IST 652 Final Project

Authors: Gary Chen, Mohamad Nayal

September 12, 2021

Trending YouTube Videos Exploration

**Topic:**

YouTube is the second most popular website. In 2017, YouTube announced that its users were watching more than 1 billion hours (about 110,000 years) of videos. In 2021, there are more than 2.2 billion users of YouTube. With a platform this large that has the potential to reach so many users we wanted to explore what are the features of the videos get the most attention.

**Data**:

The primary data source for this project is the “Trending YouTube Video Statistics” <https://www.kaggle.com/datasnaek/youtube-new/version/115?select=GB_category_id.json>. This dataset gives us information about the top trending YouTube videos from 10 countries: Canada, Germany, France, Great Britain, India, Japan, South Korea, Mexico, Rusia, and the United States.

Data about the Gross National Income of each of these countries was obtained from a secondary dataset. The Kaggle dataset “Income by Country” <https://www.kaggle.com/frankmollard/income-by-country> provided this Gross National Income or GNI information.

Our final source of data was created by pulling the official language for each country listed on Wikipedia

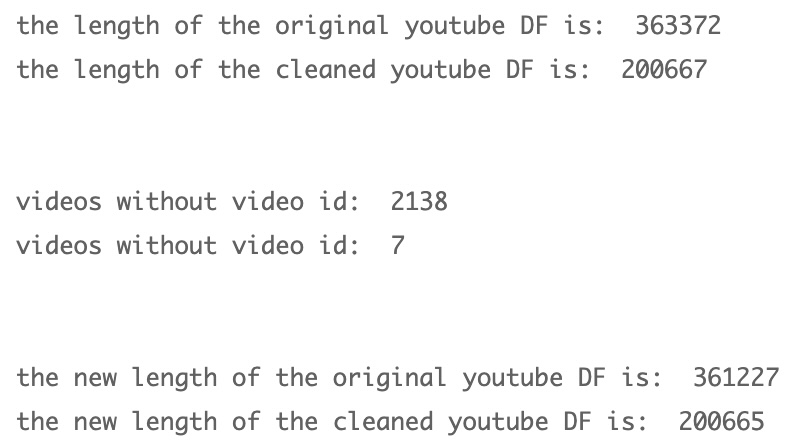
<https://en.wikipedia.org/wiki/List_of_official_languages_by_country_and_territory.> This data was used to group countries that speak the same language together.

**Data Cleaning, Exploration, and Transformation:**

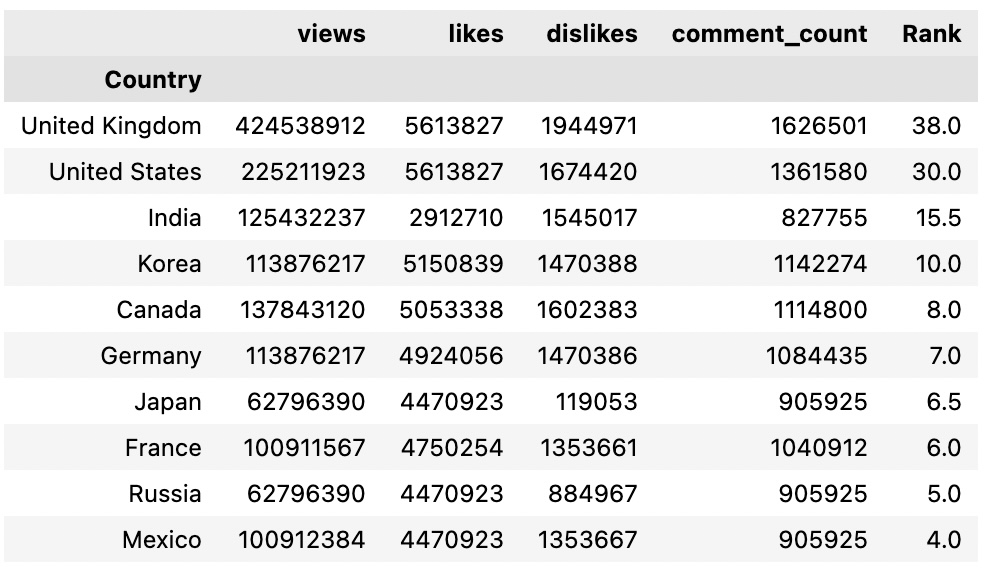
The trending YouTube video dataset was originally formatted as 10 csv files, one for each region and 10 json files also one for each region. The csv files contain information about the trending videos including: video\_id, trending\_date, title, channel\_title, category\_id, publish\_time, tags, views, likes, dislikes, comment\_count, and description. The category column of each of these dataframes contained just an integer. The json file for each of these regions only contained information about the video category id for each region. Most regions had similar video categories and category ID’s but there were some regional differences in trending YouTube categories. A dictionary from each category id and category name was created from each region. Then these values were mapped onto category id column of each imported dataframe from the csv files. A region and country column were added to each dataframe so that these dataframes could be combined. We also added Region and Country columns to denote the country source of each dataframe.

