could get pre-released movie professional reviews from New York Time. I now agree that this is a good data resource. Each movie has one professional review, and it is usually a long article. However, what I could collect now is the "brief review summary" of each movie. It is usually the first paragraph of the article.
(Because to get the full review article New York Test would check "I'm not a robot")

2 Method of title matching

1. Export the trailer data table and movie data table from MySQL into Json.
2. Create list for each field.
3. Use re.match() to match.
4. Check the format of unmatched part, if it included subtitle or series number, the matched part would not be considered.
5. store all the fields into one database.

3 Next Step

1. Collect the merged data as complete as possible.
2. Collect the remarks from New York Time.
3. Select the trailer comments in a fixed period before publish time.
4. Comments Sentiments Analysis.

4 Current Problems

1. I can't get the whole comment article in New York Time. Each movie has one professional review, and it is usually a long article. However, what I could collect now is the "brief review summary" of each movie. It is usually the first paragraph of the article.
2. We need to decide how long the fixed period should be.(Select the trailer comments in a fixed period before publish time. Like all comments one year before the trailer was published)
3. After the sentiments analysis, what we could only get is just a number of "like", "dislike", and "neutral". However, the amount of sentiment analysed trailers are much smaller than the current trailers(All current trailers have "like" and "dislike" count but not valid because of trailer published at the different time.).

Weekly Report -
— Reviews Selection Sentiment Analysis

Mike Ma

September 22

1 Brief Summary

This is a weekly report of movietrailer research from Mike. I focused on comments selection and sentiment analysis this week. In this report, I would introduce a method of comments selection I used.

2 Major problem of other papers

The only paper focusing on trailer prediction in google scholar is "Predicting movie trailer viewer's "like/dislike" via learned shot editing patterns", However, they directly use "like/dislike" in YouTube to do the prediction. However, there is a big problem is their paper. Think about this:

Movie 1 is released 3 days ago, whose trailer has 400 likes.
Movie 2 is released 3 years ago, whose trailer has 4000 likes.

However, we couldn't say movie 2 is better than movie 1, while in their paper they ignore this problem.

3 My method

To challenge their result, I don't use like/dislike only. Rather, I use the sentiments of comments in a fixed period before the movie released to do the prediction. -
Here I need the "movie release time", "trailer comments time(review time of each comment)" "sentimentssubjectivity value of comments", "likes of comments" as essences. -
Because movie information and trailer information are different tables, while I need to match "movie release time" and its respondent "trailer comments time". That's why text matching is indispensable for our research. -
Then I will introduce my method and why it solves the problem of their paper.

You would know why our time on comments selection and sentiment analysis are indispensable and valuable.

4 Method introduction

1. Get “movie release time” and “trailer comments time” from the title-matched table and comments table.

2. Transfer “trailer comments time”(like 3 days ago, 6 days ago).
0-24 hours ago — 2019-07-30 (1 days)
1 days ago — 2019-07-29 (2 days)
1 weeks ago - 2019-07-18(13 days)
1 months ago - 2019-06-02(59 days)
1 years ago - 2017-08-02(729 days)

3. Compare “movie release time” and “trailer comments time” to current data-collecting time(2019-08-01) individually. Calculate the differences:
Difference 1 = 2019-08-01 - “movie release time”
Difference 2 = 2019-08-01 - “trailer comment time”
-
4. Prerequisites for comments selection
(1)Guarantee the review of trailer before the movie released.
Difference 2 - Difference 1>0
(2)Set the comments selection period as 6 months(180 days).
0 <Difference 2 - Difference 1 <180
(3)Delate trailer info which is released less than 6 months.
Difference 2 >180
-
5. Comments selection -
Using the value of “difference 2 - difference 1”, we could normalize each Trailer review in a fixed period of time(6 months)
-
6. Sentiments analysis -
Using NLP to do sentiment analysis for each selected review. I have already used blobtext from python and it success, but I don’t know what the function of value“objectivity”.
-
Besides, there are lots of popular sentiments analysis method, I’m still learning them. In this part, we could choose popular API for sentiment analysis, like IBM or Google, in this case we don’t need to train data. Or rather, we could use NLTK to label parts of comments as training set and build a model ourselves.
-
If there is any suggestion of sentiment analysis, please let me know. I sincerely hope we could have a little discussion of sentiment analysis. NLP is cool.

Weekly Report -
Data Improvement and Plan

Mike Ma

October 2

1 Current Situation

I recollected about 82K movies information from 2013-2019 in IMDb. Previously the useful data after title-match is only 10 percents because of the incomplete movie information. However, the recollected data make the useful comments improved to more than 70 percents after data cleaning and title match.

I have sent you the invitation to my programs on Github.And you could download the data from Dropbox. When I came back to China I would write an introduction about all my data. In this 10 weeks I have achieved data collection, data cleaning, title matching, comments transfer, comments selection, and comments sentiment analysis. For all the revision ideas we discussed, I guarantee to achieve them.
Yesterday I registered my github and find it so fascinated! I have uploaded some of my previous projects and you are welcome to visit. I would arrange it very well in the future.

I know that there is only one week I could stay in this lab, I cherish every moment here and really reluctant to leave. Here I met all so kind friends and even my best friend. I have guaranteed them that I would came back every year to this lab to visit everyone. And for this project, I have a plan and sincerely hope you would like to continue be my supervisor.

2 Future Plan

Next week I would try my best to finish the correlation analysis between each feature and gross income, ratings. After that I would write a summary about the 3 months work. After I came back to China, I hope to continue to work on this project and use this project as my graduation design project(which require us to do one project under supervision in final year). I sincerely hope to continue our weekly meeting by Skype.

I plan to apply for engineer degree in Canada or the USA but I sincerely hope to continue my research project with you, professor Amirali, because I could really learn something practical and you the the rare professor who really cares about students. All my family feels so lucky that my supervisor is you. Although this semester I would be busy for my graduate application. From January next year, I could have 6 months with full concentration in this project and in my first year graduate school I could also continue on my project. One Skype meeting per week is enough for me and I would come back to our lab every year.(every semester in my graduate). Please trust me I'm serious. Zach gave me some very good material to learn deep learning. I really believe my method is more reliable than previously paper and I sincerely hope to continue finishing it. And I know that research experience is necessary for a job position. I didn't waste any moment here and have guaranteed work efficiency and time every week. Please trust me I would continue to be self-discipline if you would like to continue to be my supervisor.

