What is the Trending on YouTube Videos?

https://github.com/jiaona9994/Trending_YouTube_Video_Analysis

1. Summary:

YouTube is one of the largest world-famous video sharing websites, and it maintains a list of top trending videos on the platform. Unlike popular videos, which already been classified as "Popular" with very high viewership numbers, trending videos are the ones that a wide range of viewers might find interesting in a very short period of time, generally, the trending list will roughly be updated for every 15 minutes. Therefore, the trending videos have the potential to become popular. Since YouTube's trending videos have the potential to be popular and were viewed by a wide range of audiences around the globe, getting insights into the trending videos will have an impact on the design and evaluation of personalization services such as precise advertising.

In this project, I study what are the general features of trending videos by applying Exploratory Data Analysis (EDA) on the whole dataset and Natural Language Processing on video titles and tags, then use the results to help the YouTubers to refine their video designing so that the videos can obtain a higher probability to become trending videos or even popular videos. On top of that, I also build several machine learning models on the dataset and compare across the models, and finally come up with the best model in my case.

2. Introduction:

YouTube is the world's largest video platform with millions of concurrent users every day and vast influence on customer behavior, beliefs, and opinions. Because of that, maximizing video performance has gained tangible economic value and many startups use it to gain traction and garner interest in their products and services.

I plan to use the dataset to analyze the composition and popularity associated with different factors of trending videos on YouTube and dig in deeper to elaborate on the relationship between them. To be more specific, some trending videos are highly controversial because of its content or types. In addition to that, what I am interested in are how the culture divergence affects viewer's likes, dislikes and the overall most popular video types.

The dataset I use for this project is “Trending YouTube Video Statistics” from Kaggle, which includes months of data on daily trending YouTube videos for the USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan, and India regions, with up to 200 listed trending videos per day. For language friendly purposes, I only use data from the USA, Great Britain, Germany, Canada, France, and Mexico. For each region’s data, it includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. I employ a number of techniques from the Scikit/Learn, Numpy, Pandas, Matplotlib toolkit in Python to analyze the dataset at hand.

3. Related Works:

There are a number of works conduct on “YouTube’s Trending Video Statistics”, such as Yanni Papadakis [1] focuses solely on the US videos dataset and aims is to provide systematic data preprocessing analysis working only with this part of the dataset. Yanni Papadakis [1] ‘s research suggests that the fastest a video makes it to the "trending videos" list the more likely it is for it to increase further in popularity, which is a little bit similar to the results I got. Luc Tremsal [2] ‘s work is mainly focused on data exploration using graphic library Plotly and predictive modeling. One of the differences between his work and mine is he made a lot of effort to study “comments”, and providing an enhanced disappointment result that, indeed, watching a video doesn't mean you'll like it. This feeling can be strengthened regarding trending videos when YouTube shows you a viral video and you may have high expectations, but most of the time the recommender system might let you down, therefore, there would be a much higher probability that you dislike it or leave a (negative) comment. For both [3] and [4], they are applying the Exploratory Data Analysis (EDA) approach on the US dataset to conclude what kind of attributes does a trending YouTube video has.

4. Measurement Methodology:

4.1 Exploratory Data Analysis (EDA)

Since I want to do something different from the related works I mentioned above, in addition to do EDA on the US dataset solely, what I am interested in is how the culture divergence affects viewer's likes, dislikes, and overall most popular video types. Therefore, I decided to the analysis not only USA data, but also Great Britain, Germany, Canada, France, and Mexico. The followings are how I do the analysis and the results I got.

I do the systematic data preprocessing analysis first, employing a number of techniques from the Scikit/Learn, Numpy, Pandas, Matplotlib toolkit in Python, since before concluding results from data I need to understand key data attributes, like missing values, unique counts, and time-series trends. I will spend most of my time analyzing the impact of publishing time, video genres, video title on the trending videos across different countries. Next, visualize the results extracted from the first step.

From figure 1, it is easy to find out that the publishing time of trending videos concentrate at 4 pm - 5 pm for a day, (if want your videos have a greater possibility to be trending videos, upload them between 4 pm and 5 pm) for almost all the countries we analyze. (This finding is not very applicable to MX, let's forget it for a while), next, it might be interesting to guess the reasons for this result. I will try to see if there are any relationships between the video category and the publishing amount. I will use heatmap to analyze this relationship. The results are just like figure 2, which tell us that in CA, US, and FR, it is best if YouTubers can upload Entertainment videos at 4 pm; In DE, it is best if YouTubers can upload Entertainment videos at 5 pm; However, in GE, it is best if YouTubers can upload Music videos at 5 am, upload Music videos at 4 pm is a good choice as well. From this point, it might be interesting to figure out why it is best to upload Music video in GE instead of Entertainment videos like other countries. I filter the dataset by "days to trending" and "country", and then get the total number of trending videos for each country. Results, figure 3, can be concluded as: On average, it takes much more time for videos uploaded in GB and the US to become trending. The reason might be the total amount of views (likes and dislikes) in these two countries are much higher than in other countries.