We noticed that some videos were trending for multiple days in a row which resulted in the same video appearing multiple times in our dataframe. A function was created to partition a dataframe to group rows by their video ids to identify the latest video entry as rank 1. This was done by ranking the trending date in descending order. The function also returns a new column with the rank score for the trending date with rank 1 as the latest trending video. The function was run on all 10 dataframes and a new column was created for each using this function. Then the 10 dataframes were concatenated. A new dateframe(vidsDF) was created to contain only the latest entry of a trending video per country by filtering on the concatenated dateframe (OG\_vidsDF) that had a rank score equal to 1. The videos without video IDs were also removed because we could not be sure if they were the same video as one with a matching title or another video with an identical title when no unique identifier was available. To get an idea of how often a video was trending for multiple. Removing instances of videos that trended for multiple days decreased our dataframe from 363,372 records to 200,667 records.

After cleaning our main dataset, we wanted to bring in language and the 2018 Gross National Income per capita for each country. We chose 2018 GNI because our YouTube data points ranges from November 14, 2017 through June 14, 2018. We scraped the Wikipedia page of the list of official language with the read\_html function and then build a dictionary for languages. Then we cleaned up the formatting in a list format and created a dataframe (languageDF) for the official language spoken in each country. Then we mapped and created the Language column in our main datadrame (vidsDF). After importing the 2018 GNI csv, we merge it with the vidsDF based on country name. The df\_merge dataframe is what we used for most of the exploration and insights.



We were interested to see if the max amount of time a video trends is influenced by the Country it was posted in. The table below shows that the country with the longest trending video was the United Kingdom. The country with shortest max trending time was Mexico.



The other columns here are all the max values for their respective columns and are not related to one another. The video from the United Kingdom with that has over 424 million views may not be the video with over 5 million likes. We wanted to continue to investigate the videos that trended the longest time from each country. Those results are below



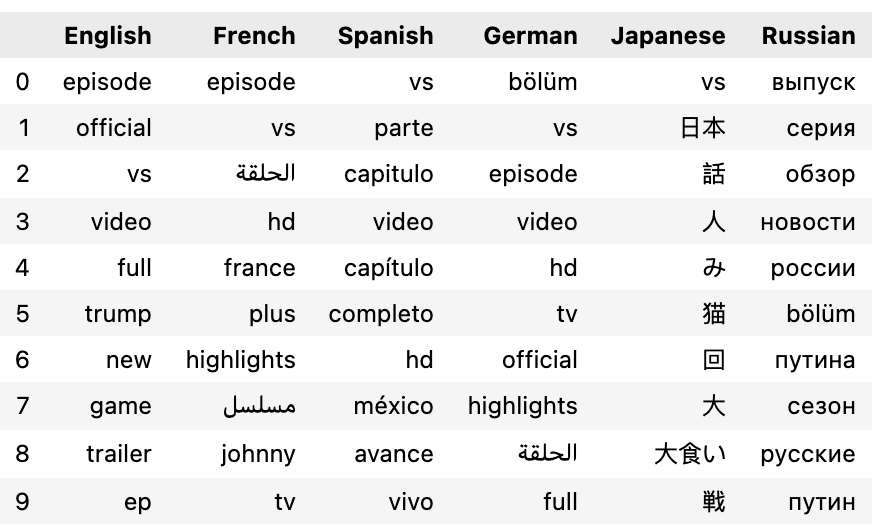
We identified the longest trending videos in our dataset from each country. In the case where more than one video was trending for the maximum amount of time, we returned the video with the more recent trending date. We noticed most of the longest trending videos by country are music videos.

**Questions:**

**Question 1: What are the most used words used in the titles of trending videos?**

To tackle this first question we decided it would be beneficial to group country’s that speak the same language together. We originally tried to pull data on the official language of each country from the UN.org but after some time decided to pull language from the Wikipedia about the UN data on the official language of each country. A dictionary was created to that listed each country and its official language. The language of each of the country for each video was mapped onto the master dataframe from the dictionary. The nltk.tokenize package was used to create a list of words from the titles of each video in the same language speaking. A dictionary was then created with the counts of how many times the word appeared and finally a sorted dataframe was created with the most used words for trending video titles in by language group. The sent\_tokenize was used to eliminate stop-words from each language. All punctuation and numbers were removed as well. For Japanese, fugashi package was used to create tokens of the Japanese text. Japanese stopwards were filtered out using the following GitHub repository <https://github.com/stopwords-iso/stopwords-ja/blob/master/README.md>

The top 10 words for each language group are below.