Summary of Trailer Data Collection v1.0

Mike Ma

Table of Contents

1. Data Introduction...1

1.1 Data Source...1

1.2 Data Introduction...3

1.3 Personal Opinion About Data...10

2. Tools and Methods...12

3. Operate Steps(Run yourself or get more data)...12

4. Extended Program...18

5. Summary...18

Four parts of my task:...18

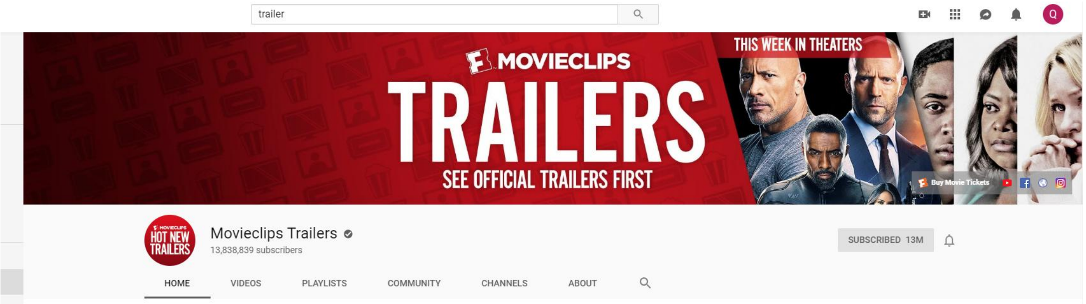
6. Problems and Improvements...19

1. Data Introduction

1.1 Data Source

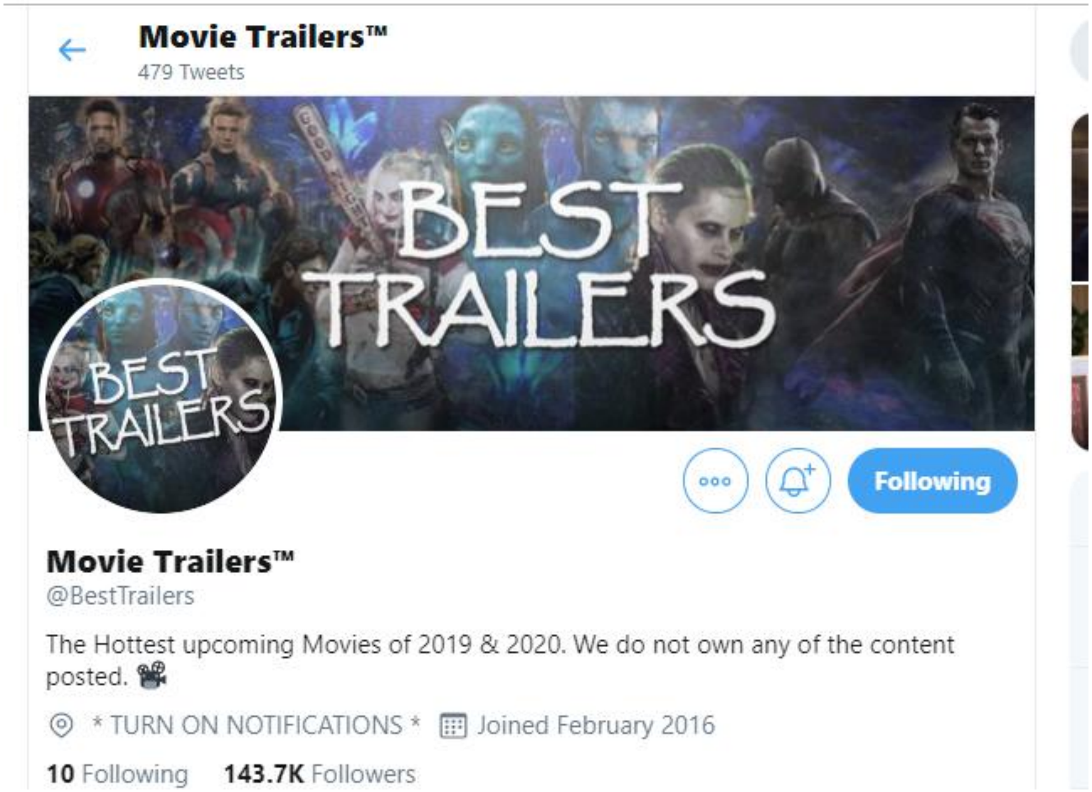
- YouTube- Movieclips Trailers (8K+ official trailers & 14M+ subscribers)

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



- Twitter

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



Movie Trailers™
479 Tweets

Movie Trailers™
@BestTrailers

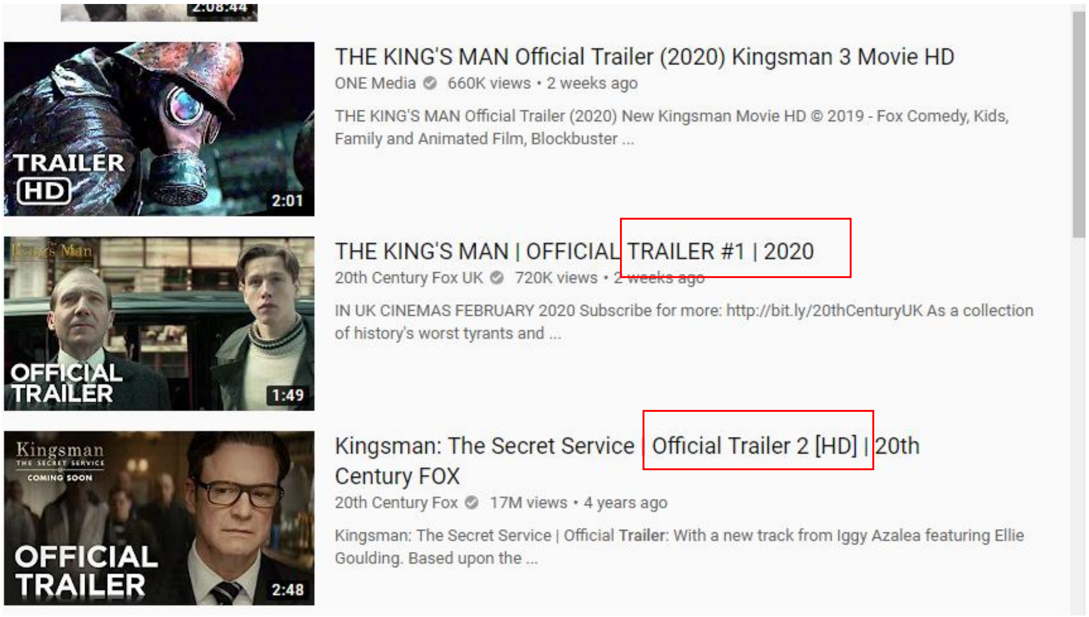
The Hottest upcoming Movies of 2019 & 2020. We do not own any of the content posted. 📽

* TURN ON NOTIFICATIONS * Joined February 2016

10 Following 143.7K Followers

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:



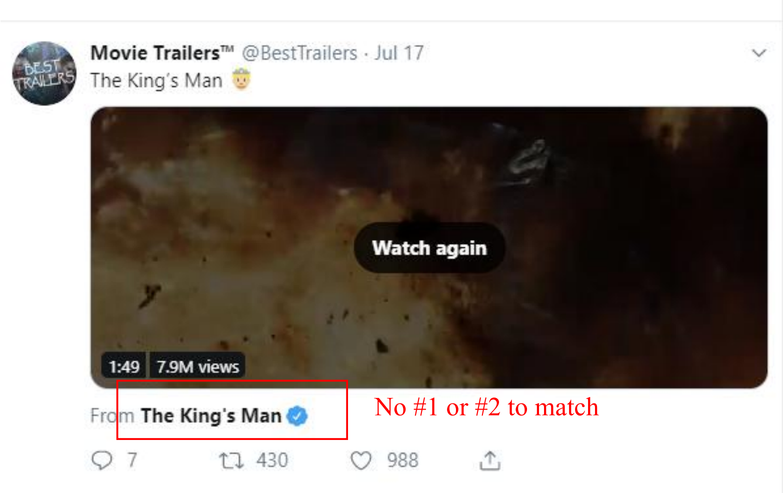
2:08:44

THE KING'S MAN Official Trailer (2020) Kingsman 3 Movie HD
ONE Media 660K views • 2 weeks ago
THE KING'S MAN Official Trailer (2020) New Kingsman Movie HD © 2019 - Fox Comedy, Kids, Family and Animated Film, Blockbuster ...