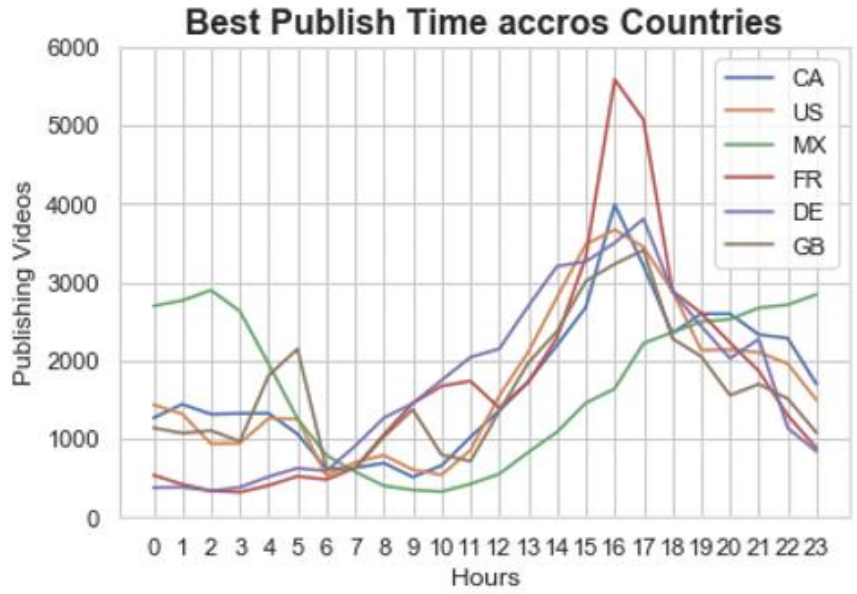

Figure 1

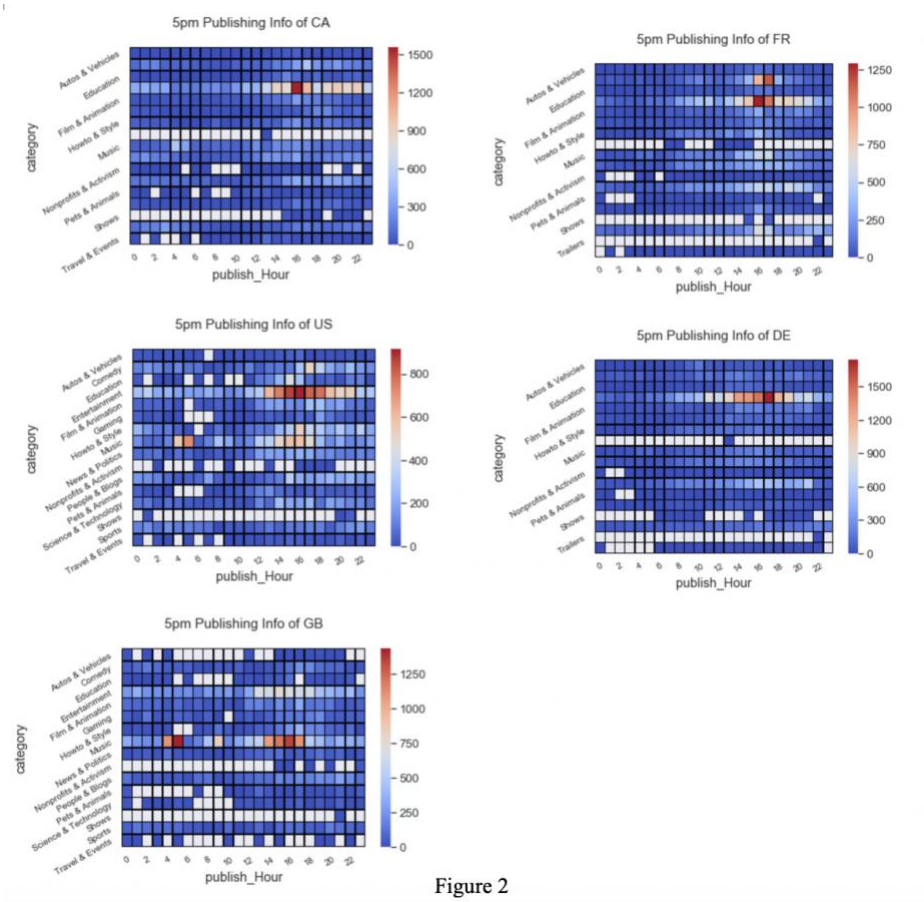

Figure 2

```
df.groupby(df['country'])['days_to_trending'].mean().to_frame()
```

