

Youtube Trending Videos Analysis Report

Yifeng Luo

Dec 17, 2018

Introduction

Most of us have been ever watched videos in YouTube (the world-famous video sharing website), which maintains a list of the top trending videos on the platform. When people do not know what video they want to watch, they could look through the trending tab to watch the current hot videos and know what is happening in the rest of the world. Trending aims to surface videos that a wide range of viewers will appreciate, so YouTube users always can find the videos they interested in from the trending list. Some trends are predictable, like a new song from a current popular artist or a new movie trailer. Others are surprising, like a viral video. The list of trending videos is updated roughly every 15 minutes. According to Variety magazine, "To determine the trending videos, YouTube considers a combination of factors including videos category, increment of views, tags and description,etc. YouTube trending system selects videos from massive videos based on a mature algorithm and specific criteria to predict a video will popular or not in the following days, then recommend them with users in trending tab. Therefore, this report will analyse and compare the features of trending video from four countries-United States, United Kingdom, Canada, India to see whether exist selection preference and what kind of video is easier to be popular among countries.

Data Resource

This dataset this research used is a daily record of the top trending YouTube videos from 11/14/2017 to 06/14/2018 in US, UK, Canada and India. It was downloaded from Kaggle. Some people scraped the data by YouTube's API and shared them in Kaggle. The dataset records the number of views, tags and description of trending videos in YouTube. Meanwhile, it includes other video information as well, like its title, category and trending date and publish date. There are many videos in trending list more than 1 day, but the data were collected daily, so it was multiple recorded. This research only keep the first day record, because the other video related information are same except the number of view, like and dislike change by time. Meanwhile, extracting some useful variables from the original dataset, such as the number of time gap between upload date and trending date and sentiment score of video description is an important step.

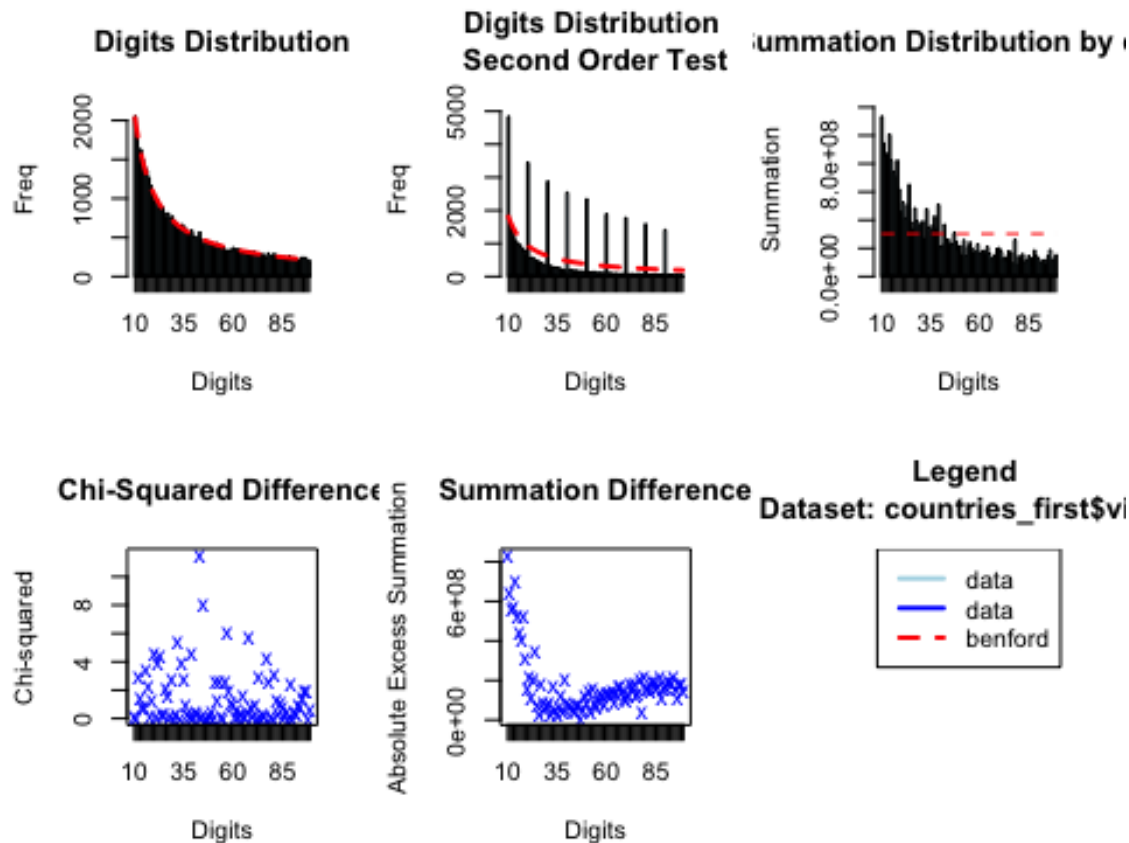
Benford's Law Analysis for Views

First-digit's law:

$$Prob(D_1 = d) = \log_{10}\left(1 + \frac{1}{d}\right)$$

Generalize,

$$Prob(D_1 = d_1, D_2 = d_2, \dots, D_m = d_m) = \log_{10}\left(1 + \left(\sum_{j=1}^m 10^{m-j} d_j\right)^{-1}\right)$$