THE KING'S MAN | OFFICIAL TRAILER #1 | 2020
20th Century Fox UK 720K views • 2 weeks ago
IN UK CINEMAS FEBRUARY 2020 Subscribe for more: http://bit.ly/20thCenturyUK As a collection of history's worst tyrants and ...

Kingsman: The Secret Service | Official Trailer 2 [HD] | 20th Century FOX
20th Century Fox 17M views • 4 years ago
Kingsman: The Secret Service | Official **Trailer**: With a new track from Iggy Azalea featuring Ellie Goulding. Based upon the ...

However in Twitter:



Movie Trailers™ @BestTrailers · Jul 17
The King's Man 🤴

1:49 7.9M views

From **The King's Man**

No #1 or #2 to match

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)

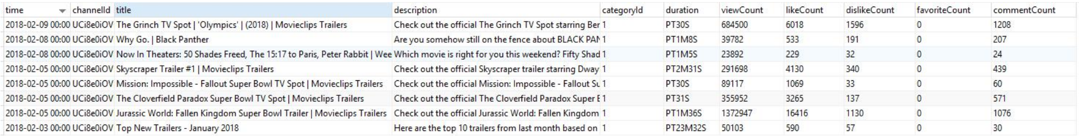
1.2 Data Introduction

- Data-table: youtube_info_all

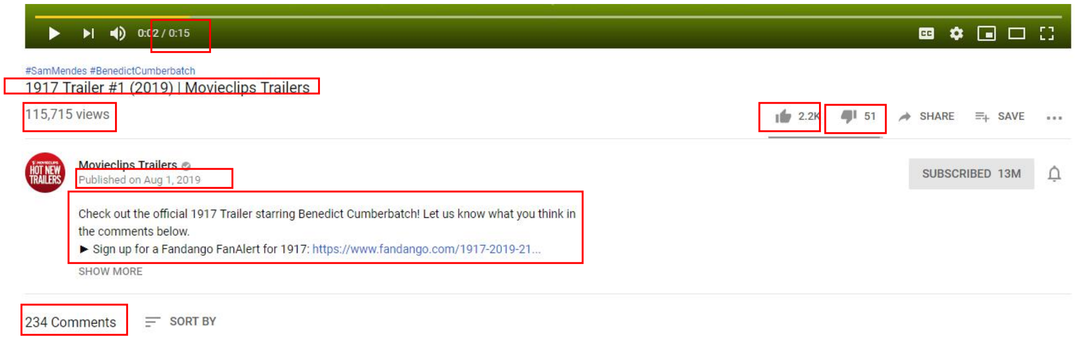
(All trailers published from 2018.1.1-2019.8.1, 700+trailers)

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



#SamMendes #BenedictCumberbatch

1917 Trailer #1 (2019) | Movieclips Trailers

115,715 views

Movieclips Trailers

Published on Aug 1, 2019

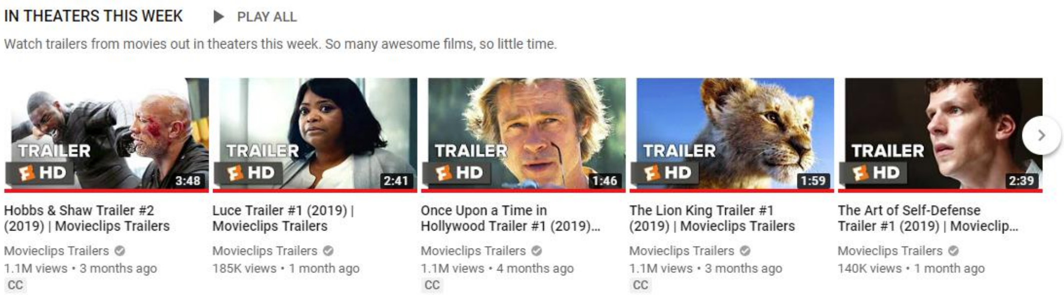
Check out the official 1917 Trailer starring Benedict Cumberbatch! Let us know what you think in the comments below.

► Sign up for a Fandango FanAlert for 1917: https://www.fandango.com/1917-2019-21...

234 Comments

- 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)



IN THEATERS THIS WEEK

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

Hobbs & Shaw Trailer #2 (2019) | Movieclips Trailers
Movieclips Trailers
1.1M views • 3 months ago

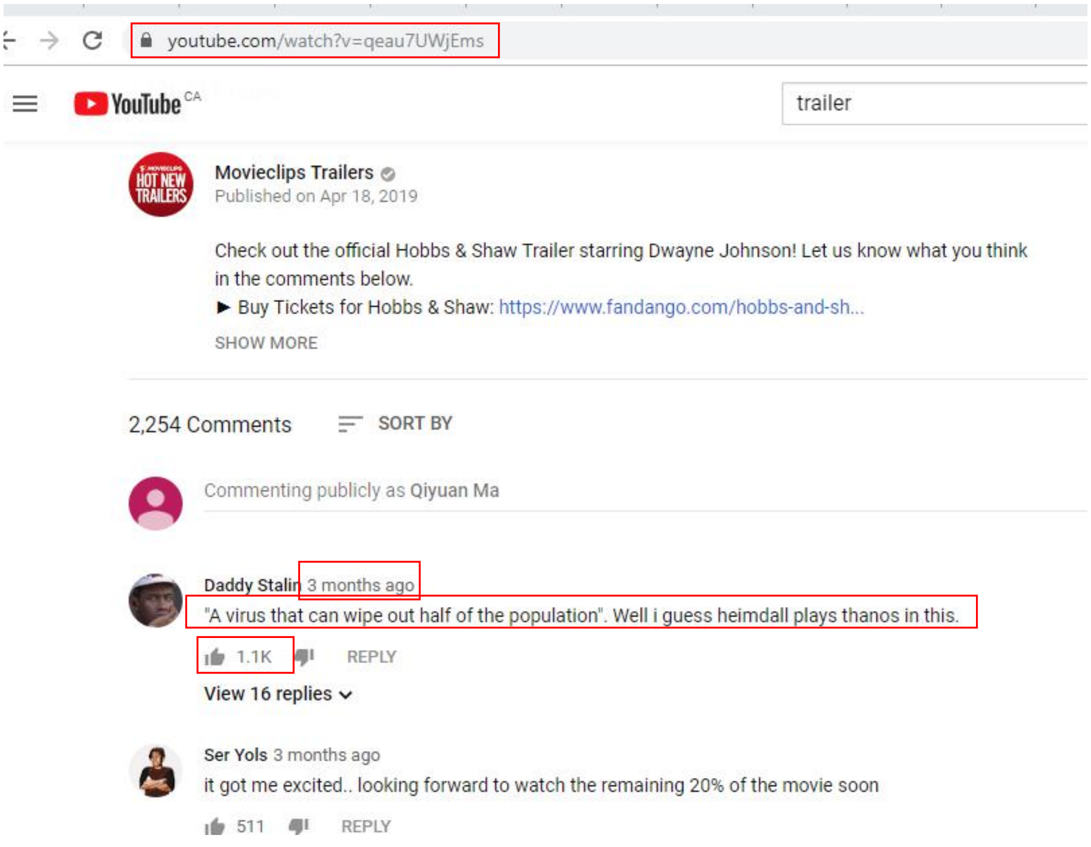
Luce Trailer #1 (2019) | Movieclips Trailers
Movieclips Trailers
185K views • 1 month ago