country	days_to_trending
CA	3.481495
DE	1.850318
FR	2.800953
GB	36.762925
MX	1.921559
US	16.810423

```
df.groupby(df['country']).sum()
```

country	views	likes	dislikes	comment_count	comments_disabled	ratings_disabled	video_error_or_removed	trending_Year	trending_Month
CA	46891975069	1618179878	82137919	206161849	583.0	279.0	27.0	82488266	213977
DE	24645115205	893395538	57059031	113774380	1044.0	672.0	14.0	82405522	213715
FR	17100897444	708144090	33188528	74624804	889.0	704.0	22.0	82171437	213240
GB	230069198174	5234962944	296250384	509346351	683.0	272.0	69.0	78522895	204103
MX	13849692994	641627186	30223385	82506287	440.0	634.0	24.0	81620523	211664
US	96671770152	3041147198	151978155	345888164	633.0	169.0	23.0	82625482	214282

Figure 3

4.2 Natural Language Processing

Next, I apply Natural Language Processing on video titles and tags to figure out what kind of words have a higher frequency in the trending videos since an attractive title is the very first thing that will catch viewers eyes and adding tags is a good approach to attract viewers and might be helpful if we can get some insights on this. In this section of the analysis, I only use the US dataset and also use WordCloud to visualize the results. According to figure 4, we can see that the top three words are trailer, official, and new in the video title, with “trailer” and “official” having a great number of more occurrences than all other words. I think this might because viewers are very interested in “official trailer” and general have the curiosity about “new” things. From figure 5, we can see that words such as “vs”, “new”, “music”, and “makeup” have higher occurrences than other words in video tags. For “vs”, I think this might because viewers are more interested in watching videos focusing on comparison or competition. “music” and “makeup” might be used to attract music-lovers and people who love doing makeup.