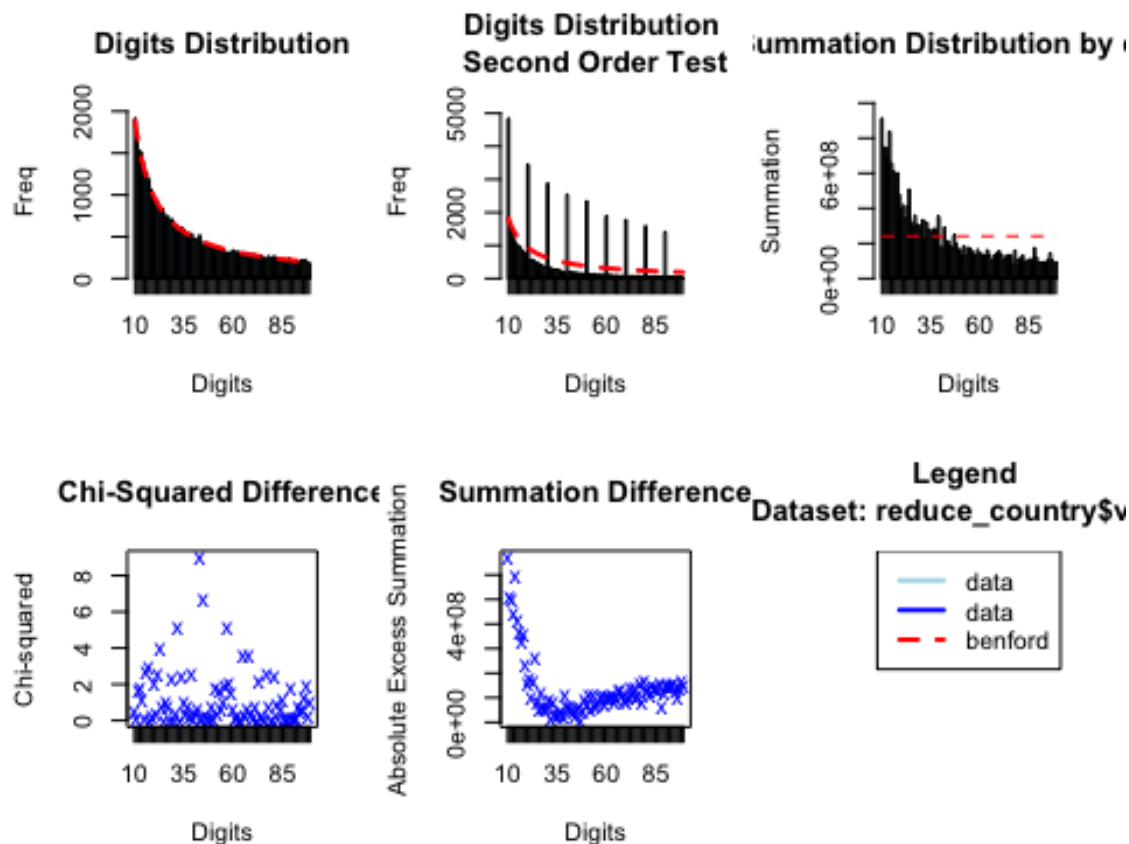
```
##
## Pearson's Chi-squared test
##
## data: countries_first$views
## X-squared = 132.09, df = 89, p-value = 0.00207
```

The original video views are in black and the expected frequency according to Benford's law is in red in the first plot. This plot shows the difference between original data and expected data. Several first two digits occurred more frequently than expected under the Benford distribution (43, 12, 20, 16, 23) as shown in the spikes. Of these, 43 is the most anomalous occurrence.

Meanwhile, this result can be verified by Chi-squared difference test. The calculated Chi-squared statistic here is 132.09 and the p-value of the test is 0.00207, which indicates that there is sufficient evidence to reject the null hypothesis of conformity to Benford's law.

However, this result probably is caused by joining 4 countries' data as one big dataset. Some videos were shared and popular in several countries, so their views are total views not for an individual country. After checking the unique video ID, there are more than 3000 rows are repeated in the dataset except the country is different.

```
## [1] 45571      2
## [1] 49070     14
```



```
##
## Pearson's Chi-squared test
##
## data:  reduce_country$views
## X-squared = 101.42, df = 89, p-value = 0.1736
```











After removing the duplicate rows, the research did the Benford test for views again to check the data accuracy. From the first plot, the red line fit better with original data. And the chi-square statistic is 0.1736, which indicates that there is not sufficient evidence to reject the

null hypothesis of conformity to Benford's law and verify the data could be real and not be manipulated.