Once Upon a Time in Hollywood Trailer #1 (2019)...
Movieclips Trailers
1.1M views • 4 months ago

The Lion King Trailer #1 (2019) | Movieclips Trailers
Movieclips Trailers
1.1M views • 3 months ago

The Art of Self-Defense Trailer #1 (2019) | Movieclip...
Movieclips Trailers
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**



youtube.com/watch?v=qeau7UWjEms

Movieclips Trailers
Published on Apr 18, 2019

Check out the official Hobbs & Shaw Trailer starring Dwayne Johnson! Let us know what you think in the comments below.
► Buy Tickets for Hobbs & Shaw: https://www.fandango.com/hobbs-and-sh...

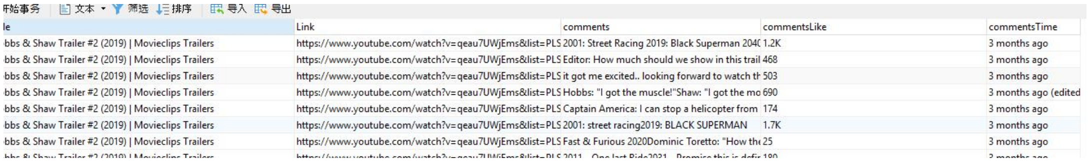
2,254 Comments

Commenting publicly as Qiyuan Ma

Daddy Stalin 3 months ago
"A virus that can wipe out half of the population". Well i guess heimdall plays thanos in this.
1.1K

Ser Yols 3 months ago
it got me excited.. looking forward to watch the remaining 20% of the movie soon
511

REPLY

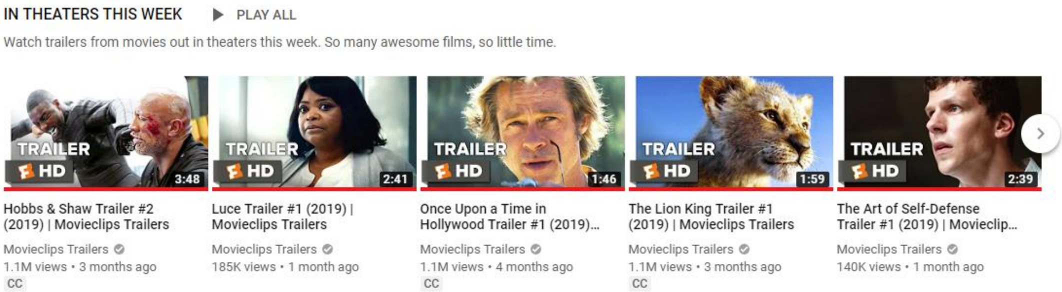


开始事务
文本
筛选
排序
导入
导出

le	Link	comments	commentsLike	commentsTime
bbs & Shaw Trailer #2 (2019) | Movieclips Trailers	https://www.youtube.com/watch?v=qeau7UWjEms&list=PLS	2001: Street Racing 2019: Black Superman 2040	1.2K	3 months ago
bbs & Shaw Trailer #2 (2019) | Movieclips Trailers	https://www.youtube.com/watch?v=qeau7UWjEms&list=PLS	Editor: How much should we show in this trail	468	3 months ago
bbs & Shaw Trailer #2 (2019) | Movieclips Trailers	https://www.youtube.com/watch?v=qeau7UWjEms&list=PLS	it got me excited.. looking forward to watch th	503	3 months ago
bbs & Shaw Trailer #2 (2019) | Movieclips Trailers	https://www.youtube.com/watch?v=qeau7UWjEms&list=PLS	Hobbs: "I got the muscle!"Shaw: "I got the mo	690	3 months ago (edited
bbs & Shaw Trailer #2 (2019) | Movieclips Trailers	https://www.youtube.com/watch?v=qeau7UWjEms&list=PLS	Captain America: I can stop a helicopter from	174	3 months ago
bbs & Shaw Trailer #2 (2019) | Movieclips Trailers	https://www.youtube.com/watch?v=qeau7UWjEms&list=PLS	2001: street racing2019: BLACK SUPERMAN	1.7K	3 months ago
bbs & Shaw Trailer #2 (2019) | Movieclips Trailers	https://www.youtube.com/watch?v=qeau7UWjEms&list=PLS	Fast & Furious 2020Dominic Toretto: "How the	25	3 months ago
bbs & Shaw Trailer #2 (2019) | Movieclips Trailers	https://www.youtube.com/watch?v=qeau7UWjEms&list=PLS	2011 – One last Ride2021 – Promise this is defir	180	3 months ago

- 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
Movieclips Trailers
1.1M views • 3 months ago
CC

Luce Trailer #1 (2019) | Movieclips Trailers
Movieclips Trailers
185K views • 1 month ago

Once Upon a Time in Hollywood Trailer #1 (2019)...
Movieclips Trailers
1.1M views • 4 months ago
CC

The Lion King Trailer #1 (2019) | Movieclips Trailers
Movieclips Trailers
1.1M views • 3 months ago
CC

