Analysis of a YouTube trending video dataset

Alberto Ranz
Ironhack Data Analysis
FT RT May 2021

1.- Introduction

YouTube is the biggest online platform for video sharing. It was founded in 2005, and since then it has grown massively. In 2020 it had 500 million users only in India, followed by the 200 million users in USA. Around 2.3 billion people in the world access YouTube monthly, which is equivalent to around a third of the global population. The most subscribed channel (T-Series) belongs to the music category and had 176 million subscribers as by Feb 2021. The platform generated $19.7 billion revenue in 2020 (30.4 % yearly increase).

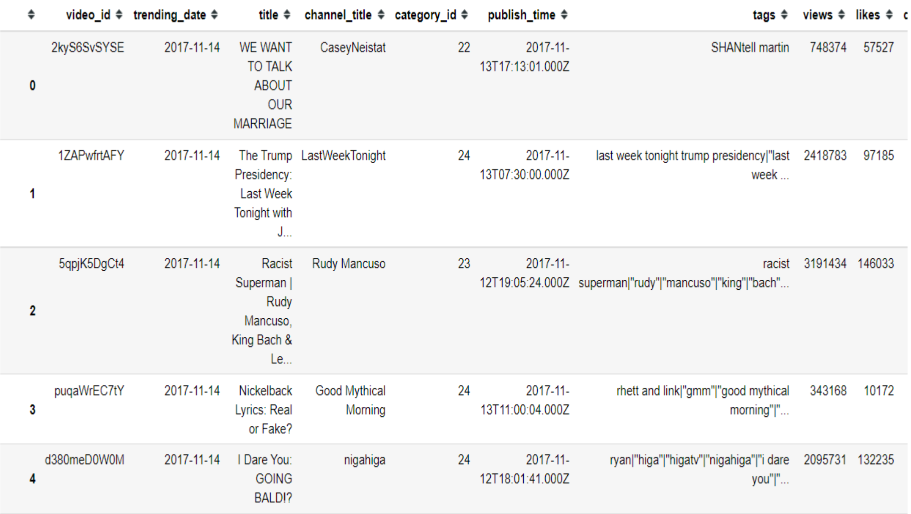
In the Trending section, the users can find a daily selection of videos that YouTube considers relevant. The exact way the platform selects these trends remains unknown, and it has sometimes been critisized for showing controversial content. But according to Google, it follows some rules. Trending tries to present videos that are: attractive to most users, non-polemic, topical, diverse and surprising. For this purpose, it takes in account the number of views and the speed of their growth, the publication date, a comparison with other videos of the same channel, and many more.

This analysis can provide some insight on how YouTube Trending behaves. The conclusions drawn could help to explain not only the functioning of the platform, but also the ways in which people in each country make use of it.

2.- Dataset

The data used for this project was taken from Kaggle, an internet site for dataset sharing, and consist on 10 csv files (one for each country) and another 10 json files which contain the category names and some extra information.

Interesting information for this analysis is: trending date, publication date, category id, views, likes, dislikes and number of comments. Other columns could be also to consider, depending on the kind of analysis to perform and the tools avaliable.



	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	c
0	2kyS6SvSYSE	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13T17:13:01.000Z	SHANtell martin	748374	57527	
1	1ZAPwfrtAFY	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13T07:30:00.000Z	last week tonight trump presidency|"last week ...	2418783	97185	
2	5qpjK5DgCt4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146033	
3	puqaWrEC7tY	2017-11-14	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-11-13T11:00:04.000Z	rhett and link|"gmm"|"good mythical morning"|"...	343168	10172	
4	d380meD0W0M	2017-11-14	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	2095731	132235	

Fig. 1: An example of a table of the dataset. Each table having 16 columns.

The data was scraped from the YouTube API, and includes trends from November 2017 to June 2018, making six months in total. Because of this, it was not possible to perform analysis based on the time of the year. Some of the videos were trend repeatedly for several days, therefore a cleaning of the tables had to be done to avoid errors.

3.- Goals

The main goal of this analysis is to obtain reliable information about YouTube. Although this dataset does not allow to obtain highly significant conclusions, we can partially observe the conditions that a video must fulfill in order to be selected as trend. We can also see which categories are more likely to be trend, as well as the ones more outstanding in view counts, likes, dislikes and comment number. This can give a small hint of which categories are overall more important in the platform and the ones that users prefer the most.