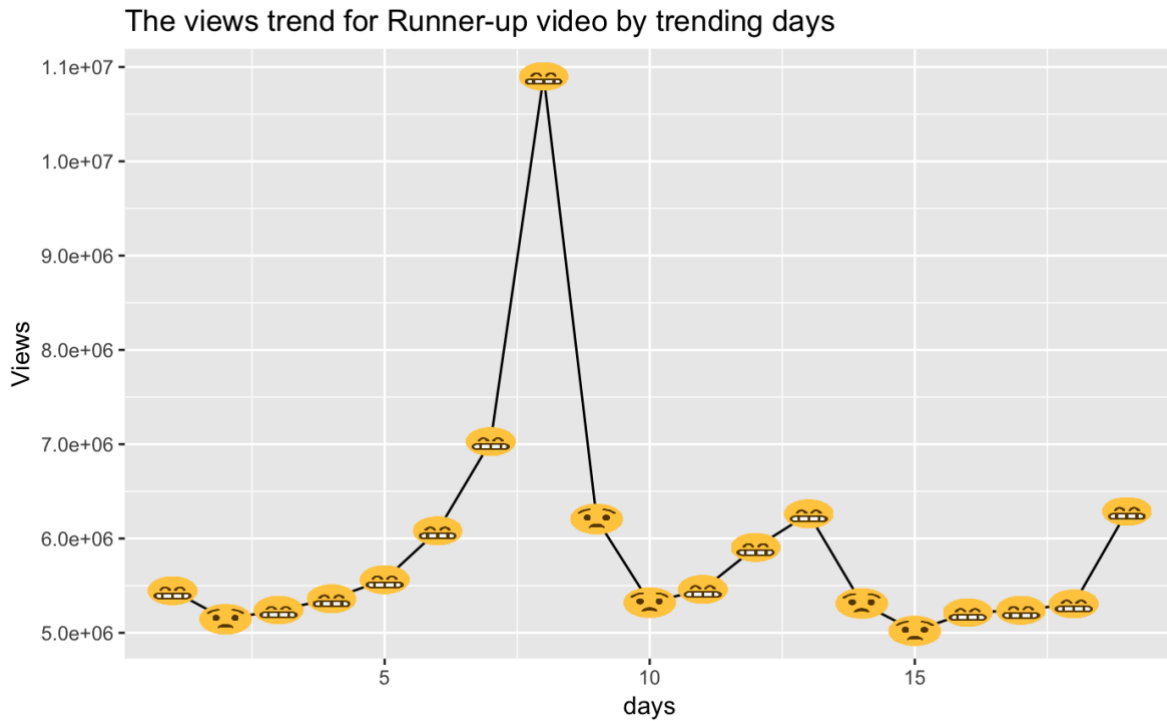
Visualization – Exploratory Data Analysis

This part aims to find the features of trending video and the differences among countries by varieties of plots.