The above table was translated using Google Translate. It was discovered in all 6 regions had English words in the titles of top trending videos, the French speaking region had Arabic word, the German speaking region had Arabic and Turkish words, and the Russian speaking region has Turkish words.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| English | French | Spanish | German | Japanese | Russian |
| episode | episode | vs | episode | vs | release |
| official | vs | part | vs | Japan | series |
| vs |  | episode | episode | Talk | overview |
| video | hd | video | video | Man | news |
| full | france | episode | hd | Fruit | Russia |
| trump | plus | full | tv | Cat | episode |
| new | highlights | hd | official | Times | путина |
| game | johnny | méxico | highlights | Big | сезон |
| trailer | Tv show | trailer | episode | Gluttony | русские |
| episode | tv | live | full | Battle | путин |

The token list from each of the language groups was used to create a word cloud using the wordcloud package. For Japanese, the MeCab package was also required.

|  |  |
| --- | --- |
| English | French |
| Spanish | German |
| Japanese | Russian |

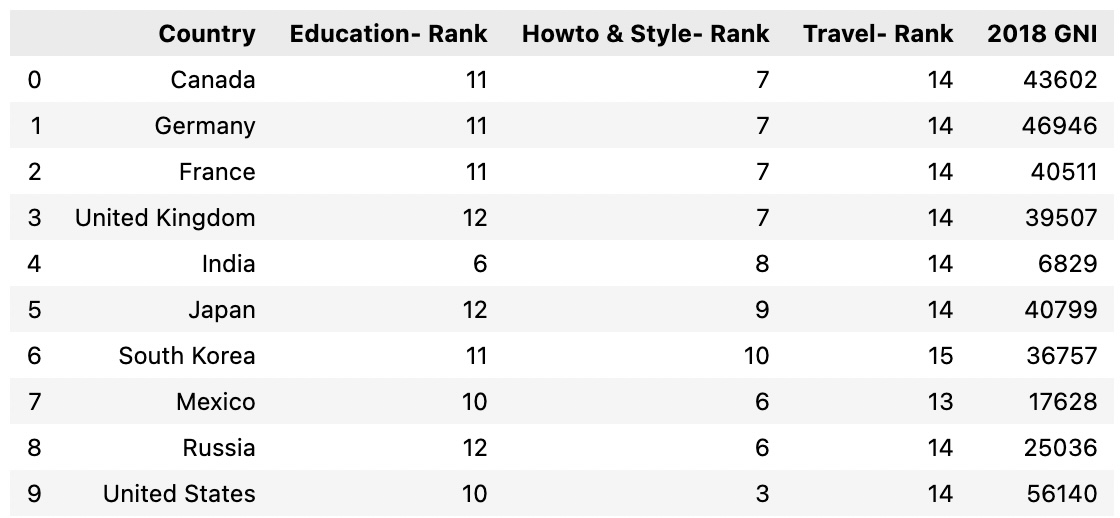
From the French word cloud we can see that the musician Johnny Hallyday was trending during this time. Hallyday passed away within the timeframe of this data which might have caused him to be trending. From the Spanish word cloud, we can see the professional soccer team Cruz Azul was trending as well as liga or league. The Spanish word cloud also shows the Amlo, a former Mexican president, who died in 2018 was trending during that time. From the Japanese word cloud, we can see that Japanese has beautiful aesthetic.