4.- Analysis

This part of the project could be divided into three parts:

4.1.- Preparation and Data Cleaning: as some of the videos appear several times in different rows, new data frames were created, which dropped the repeated rows for same video id, only keeping the first one of each. By making copies of the data frames, no information was lost. Other data frames were created containing the specific columns needed for the later plotting in Tableau. This is something to consider, because working with such big tables in Tableau needs great amounts of time for computing, and preparing the data beforehand makes the task easier and faster.

4.2.- Coding: after arranging all the tables, the main calculations can be done. The principal ones were sum of columns filtering by categories and other columns, mean calculations, median calculations, and other basic operations. In the case of the days that videos took from publication to trend, some tables had big outliers. The USA table had trending videos with several years of age, which means that the average is not the appropriate measuring. Therefore, a calculation of the median was performed, achieving more homogeneous values between all the countries.

4.3.- Plotting: the plots were created mainly in Tableau. Some plots were done as a test with matplotlib but due to its own limited possibilities, they could not display the information clearly, except for the correlation matrix. This plot shows the correlation level between the main variables: views, likes, dislikes and comments (Fig.2). The principal idea that could be deducted from this graph is that views and likes are highly correlated (0.76), as well as likes and comments (0.81).

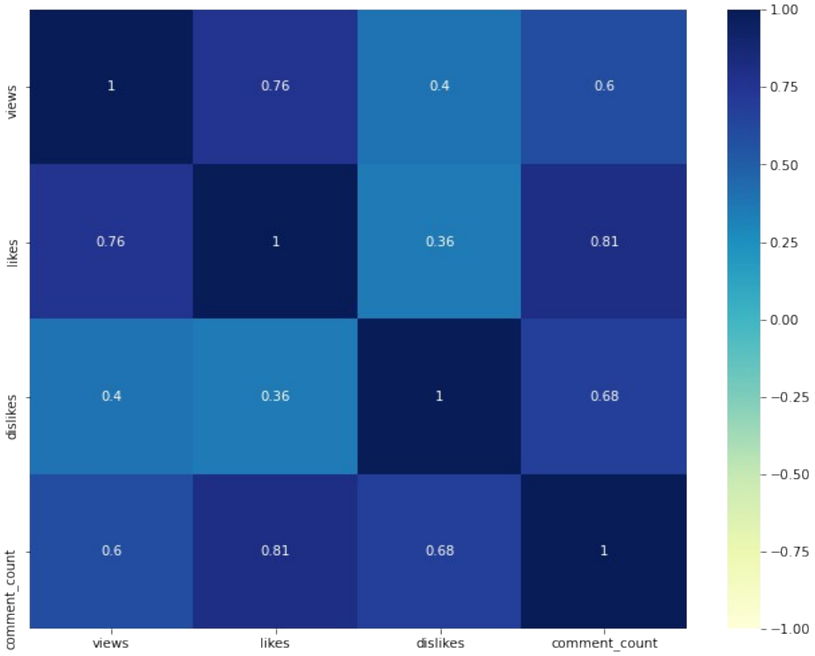


Fig. 2: Correlation Matrix

The views bar chart (Fig. 3) indicates that the most viewed category is 24. Entertainment, followed by 10. Music. This is actually a logical conclusion. Most people use the platform for entertainment purposes, or for listening to music and watching musical videos. Within the leading category, Canada is surprisingly the country where most views come from. For the music category, Great Britain has the bigger amount of views, followed by Canada.