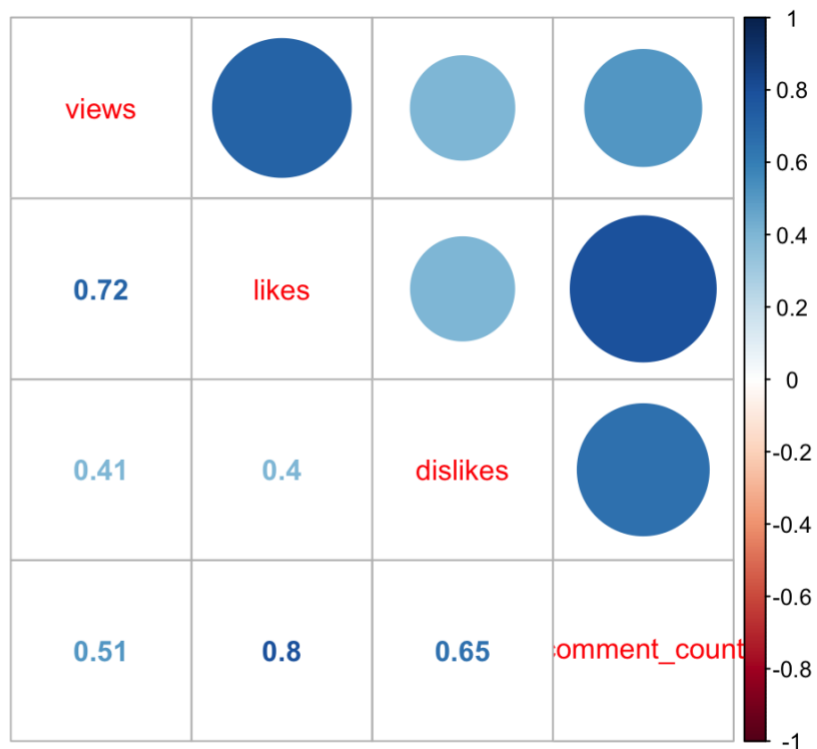
1) Views

	image	country	category	title	views
1		GB	Music	Luis Fonsi, Demi Lovato - Échame La Culpa	143408235
2		GB	Music	Becky G, Natti Natasha - Sin Pijama (Official Video)	88568646
3		US	Music	Maluma - El Préstamo (Official Video)	48431654
4		GB	Music	Sebastián Yatra - Por Perro ft. Luis Figueroa, Lary Over	47669287
5		GB	Music	Ozuna - Única (Video Oficial) A U R A	42923278
6		US	Music	BTS (방탄소년단) 'FAKE LOVE' Official MV	39349927
7		GB	Music	BTS (방탄소년단) 'FAKE LOVE' Official MV	39349927
8		CA	Music	BTS (방탄소년단) 'FAKE LOVE' Official MV	39349927
9		GB	Music	Rkm & Ken-Y X Natti Natasha - Tonta [Official Video]	39118664
10		US	Music	TWICE What is Love? M/V	38873543

This table shows the top 10 views of trending video from Dec 2017 to Jun 2018. It is clear to see all of these videos are music video and most of them come from UK and US, which means music genre are easier to attract viewers' attention, so their views are higher than other trending videos'. Meanwhile, viewers in US and UK are more likely to watch MV in YouTube.

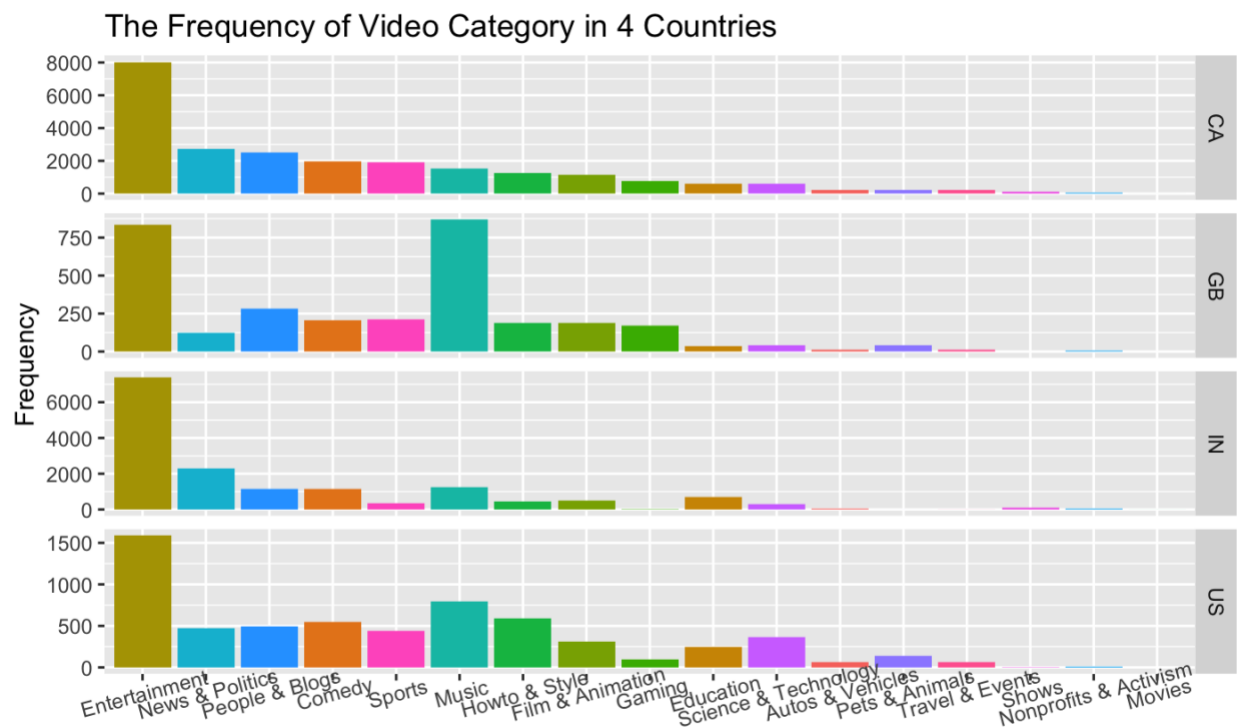


The line graph shows the changes of views increment of the runner-up video in the above table during the trending period. The x-axis is the trending days and y-axis is the incremental views. If the video increment is larger than last day, the node shows a smile face and vice versa. Overall, the views increased rapidly from day 3 to day 9 and reached a pick in day 8.



From the correlation plot, it is clear to notice that the amount of view, like, dislike and comment highly correlate.

