YouTube Trending Video Visualization

Chen Liang-Yu
Hong Kong University of Science and Technology
Student ID: 20387832
lchenbm@connect.ust.hk

Chen Yutong
Hong Kong University of Science and Technology
Student ID: 20412261
ychendm@connect.ust.hk

Huang Xuhua
Hong Kong University of Science and Technology
Student ID: 20329347
xhuangat@connect.ust.hk

Zhang Yi
Hong Kong University of Science and Technology
Student ID: 20413734
yzhangfg@connect.ust.hk

ABSTRACT

This visualization is for YouTube trending videos. Our main goals are to analyze the difference among countries' preference on video categories, summarize the trend of video content as well as figure out the category popularity pattern.

Index Terms: YouTube Trending Video Visualization, Interactive Map, Word Could, Stacked Graph, Chord Diagram, Parallel diagram, D3.js, React, Python, Lodash

1 INTRODUCTION

YouTube is the world's largest video search engine and sharing platform with over 1.5 billion users around the world. Trending videos are automatically filtered regularly based on the popularity of videos in each region. It aims to surface videos that a wide range of viewers would find interesting. Visualization of the trending video data enable us to have a deeper insight into the pattern of popular categories and video content. In this project, we visualized YouTube trending video data from Kaggle. This is a multivariable dataset from November 2017 to June 2018 and can be classified the features into four categories, which are numeric data, temporal data, geographic data and text data. It provides the daily data of trending video's number of likes/ dislikes/ reviews/ comments, text of title and description, region, etc. We mainly focus on 3 dimensions to analyze the YouTube trending dataset, including country analysis, video title analysis and video category analysis as well. In the following sessions, visualization details and findings of the above 3 dimensions will be shown, respectively.

2 INTERACTIVE WORLD MAP ANALYSIS

In this section, we will introduce how we conduct the interactive world map analysis as well as some findings. The goal for this task is to let user visualize the difference of month, category, and numeric feature such as number of likes and dislikes between countries.

2.1 World Map Visualization

We first preprocess the data with python to aggregate the target features and store it in JSON. We then make use of React to create filter options and utilize props to pass the user's selection to the world map component rendered by D3.JS. The color of the map is in linear scale and single hue, if the country is not in the dataset, we will set their value to zero and the color would be gray. We also add animation to smooth the transition when switching to different contents. Figure 1. displays the overall view.

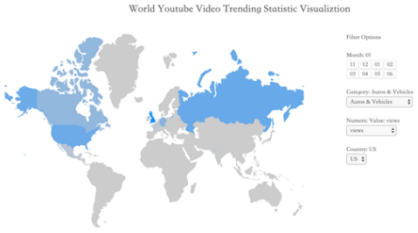

Figure 1. Overall View of the World Map Visualization

The left part of Figure 1 is the world map component, and the right part is the filter options. We design the whole webpage with CSS to make it more user friendly. Moreover, we design a tooltip so that user can figure out the exact numeric values when they hover on the country.

Besides the existing numeric values, we also calculate several rates that users could visualize, and they are displayed in different colors.

$$favor\ rate = \frac{\#\ of\ likes}{\#\ of\ views}$$

$$toxic\ rate = \frac{\#\ of\ dislikes}{\#\ of\ views}$$

$$comment\ rate = \frac{\#\ of\ comments}{\#\ of\ views}$$

Figure 2. shows the example views of favor rate, toxic rate and comment rate accordingly.