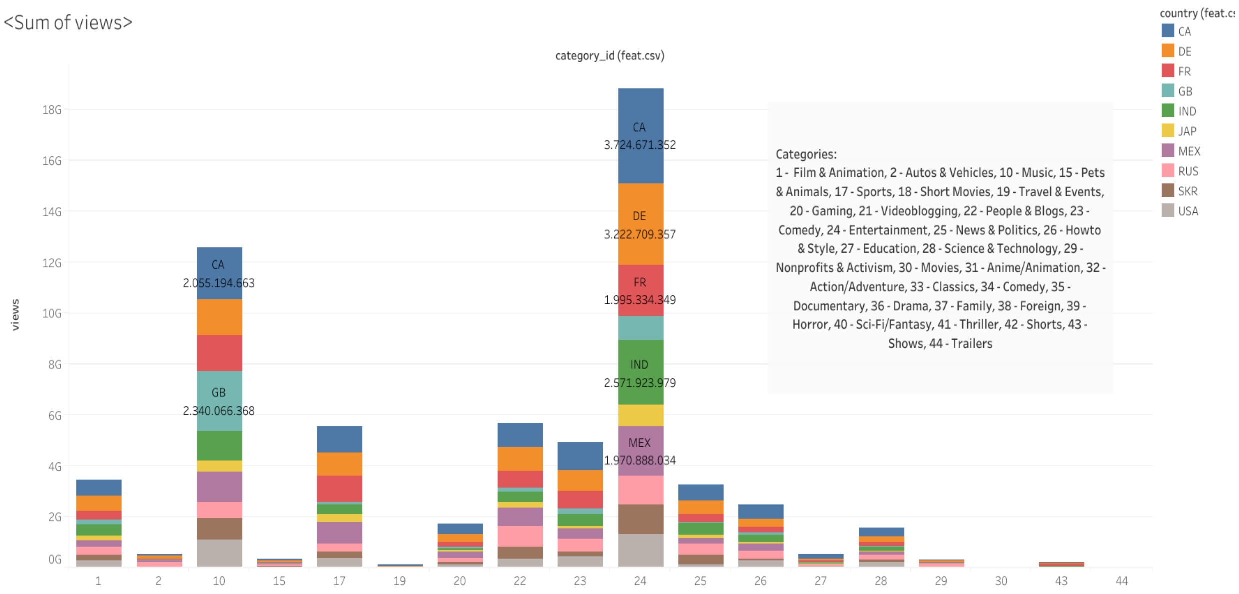


Fig. 3: Sum of views per category.

A similar chart for likes (Fig. 4) shows, on the other hand, that the music category gets the biggest total amount of them, being Canada again, the leader and Great Britain the second. The second category is now 24. Entertainment, with a total of 600 M likes, of which, most of them come from Canada firstly, and Germany secondly.

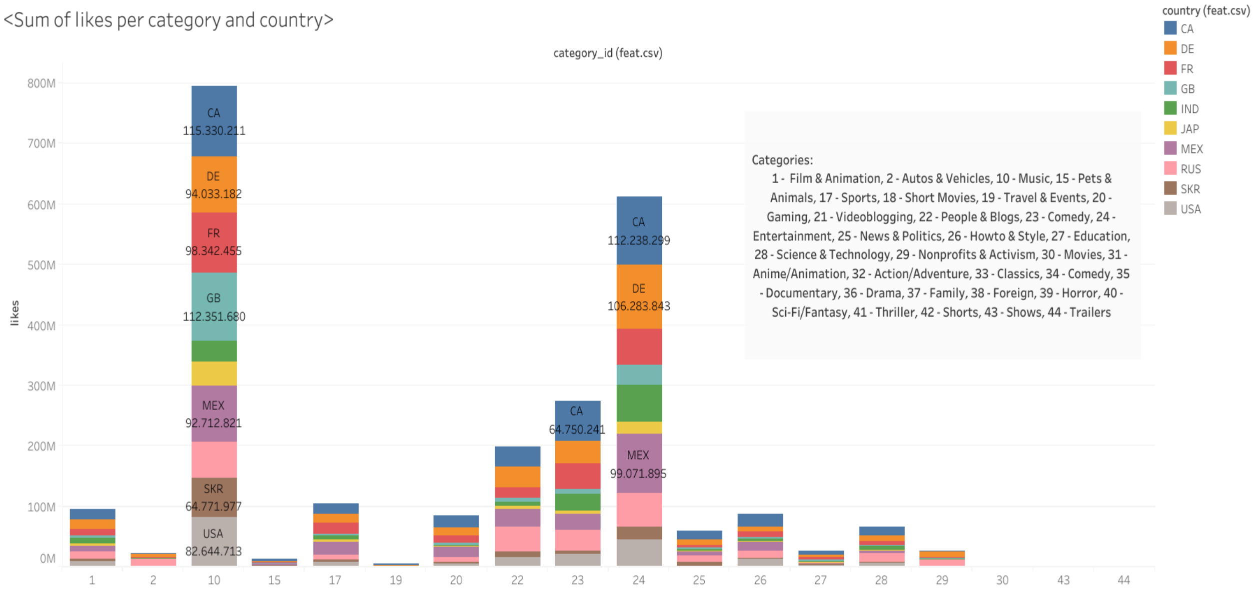


Fig. 4: Likes per category.