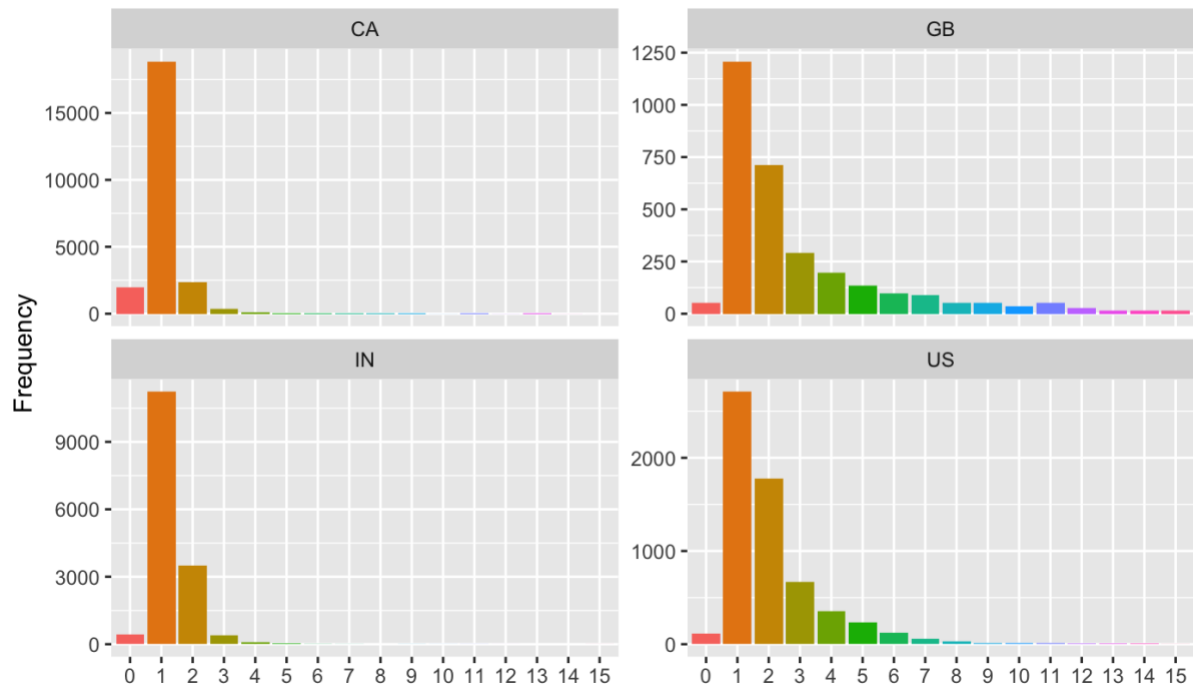
2) Video Category



This plot shows the difference of video category of trending videos by each country. Overall, the entertainment video has the highest frequency. The music videos are more popular in US and UK, especially in UK. On the contrary, the frequency of news and politics video in UK is lower than other three countries. A thing should be noticed is the science and technology and endcation videos attract more attention than other three countries. On the other hand, the sport videos are not popular in India