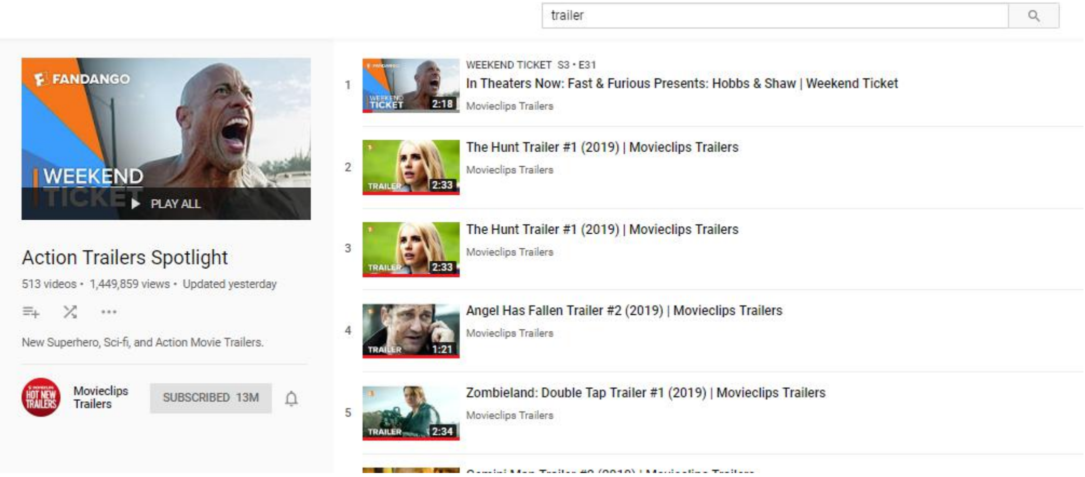
The Art of Self-Defense Trailer #1 (2019) | Movieclip...
Movieclips Trailers
140K views • 1 month ago

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).



trailer

Action Trailers Spotlight
513 videos • 1,449,859 views • Updated yesterday
New Superhero, Sci-fi, and Action Movie Trailers.
Movieclips Trailers SUBSCRIBED 13M

1 WEEKEND TICKET S3 • E31
In Theaters Now: Fast & Furious Presents: Hobbs & Shaw | Weekend Ticket
Movieclips Trailers

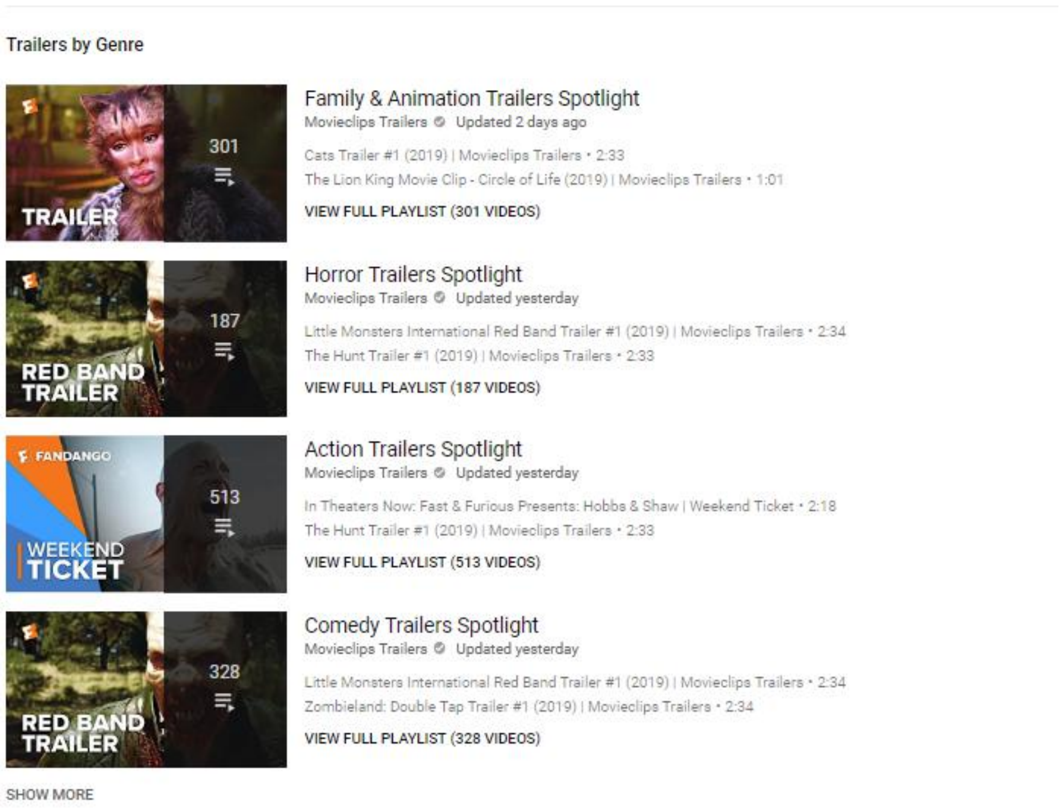
2 The Hunt Trailer #1 (2019) | Movieclips Trailers
Movieclips Trailers

3 The Hunt Trailer #1 (2019) | Movieclips Trailers
Movieclips Trailers

4 Angel Has Fallen Trailer #2 (2019) | Movieclips Trailers
Movieclips Trailers

5 Zombieland: Double Tap Trailer #1 (2019) | Movieclips Trailers
Movieclips Trailers

Gemini Man Trailer #2 (2019) | Movieclips Trailers



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

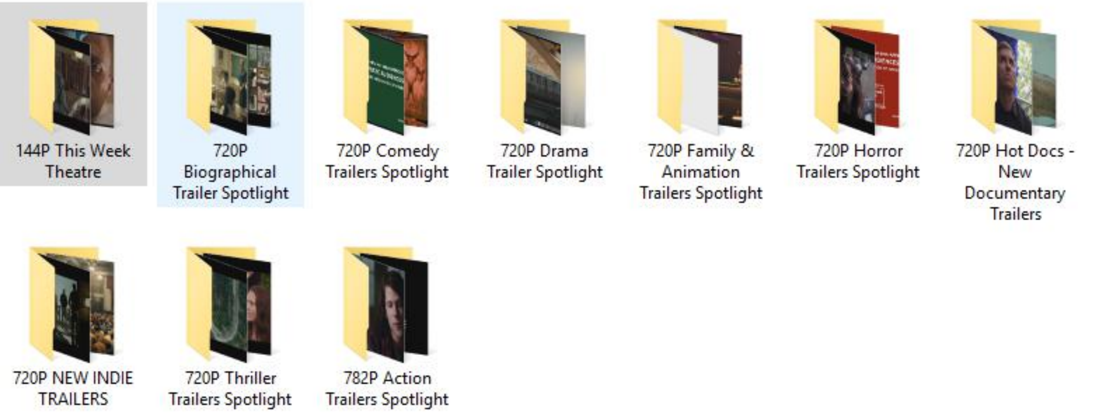
144P Low

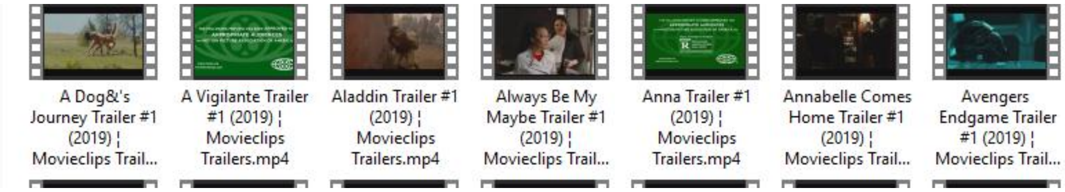
128P Kb/s

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.