The chart of dislikes (Fig. 5) presents some change. Entertainment is again the leader, where germans seem to be the most dislikers of all. In the music category Canada is the leader in dislikes. The rest of the categories still remain less important in comparison.

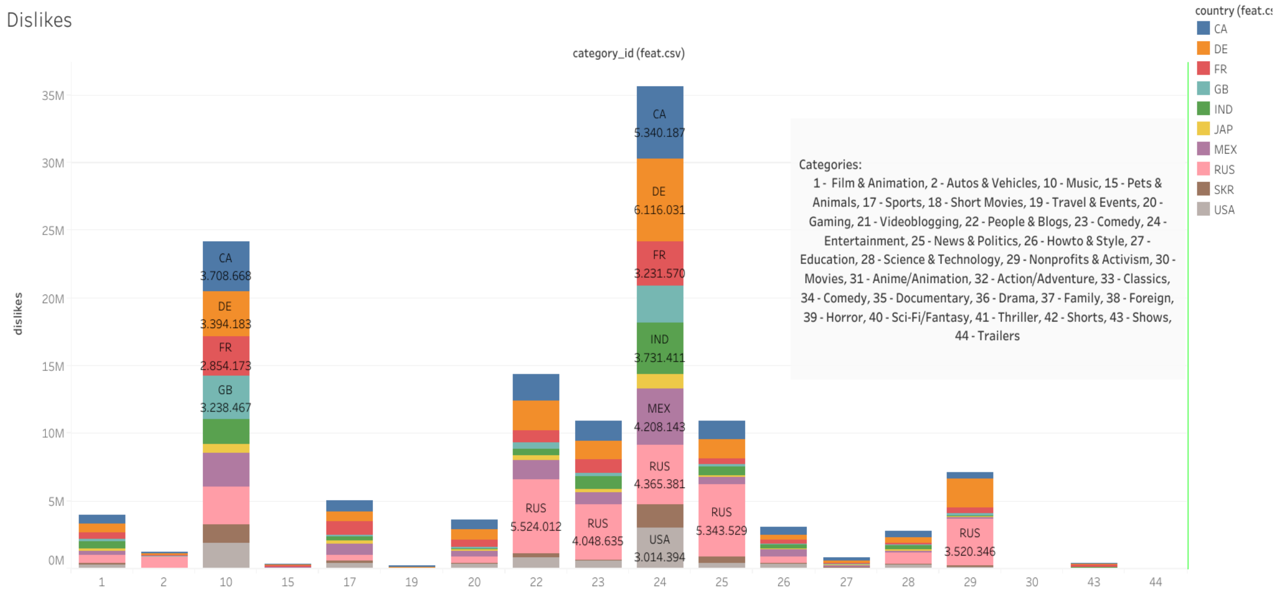


Fig. 5: Dislikes.

In the comment count bar chart (Fig. 6), we get bigger numbers than in the ones before, but we get the same leaders for categories: Entertainment and Music. For the first one, the biggest number of comments come from Canada, surprisingly followed by Mexico. In the music category, the first country is Canada again, and the second is Great Britain.