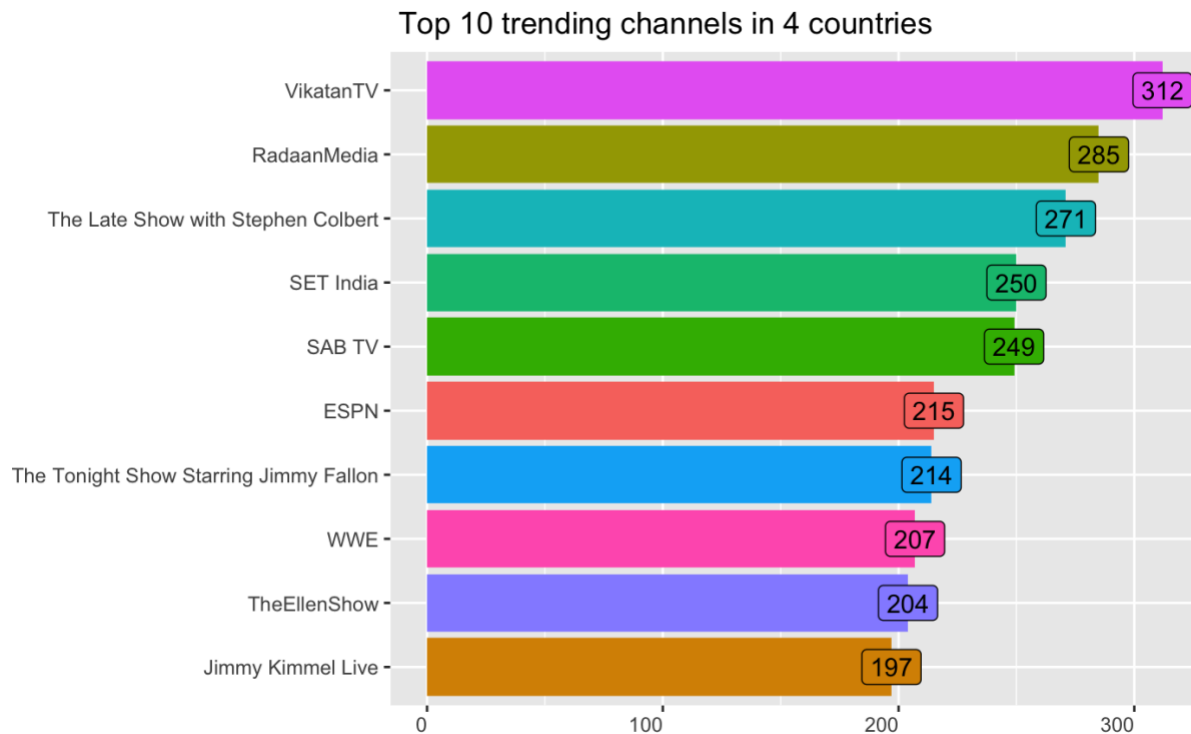
3) Time Gap

The Frequency of Time Gap between Video Upload and Trending



The facet plot shows the distribution of time gap between video upload date and trending start date. Overall, it seems that the videos never trend in the same day it is published and most of video trended between 1 to 3 days after uploading.

4) Channels



The bar chart shows the top 10 trending channels among US, UK, CA and IN. Most of them are TV channel and talk show (EllenShow, Fallon show and Kimmel show).

5) Tags

This part will use wordcloud to analyse the tags attached in trending videos to find features and differences between countries.



Here is the wordcloud about the top 100 frequent tags in US trending videos. The highest frequency words are "funny" and "comedy". Meanwhile, it shows some interesting tags, like "nba", "basketball", "food", "pop" and "science". These tags are highly relevant to American life.

video. Meanwhile, there are some highly relevant tags with India, like “bollywood”, “hindi” and “zee5”.

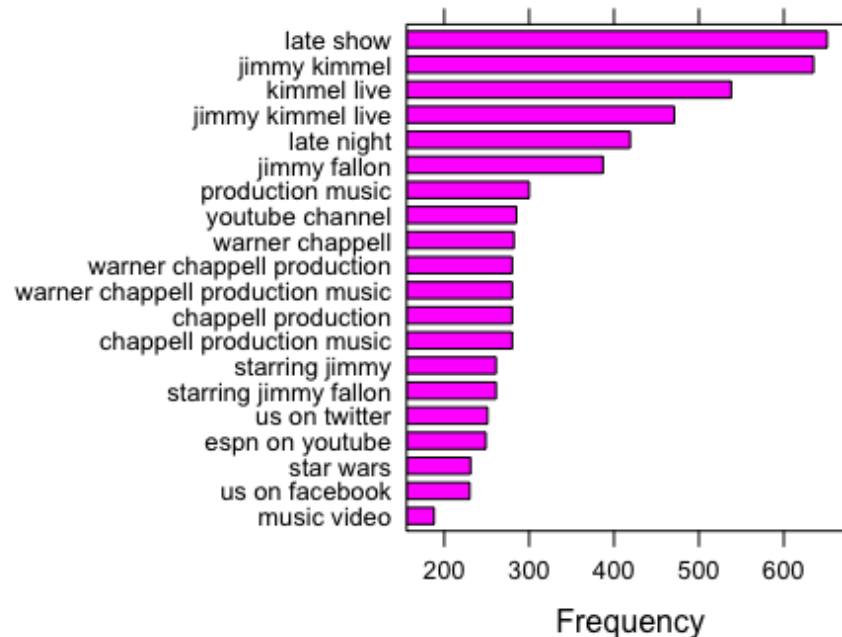


Here is the wordcloud about the top 60 frequent tags in UK trending videos. Except “music” and “funny”, “music” is a highly frequency tag, which correspond with above analysis result, British really like to watch music video in YouTube. Moreover, there are many music-related tags, like “rap”, “pop”, “trailer” and “hip pop”.

6) Description

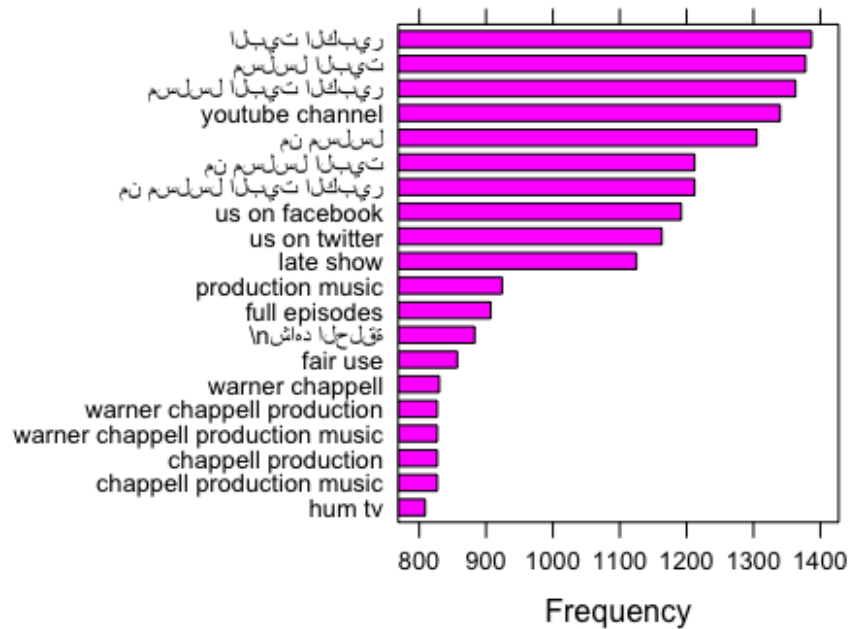
In this part, this report will explain the text mining and sentiment analysis of video description. In the text mining part, this research applied RAKE algorithm to extract noun phrases from the description. RAKE short for Rapid Automatic Keyword Extraction algorithm, is a domain independent keyword extraction algorithm which tries to determine key phrases in a body of text by analyzing the frequency of word appearance and its co-occurrence with other words in the text. In the sentiment analysis part, this research applied “bing” lexicon to get the frequency of positive and negative words and used “sentimentr” package to calculate the sentiment score for each video description.