144P This Week Theatre
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
782P Action Trailers Spotlight


A Dog&'s Journey Trailer #1 (2019) ¦ Movieclips Trail...
A Vigilante Trailer #1 (2019) ¦ Movieclips Trailers.mp4
Aladdin Trailer #1 (2019) ¦ Movieclips Trailers.mp4
Always Be My Maybe Trailer #1 (2019) ¦ Movieclips Trail...
Anna Trailer #1 (2019) ¦ Movieclips Trailers.mp4
Annabelle Comes Home Trailer #1 (2019) ¦ Movieclips Trail...
Avengers Endgame Trailer #1 (2019) ¦ Movieclips Trail...

- 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

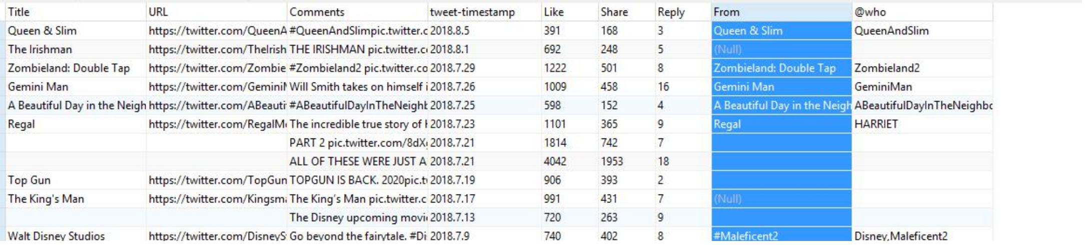
LikeCount

ShareCount

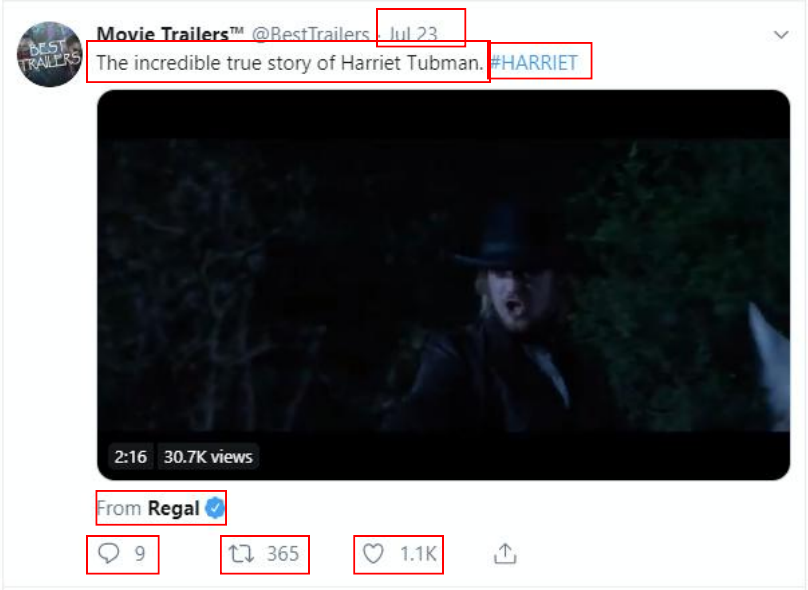
ReplyCount

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/GeminiM	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	#ABeautifulDayInTheNeighb	2018.7.25	598	152	4	A Beautiful Day in the Neigh	ABeautifulDayInTheNeighbc
Regal	https://twitter.com/RegalM	The incredible true story of H	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



Movie Trailers™ @BestTrailers · Jul 23

The incredible true story of Harriet Tubman. #HARRIET

From Regal

9 365 1.1K

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

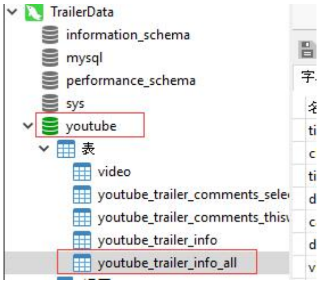
DislikeCount

FavoriteCount

commentCount

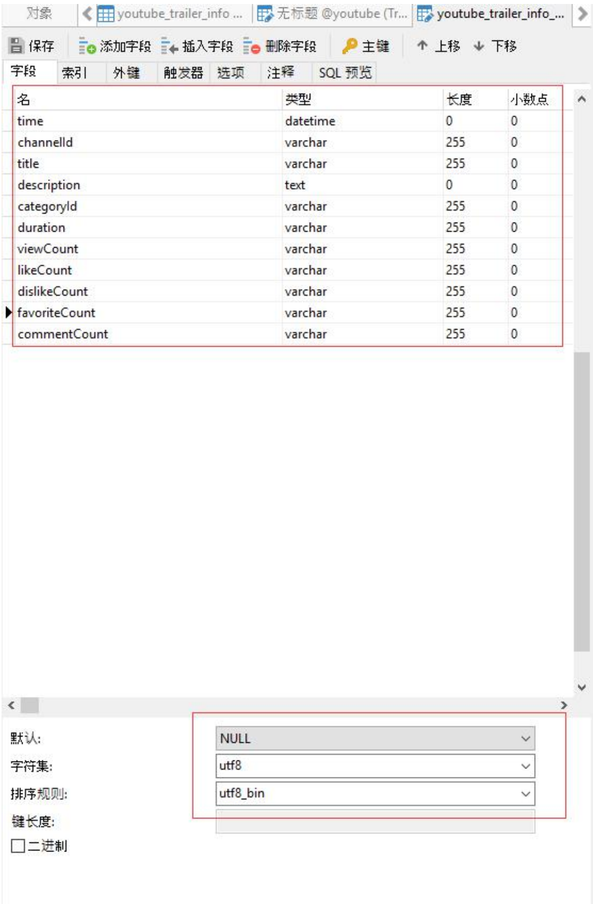
Steps:

1. Download Navicat, and Mysql, Build Connection in the Navicat.
2. Build a Database named youtube, build a datatable named youtube_trailer_info_all



- TrailerData
  - information_schema
  - mysql
  - performance_schema
  - sys
  - youtube
    - 表
      - video
      - youtube_trailer_comments_sele
      - youtube_trailer_comments_thisv
      - youtube_trailer_info
      - youtube_trailer_info_all

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



名	类型	长度	小数点
time	datetime	0	0
channelId	varchar	255	0
title	varchar	255	0
description	text	0	0
categoryId	varchar	255	0
duration	varchar	255	0
viewCount	varchar	255	0
likeCount	varchar	255	0
dislikeCount	varchar	255	0
favoriteCount	varchar	255	0
commentCount	varchar	255	0

默认: NULL

字符集: utf8

排序规则: utf8_bin

键长度:

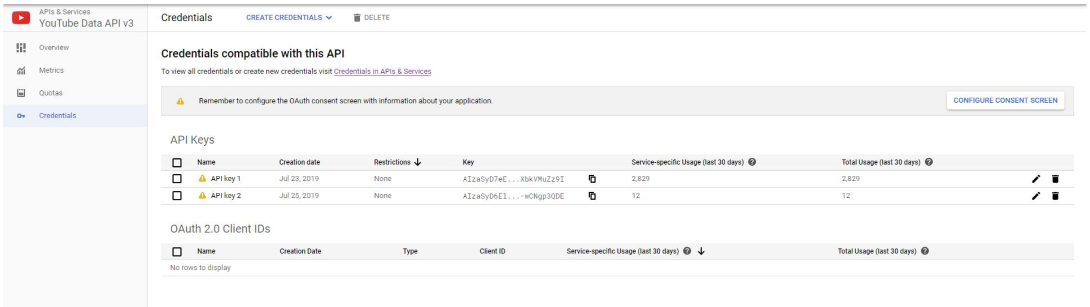
□二进制

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



Credentials

Credentials compatible with this API

To view all credentials or create new credentials visit Credentials in APIs & Services

Remember to configure the OAuth consent screen with information about your application.