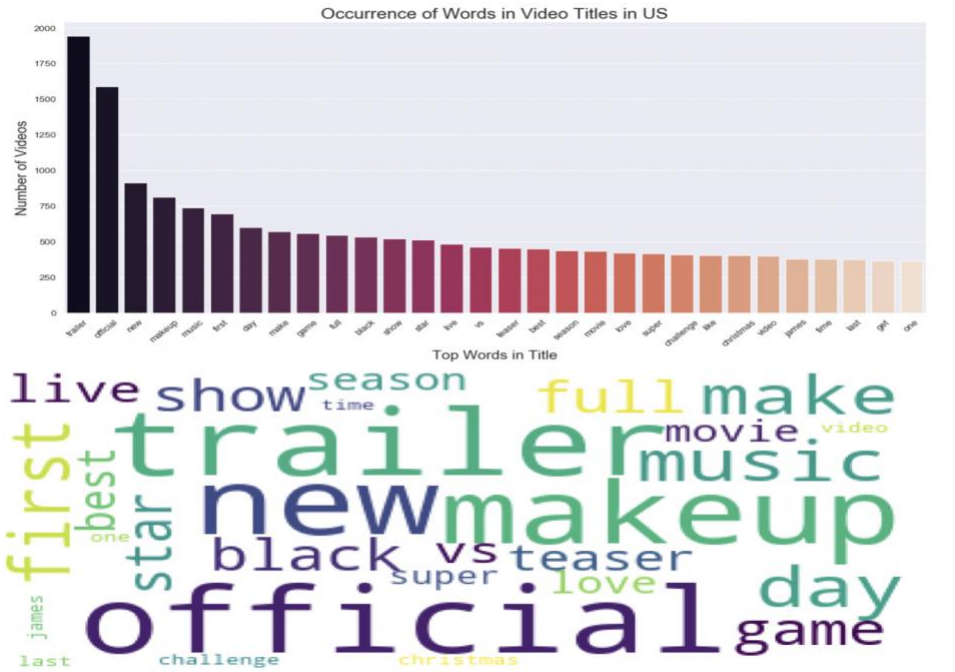

Figure 4

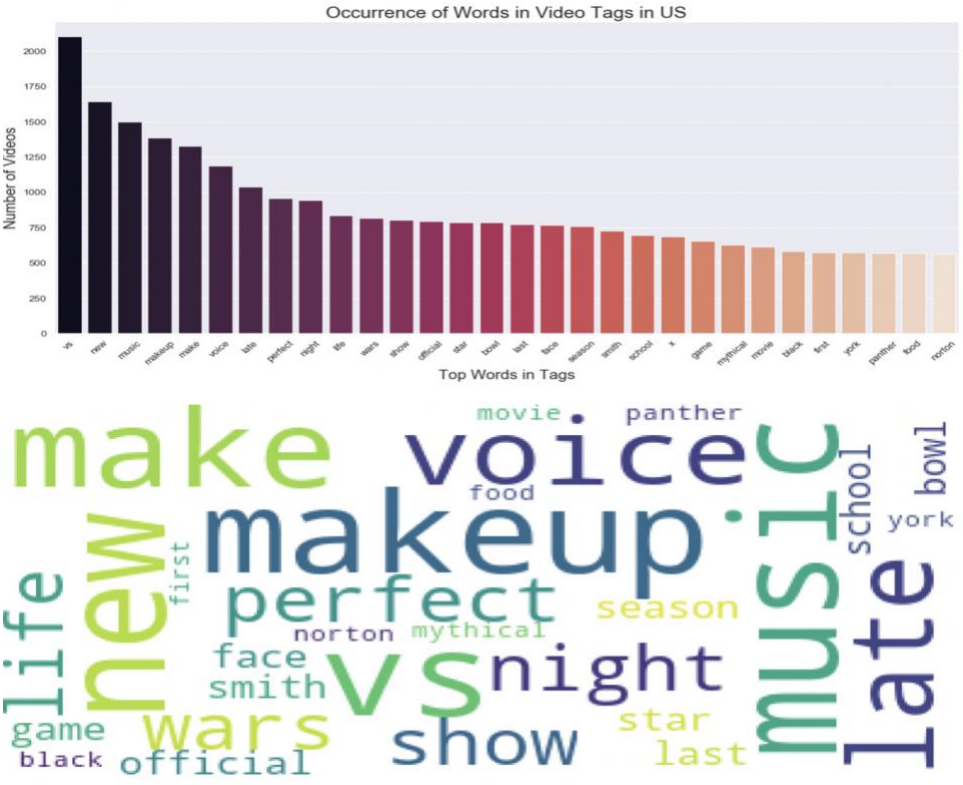

Figure 5

4.3 Machine Learning Models

I build machine learning models to predict the trending videos by splitting the existing dataset to the training dataset and test dataset. The models I am planning to use are Linear Regression, Ridge Regression, K-Nearest Neighbors, Decision Trees, and Random Forest. Since I train several classifiers for the same set of features, I evaluate the performance of those classifiers and chose a better one for each set of features. For example, as the number of views, likes, and dislikes can be used as the indicators of the "popular" videos, I use Linear Regression and Random Forest to predict those features individually and then conclude that which one is the better classifier for each feature. Since there are no separate test datasets available for me to evaluate models' performances, I will do the evaluation by seeing how well they model the available data.

Before starting to train the machine learning models, I did some data mining works again, such as changing all the features to numerical, implement matrix correlation analysis on all the features, and then drop the uncorrelated ones. Figure 6 is the correlation matrix, according to that, I dropped "category", "video_id", "trending_date", "trending_time", "publish_time", "tags", "title", "description", and "channel_title" columns as they are not correlated.

First, training models to predict the indicators of "popular" videos, which are views, likes, and dislikes. The classifiers I used for this step are Linear Regression and Random Forest, the results are like Linear Regression perform better on predicting views and likes, and Random Forest perform slightly better on dislikes, however, those two models both don't have a very low Mean-Square-Error value.

Second, training Linear Regression, Decision Tree, and K-Nearest Neighbors models to predict days to trending as days to trending can be used as the indicators of the engagement of videos. The observations for this step is very interesting. It is not surprising that Linear Regression fits will, Decision Tree accuracy is not bad as well, but the depth of the tree is a little bit high, there is a trade-off between the depth of the tree (the complexity of the model) and the prediction accuracy. KNN is the most interesting part, see figure 7, the accuracy of the model which uses the whole dataset is very low, however, when I created and tested KNN model by varying the number of trending days from 2 to 6, the accuracy varies. Due to the nature of using a classifier, random guessing would give an accuracy of 50% (0.5). We see that the accuracy drops to a minimum of around 50% when using prediction for at most 6 trending days. This tells us that trying to classifier

if a video will trend at most 5-7 trending days is the hardest as the accuracy becomes close to random guessing.