**Question 2: What categories are the most popular for each country?**

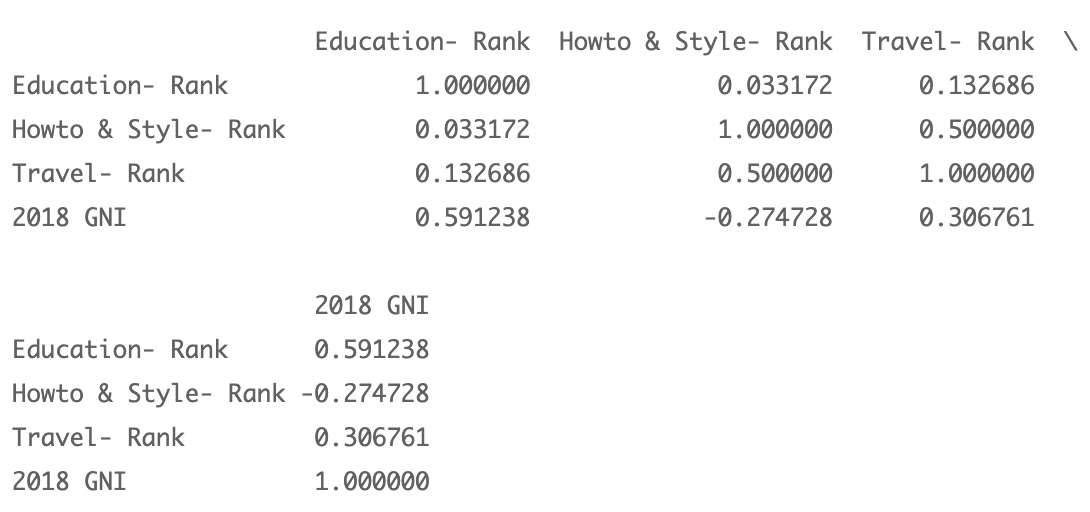
Below are the overall category counts in our dataset.



Entertainment, People & Blogs, News &Politics, Sports, Music, and Comedy were the usual top 6 trending categories for the 10 countries in our dataset. There were some outliers such as Education being ranked 6th in India, Pets & Animals ranked 5th in Japan, and Howto & Style ranked 3rd in the United States. We wanted to explore some of these outlier categories and get additional insight into the non-top 6 most popular categories. We chose Education, Howto & Style, and Travel categories to do a comparison with each country’s 2018 Gross National Income (GNI) to answer our next question. Below is a data frame which contains our 10 countries, their 2018 GNI, and their respective ranks of the Education, Howto & Style, and Travel categories.



**Question 3: Does the wealth (GNI) of a country influence what type or category of videos trend?**



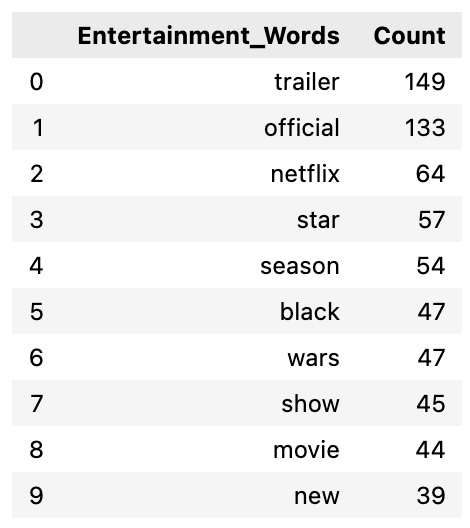
We did a correlation analysis with the data frame above. Since a higher rank score means it is weighted lower, we should view rank values as inverse. The –0.275 magnitude between Howto & Style and 2018 GNI means that Richer countries will tend to watch more Howto & Style videos. The Education and 2018 GNI 0.59 magnitude shows that poorer countries will watch more educational videos. Travel videos are ranked extremely low across all countries, so it is hard to get insight on if GNI affects Travel videos. Overall, the most popular videos are in the top 6 categories, but it is interesting that GNI can have influence on the smaller categories.

**Question 4: What combination of words in what category would help a title to trend globally?**

We wondered if we wanted to exploit what we now know about the YouTube trending video behavior what type of video would we create and how would we title the video. We discovered that the number one category of trending videos across all regions was ‘Entertainment’. A made a list of all the titles across all regions that were under the video category of Entertainment. These words were then tokenized and ranked in a similar method as in question 1. The top 10 words found in the titles of trending Entertainment videos are below.



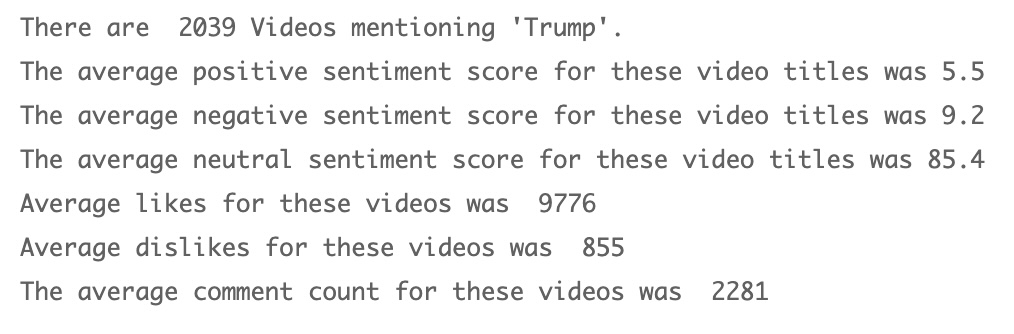
The Arabic words in the list translate to ‘episode’ and ‘tv program’ and the Turkish word also translates to ‘episode’. From these results it seems people around the world rely on YouTube to either stream tv shows or discuss fan theories about popular tv shows. Out of curiosity we decided to repeat this process but only look at videos in the entertainment category from the United States. Those results are below.