API Keys

□	Name	Creation date	Restrictions ↓	Key	Service-specific Usage (last 30 days)	Total Usage (last 30 days)
□	API key 1	Jul 23, 2019	None	AIzaSyD7eE...XbkVMuZz9I	2,829	2,829
□	API key 2	Jul 25, 2019	None	AIzaSyD6El...-wCNgp3QDE	12	12

OAuth 2.0 Client IDs

□	Name	Creation Date	Type	Client ID	Service-specific Usage (last 30 days) ↓	Total Usage (last 30 days)

No rows to display

   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

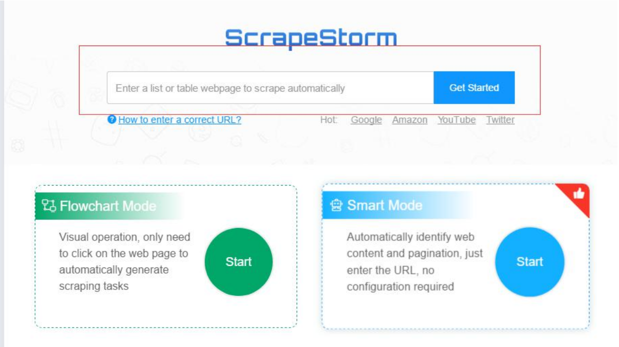
  CommentsLike(how many likes this comment get)

  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

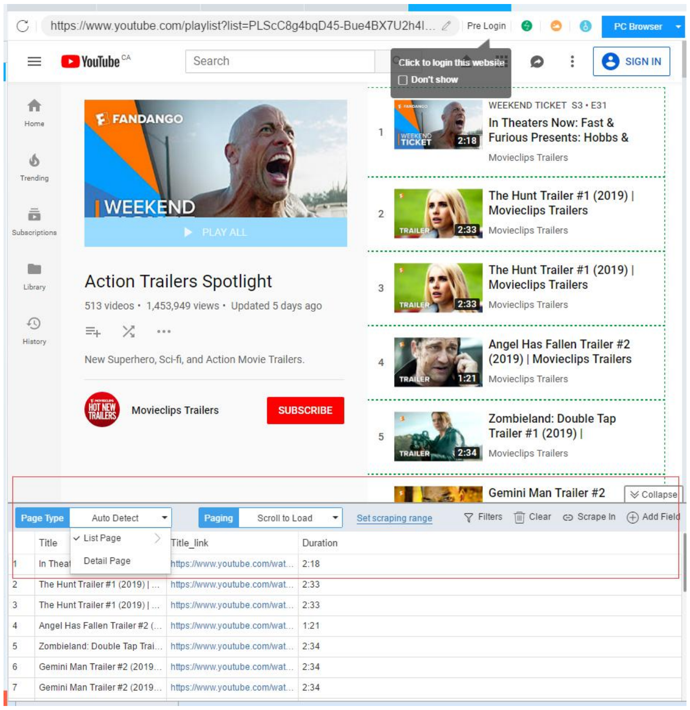
2. Copy the url(https://www.youtube.com/playlist?list=PLScC8g4bqD45-Bue4BX7U2h4IkzEV3hcL) and get started


ScrapeStorm
Enter a list or table webpage to scrape automatically
Get Started
How to enter a correct URL?
Hot: Google Amazon YouTube Twitter
Flowchart Mode
Start
Smart Mode
Start

Visual operation, only need to click on the web page to automatically generate scraping tasks

Automatically identify web content and pagination, just enter the URL, no configuration required

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



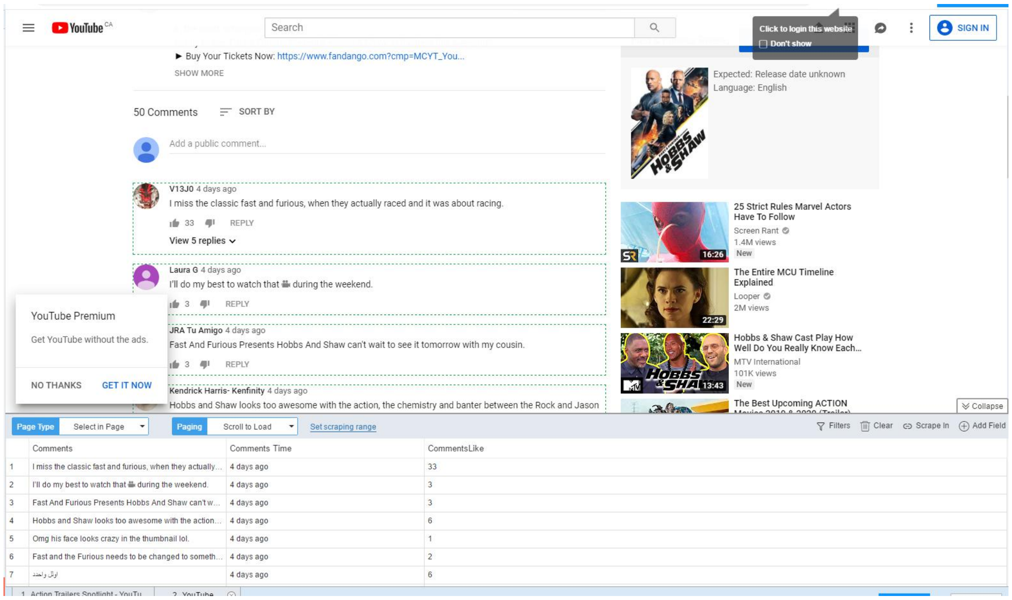
Page Type: Auto Detect (✓ List Page / Detail Page)

Paging: Scroll to Load

	Title	Title_link	Duration
1	In Theat	https://www.youtube.com/wat...	2:18
2	The Hunt Trailer #1 (2019) | ...	https://www.youtube.com/wat...	2:33
3	The Hunt Trailer #1 (2019) | ...	https://www.youtube.com/wat...	2:33
4	Angel Has Fallen Trailer #2 (...	https://www.youtube.com/wat...	1:21
5	Zombieland: Double Tap Trai...	https://www.youtube.com/wat...	2:34
6	Gemini Man Trailer #2 (2019...	https://www.youtube.com/wat...	2:34
7	Gemini Man Trailer #2 (2019...	https://www.youtube.com/wat...	2:34