Finally, apply Linear Regression and Ridge Regression to predict the number of tags as using good tags will have a positive impact on video performance. We have already gotten knowledge of what kind of words should be included in the tags of a trending video, and it is time to analyze the number of tags for trending videos. Since our analysis contains multiple relevant columns, a straightforward linear regression might not be sufficient to detect or show a relationship between the variables. With this, Ridge Regression might possibly serve as a better algorithm for linear relationships between multiple variables. Ridge Regression is a technique for analyzing multiple regression data and understanding the multicollinearity between various variables. Multicollinearity is the existence of near-linear relationships among the independent variables. It can be concluded from the results that the performance of Ridge Regression is better than simple Linear Regression.

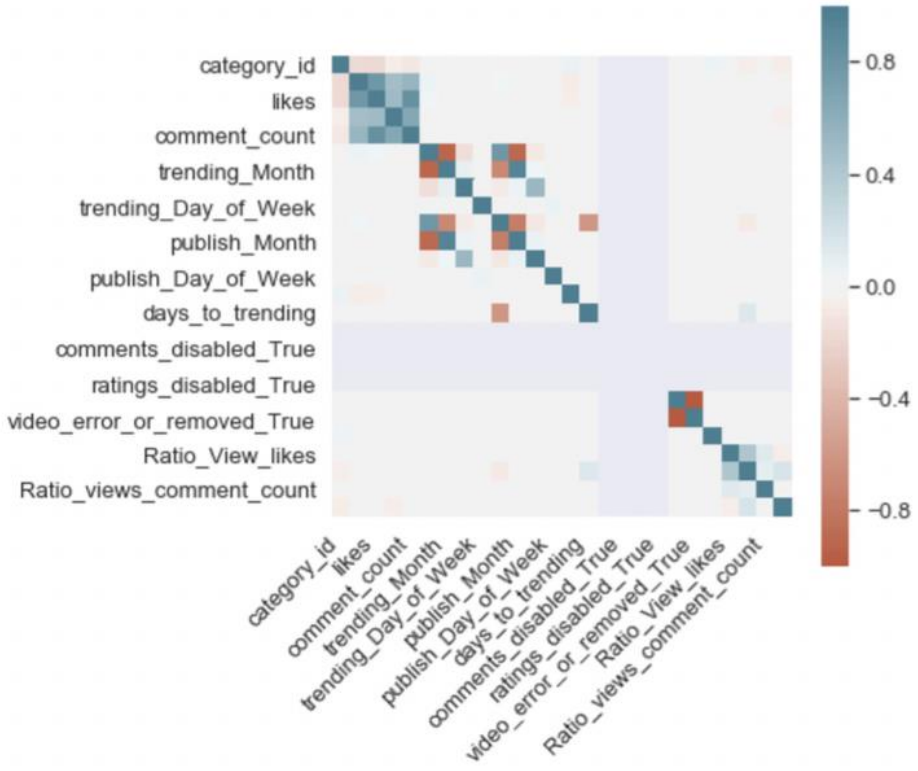

Figure 6

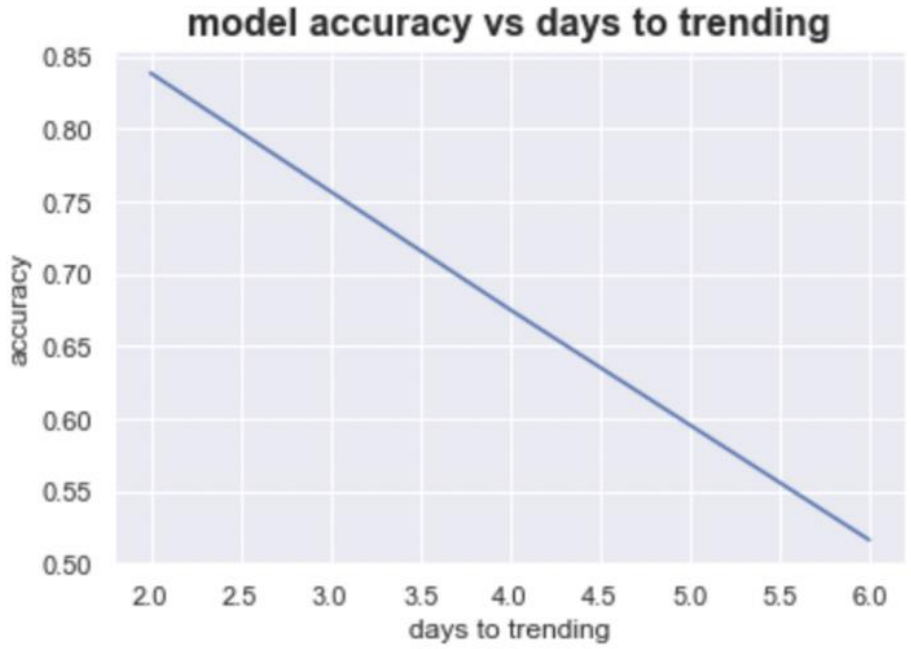

Figure 7

5. Conclusion and Future Works:

5.1. If YouTubers want their videos to have a greater possibility to be trending videos, upload them between 4 pm and 5 pm.

5.2. In CA, the US, and FR, it is best if YouTubers can upload Entertainment videos at 4 pm; In DE, it is best if YouTubers can upload Entertainment videos at 5 pm; However, in GE, it is best if YouTubers can upload Music videos at 5 am, upload Music videos at 4 pm is a good choice as well.

5.3. On average, it takes much more time for videos uploaded in GB and the US to become trending.

5.4 The total amount of views (likes and dislikes) in GB and US is much higher than in CA, FR, AND MX.

5.5 The top three words in trending videos' titles are a trailer, official, and new, with "trailer" and "official" having a great number of more occurrences than all other words. 5.6 Words such as "vs", "new", "music", and "makeup" have higher occurrences than other words in video tags.

5.7 When predicting "view"', "likes", and "dislikes", Linear Regression is a useful model.

5.8 When predicting "days to trending", all the models perform not bad after I normalize the dataset. An interesting finding is that when applying KNN, the accuracy of the model which uses the whole dataset is very low, however, when I created and tested KNN model by varying the number of trending days from 2 to 6, the accuracy varies. Due to the nature of using a classifier, random guessing would give an accuracy of 50% (0.5). We see that the accuracy drops to a minimum of around 50% when using prediction for at most 6 trending days. This tells us that trying to classifier if a video will trend at most 5-7 trending days is the hardest as the accuracy becomes close to random guessing.