Keywords - simple noun phrases



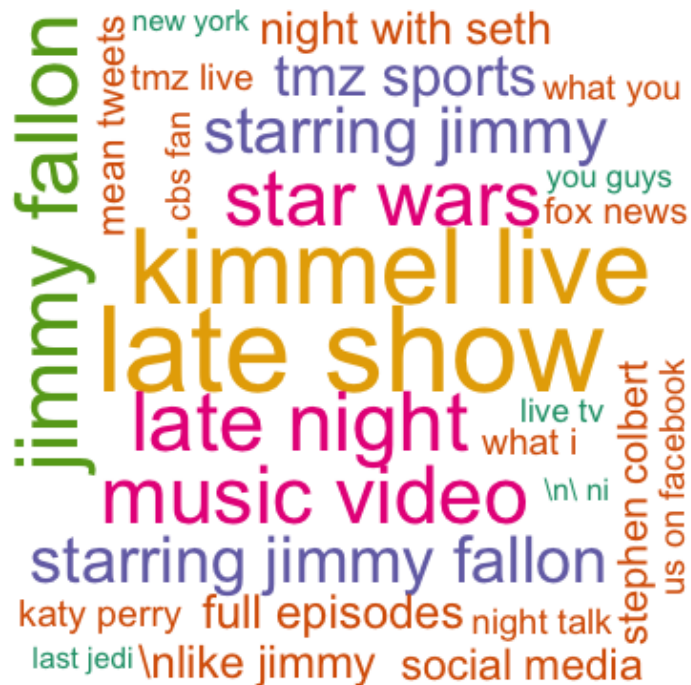
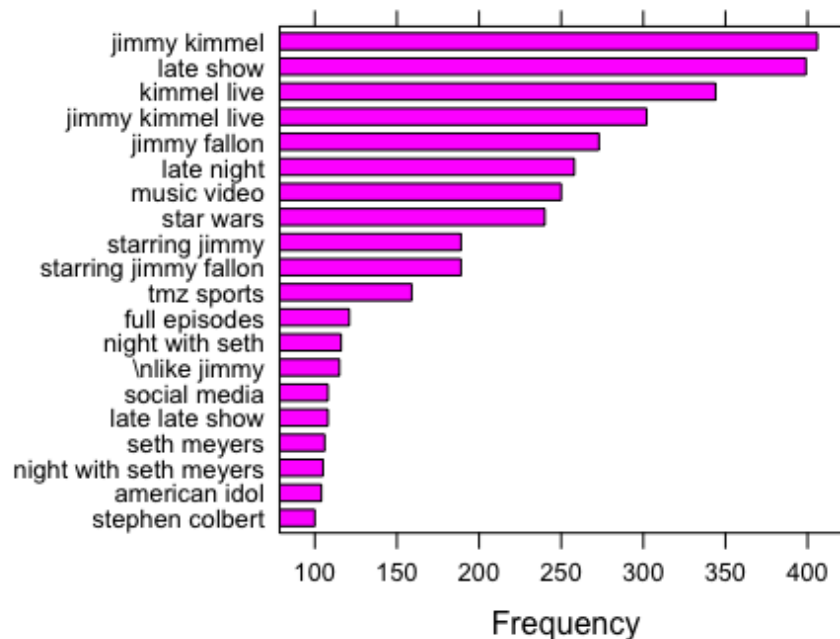
From the wordcloud, it is clear to see Jimmy Kimmel and Jimmy Fallon are really popular in US. Most of trending video mentioned “jimmy follon” and “jimmy kimmel” in their description. Meanwhile, “warner chappell”, “production music” and “music video”, these knid of music related phrases usually were mention in description as well. (Note: Warnner Chappell is a music production company.)