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



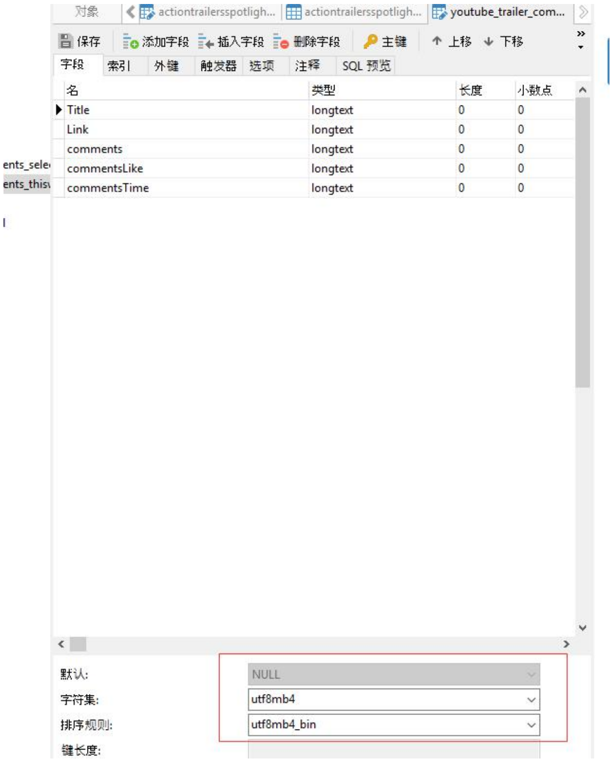
Page Type: Select in Page

Paging: Scroll to Load

	Comments	Comments Time	CommentsLike
1	I miss the classic fast and furious, when they actually...	4 days ago	33
2	I'll do my best to watch that 🎥 during the weekend.	4 days ago	3
3	Fast And Furious Presents Hobbs And Shaw can't w...	4 days ago	3
4	Hobbs and Shaw looks too awesome with the action...	4 days ago	6
5	Omg his face looks crazy in the thumbnail lol.	4 days ago	1
6	Fast and the Furious needs to be changed to someth...	4 days ago	2
7	اول واحدد	4 days ago	6

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)



名	类型	长度	小数点
Title	longtext	0	0
Link	longtext	0	0
comments	longtext	0	0
commentsLike	longtext	0	0
commentsTime	longtext	0	0

默认: NULL

字符集: utf8mb4

排序规则: utf8mb4_bin

键长度:

- 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

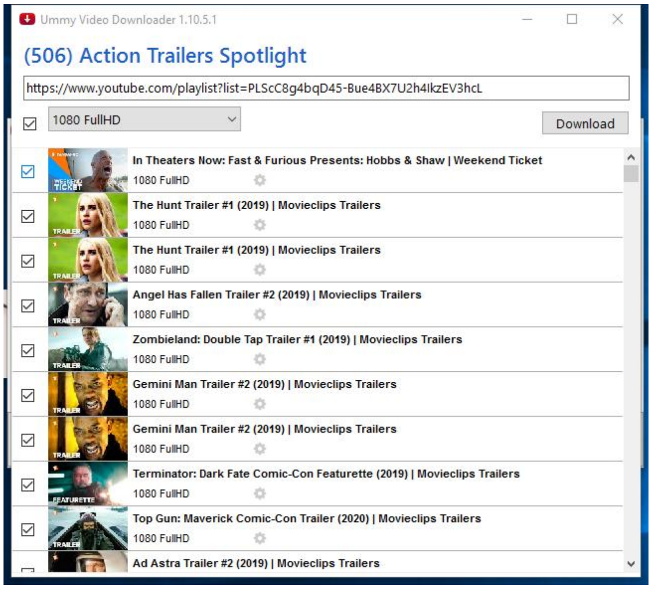
720P NEW INDIE TRAILERS

720P Thriller Trailers Spotlight

720P Action Trailers Spotlight

Steps:

1. Install Ummy Video Downloader
2. Copy URL()



Ummy Video Downloader 1.10.5.1

(506) Action Trailers Spotlight

https://www.youtube.com/playlist?list=PLScC8g4bqD45-Bue4BX7U2h4lkzEV3hcL

☑ 1080 FullHD Download

☑ In Theaters Now: Fast & Furious Presents: Hobbs & Shaw | Weekend Ticket — 1080 FullHD

☑ The Hunt Trailer #1 (2019) | Movieclips Trailers — 1080 FullHD

☑ The Hunt Trailer #1 (2019) | Movieclips Trailers — 1080 FullHD

☑ Angel Has Fallen Trailer #2 (2019) | Movieclips Trailers — 1080 FullHD

☑ Zombieland: Double Tap Trailer #1 (2019) | Movieclips Trailers — 1080 FullHD

☑ Gemini Man Trailer #2 (2019) | Movieclips Trailers — 1080 FullHD

☑ Gemini Man Trailer #2 (2019) | Movieclips Trailers — 1080 FullHD

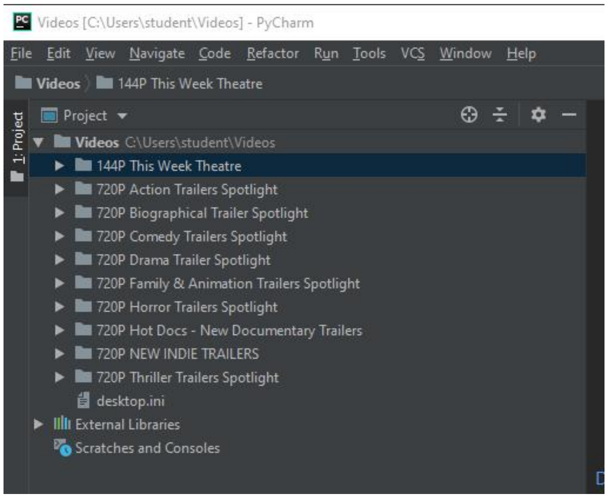
☑ Terminator: Dark Fate Comic-Con Featurette (2019) | Movieclips Trailers — 1080 FullHD

☑ Top Gun: Maverick Comic-Con Trailer (2020) | Movieclips Trailers — 1080 FullHD

Ad Astra Trailer #2 (2019) | Movieclips Trailers

3. Choose Quality and Download.

4. Open these videos in python when you need to do analyse(File->Open->Videos)



Videos [C:\Users\student\Videos] - PyCharm

File Edit View Navigate Code Refactor Run Tools VCS Window Help
Videos 〉 144P This Week Theatre
1: Project
Project

- Videos C:\Users\student\Videos
  - 144P This Week Theatre
  - 720P Action Trailers Spotlight
  - 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
  - desktop.ini
- External Libraries
- Scratches and Consoles

- 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” — All comments of about 10 trailers) 8,500+ rows

    (“This Week Theatre” — Selected comments of about 10 trailers) 800+ rows

    (One category “Action Trailers Spotlight” — 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.