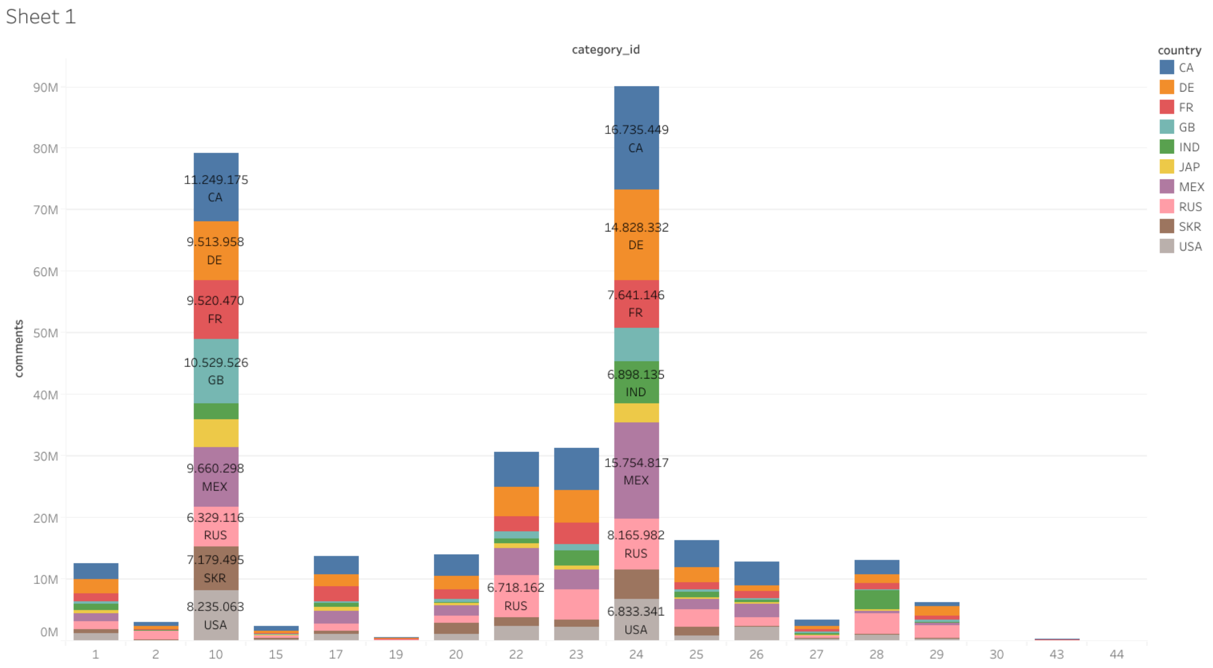


Fig. 6: Comments

The chart for median of days from video publication until it trends (Fig. 6), show that in most cases, videos take only one day to be selected. This is more obvious in the linear distribution (Fig. 7).