Keywords - simple noun phrases



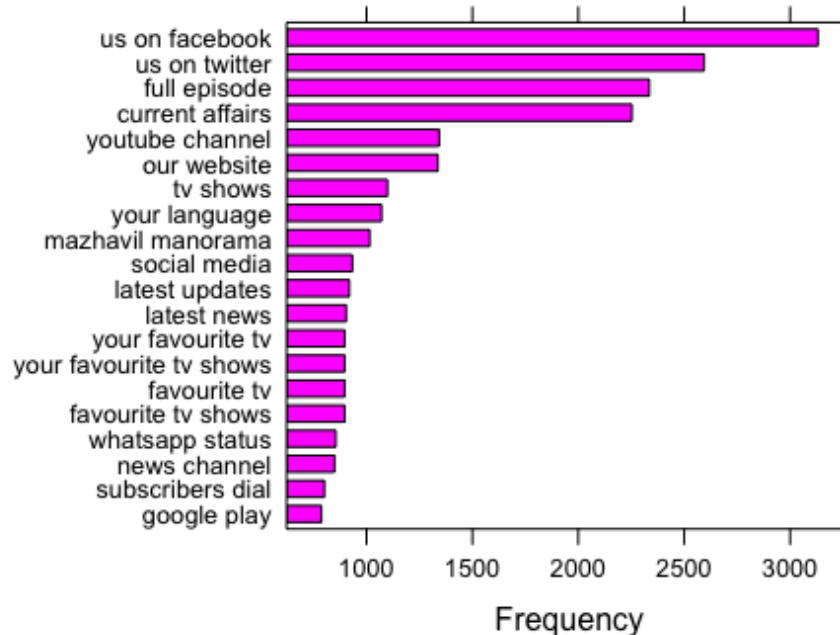
The wordcloud shows the highly frequency noun phrases in description of Canadian trending video. There are some languages from different countries in the wordcloud which means many inter-culture videos were trend in CA. And the “warner chappell” were always mentioned as well. On the other side, video creators always mentioned other social medias in the video description to attract followers, like “tweets” and “facebook”.

Keywords - simple noun phrases

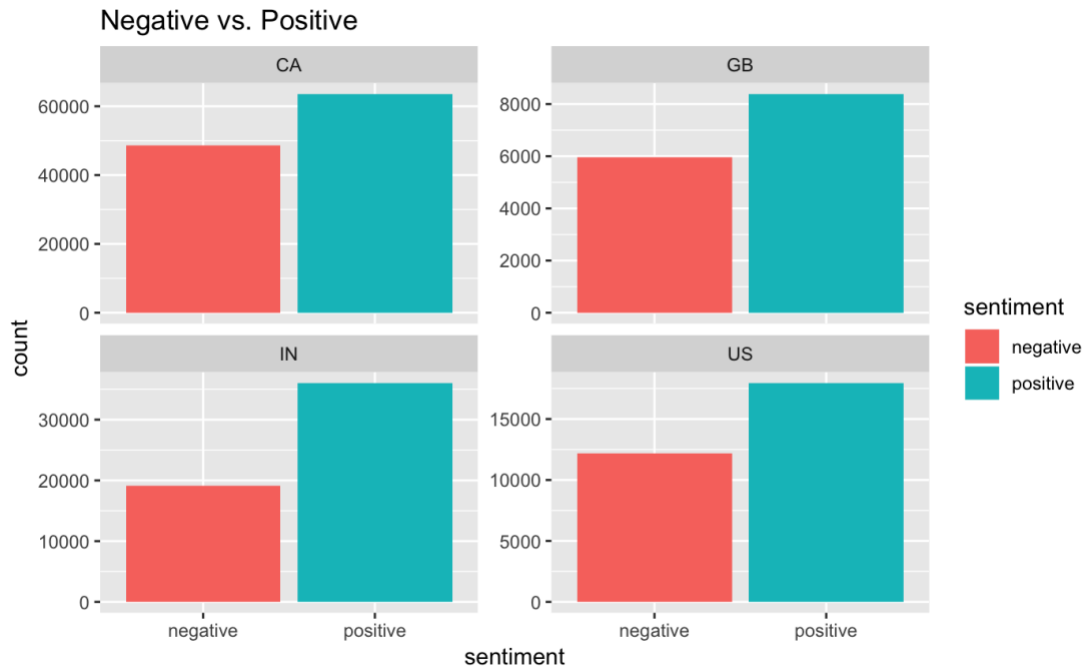


Again, “jimmy kimmel” is the most frequent phrase in description of UK trending video. Meanwhile, some interesting phrases were mentioned, such as, “star wars”, “tmz sparts”, “American idol” and “katy perry”.

Keywords - simple noun phrases



Indian trending video description always mentioned “current affairs” and “full episode”. Therefore, current affairs and drama are easier to attract people’s attention in India to some extent.



This bar chart shows the frequency of negative and positive word in video description among 4 countries. It is clear to see the positive word has a higher frequency than negative word. A good video should share a positive thought and attitude.



The distribution of sentiment score of description in these 4 countries indicates the most of sentiment score are higher than 0, which means the description is positive in overall.

Acknowledge

The data could not have been created without the hard work of the person who grasped the data from YouTube. They actually did a lot of work of collecting all the necessary metrics of the video records. And thanks to the everyone who shared their great ideas and EDA process in Kaggle, which inspires this study to dig in deeper.