5.9 When predicting "Tags count", Ridge Regression gives us better results than simple Linear Regression because of the features we used to analyze are relevant to each other.

Since the performance of the models predicting "views", "likes", and "dislikes" is not very good, and I didn't compare the differences between a wide range of normalization approaches, for

future works, these two basic shortcomings of my analysis can be improved. In addition, one of the futures works we can do is using the YouTube API to do the data crawling by ourselves. There are two main reasons I want to do that: one is that data obtained by data crawling is much more updated, and size and features of the dataset can be controlled by our determination, the other reason for that is we can try to analysis a new dataset which is full of non-trending videos, and compare the results to the trending videos dataset. In this way, the suggestions made to the Youtubers could be much more meaningful.

6. Efforts made to the project:

From this project, I learned the process of doing a data analysis project, analyzing the related works, and how to implement what I learned in class to real-world applications. In my opinion, I think I have already possessed the skills to do Exploratory Data Analysis, Natural Language Processing, WordCloud, training different machine learning models to fit the dataset, and for any given model, using proper packages or approaches to tuning parameters (I did this when train the Decision Tree, Random Forest, and K-Nearest Neighbors classifiers). Moreover, through this project, my understanding of Scikit/Learn, Numpy, Pandas, Matplotlib toolkit in Python is much deeper, besides those libraries, I studied how to use NLTK to implement Natural Language Processing, which is very interesting.

7. Bibliograph:

Data Content: (from Kaggle)
It includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.

EDIT: Now includes data from RU, MX, KR, JP and IN regions (Russia, Mexico, South Korea, Japan and India respectively) over the same time period.

Each region's data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

The data also includes a category_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

Related Works References:

[1] https://www.kaggle.com/yanpapadakis/trending-youtube-video-metadata-analysis
[2] https://www.kaggle.com/lepuppy/youtube-trending-videos-interactive-eda/notebook
[3]https://www.kaggle.com/quannguyen135/what-is-trending-on-youtube-eda-with-python/notebook
[4] https://www.kaggle.com/mlenzovet/newpublish-time-strange-distribution/