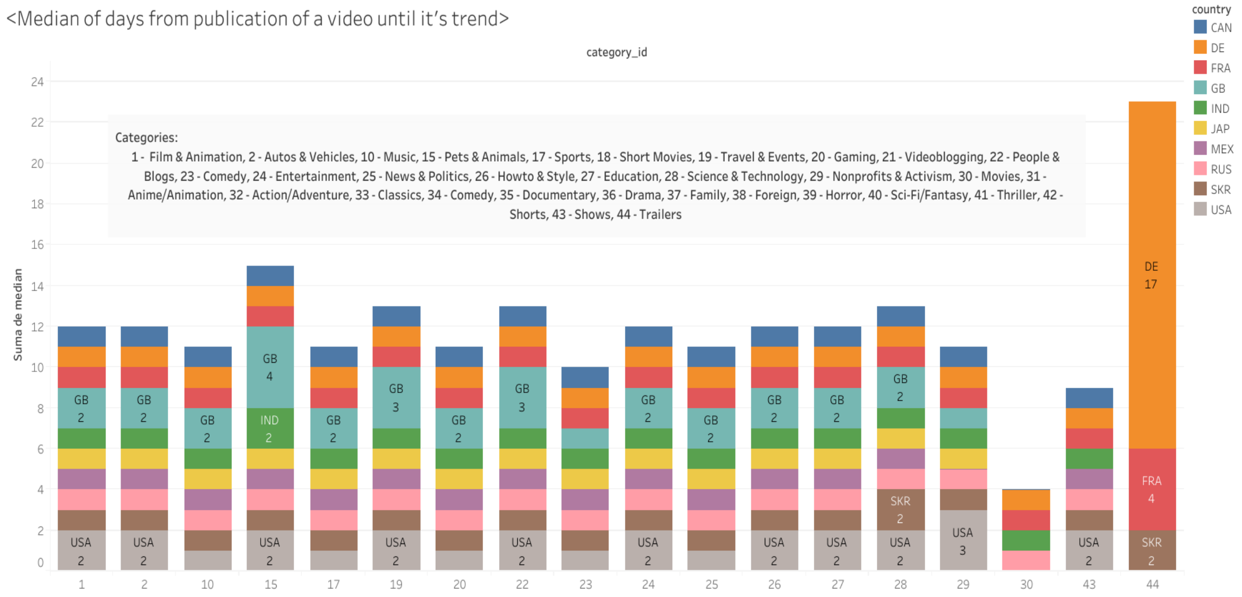


Fig. 6: Median of days from publication until trend, bar chart.

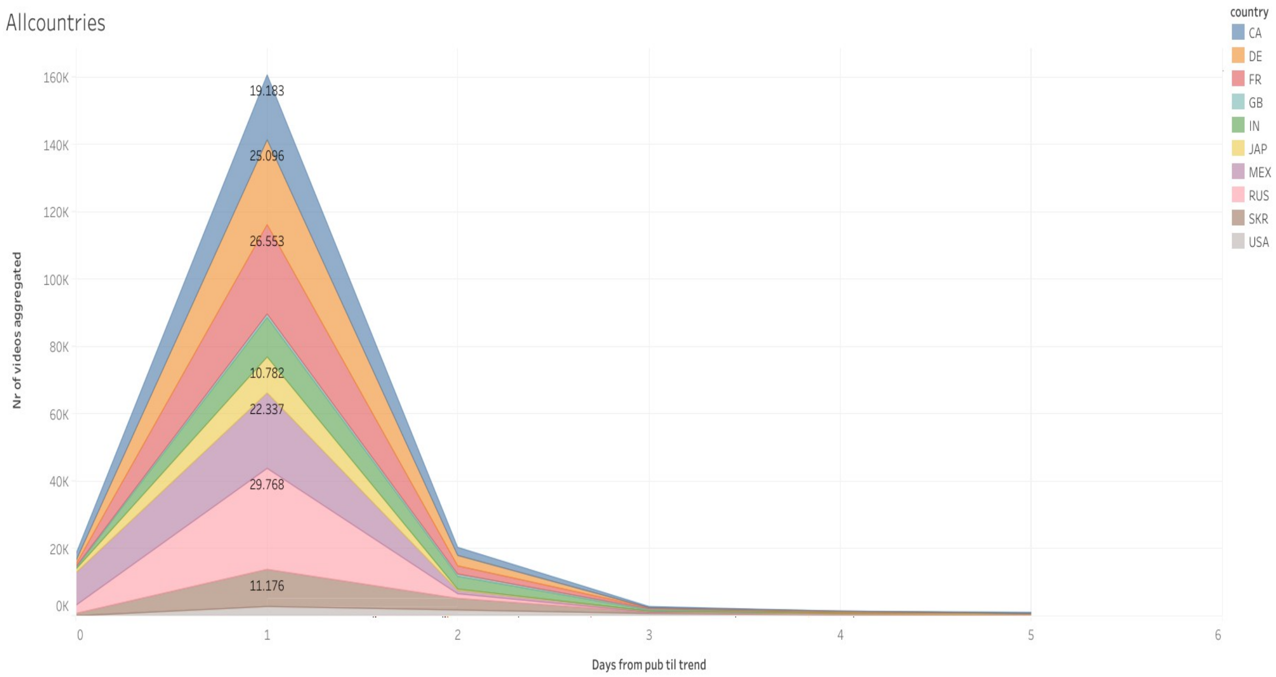


Fig. 7: Median of days, linear plot.

5.- Conclusions

The data indicate that Canada has a very important activity in the platform, as well as Great Britain, and USA is not as important as could be logically expected, having in mind that it's the second country for number of users according to external sources.

Most of the videos with big amounts of likes have a directly proportional number of comments, as there is big correlation. The same happens with views and likes. The dislikes count does not seem to have a correlation with other variables, which means that well viewed videos could have either lots of dislikes, or very few.

The age of a video in the moment of first trend, is also mostly one day, which confirms that YouTube selects mainly videos that are relatively new, and therefore more representative of the tendency and fashion of that moment.