This table might indicate that people in the US are more interested in movies than tv shows. Interestingly Star Wars may have been trending at the time this data was collected in late 2018.

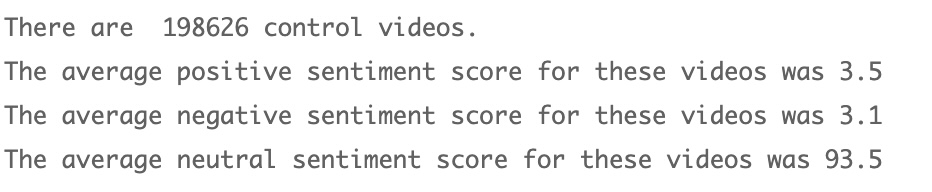
**Question 5: Do more trending videos contain a positive or negative image of Trump?**

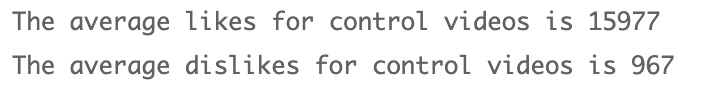
‘Trump’ appeared in more than one of our word clouds. He is often described as a divisive politician so we wanted to know if we could see any interesting patterns in the way he was talked about in the titles of videos that he was mentioned in. A separate dataframe was created that contain all videos where the title mentioned ‘Trump’ in any country. A sentiment analysis was conducted to see the mean positive and negative score for each video. The mean of likes, mean dislikes, and mean count of comments was calculated. The output is below.



Our analysis shows that the mean negative score is nearly twice as high as high as the positive score. However, both scores are low at 5.5 and 9.2. The most surprising piece of data was that the neutral sentiment for these videos was over 85%. We expected the sentiment analysis to be close to 50% positive and 50% negative but this was far from the case. One potential reason for this is many news videos which may be required to be titled with a neutral tone. This made us wonder how the sentiment analysis of all other videos in our data frame looked by comparison.

The read out from the comparison videos is below.





Based on these results videos mentioning ‘Trump’ have titles with a lower neutral score than other videos but further analysis needs to be done to elucidate the difference.

**Conclusions:**

We were successfully able to pull both semi structured and structured data in from two Kaggle datasets and from multiple Wikipedia webpages. We combined this data into one cohesive dataset to learn about features of trending YouTube videos. We learned that videos in the United Kingdom and the United States trend significantly longer than anywhere else in the world. Each has trending videos that continued to trend for twice as long or more than the rest of the world. We learned that the most used words in the titles of trending videos include words used to describe tv series such as ‘episode’.

We learned of a Mexican politician who was trending, a French musician who was trending, a Mexican soccer team that was trending, and we saw that former president Trump was trending. We learned that the ranking of video categories is remarkably similar across all countries, but it is not identical. The GNI of country is positively correlated with how much HowTo and Style content is watched and negatively correlated with how much Educational content is watched on YouTube.

Creating a video in the Entertainment category and using the word episode in the title is associated with trending more. In the US using ‘Star’ ‘Wars’ in the title of an entertainment category video is more likely to make it trend. Videos with “Official” and “Episode” in its title trend to be more popular and success in all countries. We commonly see the words official, episode, trailer, versus (vs. / v.), and part in many languages.

Finally, we learned that most of the sentiment in the title of videos mentioning ‘Trump’ is neutral but there is twice as much negative sentiment as positive sentiment videos. The average likes for these videos is 10 fold higher than the dislikes. The ratio of likes to dislikes in videos about Trump is lower than the ratio of likes to dislikes in non-Trump videos. There are many likes for Trump videos but compared to the control group, it has proportionally more dislikes

**Team member contributions:**

Both authors contributed to importing the data from various sources and cleaning the data to produce a cohesive data structure that combines all the data. Mo worked on questions 1, 4, 5 involving all the title parsing, word clouds, and sentiment analysis. Gary worked on questions 2 and 3 to identify top video categories by country, looking at how GNI affects categories in each country and identified longest trending videos in each country.