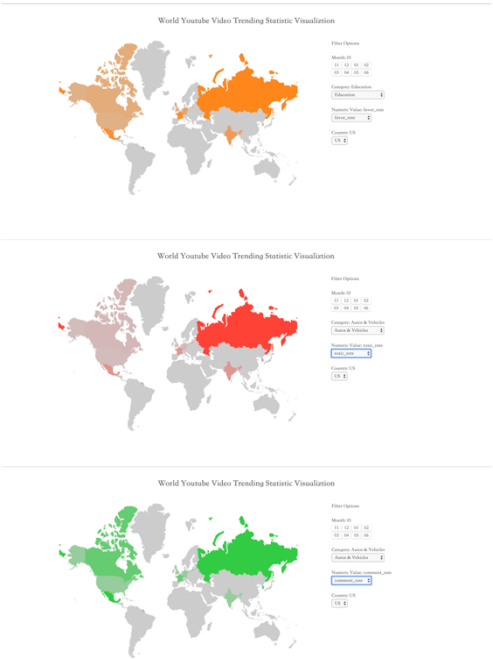

Figure 2. Favor Rate, Toxic Rate and Comment Rate

Since this is an interactive visualization, users could create numerous of combinations and obtain insights accordingly. Here we just list some interesting findings.

1. During Winter and Summer, Gaming is more popular in the U.S. compare to other countries.
2. Music are very popular in England, U.S. and Canada comparing to other countries like Russia.

3. Russia is more interested in Autos & Vehicles, and it is the most toxic country to many categories for most of the time.

3 VIDEO TITLE ANALYSIS

In this section, we use word cloud of video title and staked graph to study users' content preference in a more specific level than category. Due to work efficiency and language limitation, we filter out the non-English speaking countries and non-English videos, which means our dataset is trimmed down to English videos in US and Canada. We've found that there're more non-English trending videos in Canada. It is probably because there are more non-English speakers, like Chinese immigrants, in Canada.

3.1 Word Cloud

We first analysis the content trend using word cloud of video title. We extract words in title expect some less meaningful words, like is, are, of, an, etc. Next, we assign the likes/dislikes number of the title to every word in that title. Then we add up all the likes/dislikes number of each word in one month. We generate two graphs for data of every month in each country. Red graph represents the content that users like most and blue graph represents the content that users dislike most. To improve the intuition of our visualization, we mask the word cloud with map shape of each country. Our encoding system is based on the word size and the color saturation. Words with more likes/dislikes will have larger size and higher color saturation.

Figure 3 shows the word cloud for trending video title in January 2018 and February 2018 in both US and Canada. Here are our key findings. First, the most popular videos are usually music videos. As you can see, the large words, like taylor, Sheeran, are music related. Secondly, the high frequency words usually come from one video, you can search the big words in one graph in YouTube and easily find the corresponding videos. For example, searching "game, future, taylor" will lead you to the music video named "Taylor Swift - End Game ft. Ed Sheeran, Future" published in Jan 18, 2018. Also, trending videos in Canada and US have high similarity. Maybe this is because the two countries have similar culture and geographic location.

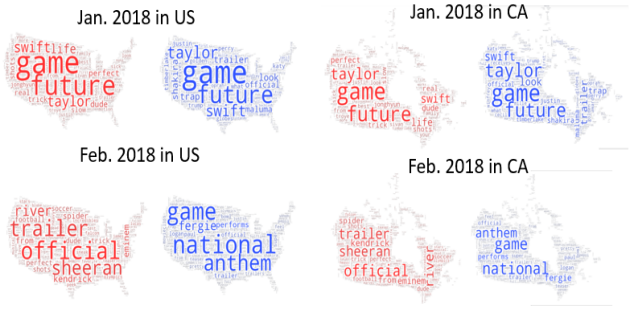

Figure 3: The word cloud of video title

3.2 Staked Graph

As a complementation of word cloud which is lack of details, our next function is to use staked graph to show the daily likes and dislikes of words in every word cloud. Like color encoding in word cloud, red represents likes and blue represents dislikes. In addition to tackle the eye beat memory issue, we can show four graphs at one time.

In figure 4, we show the example of graphs of four words in January 2018 in US. The date and indistinguishable graphs of "taylor" and "game" indicate that a large portion of likes and dislikes of these two words come from one video mentioned above. The same is true for "perfect" and "life". They come from "Real Life Trick Shots 2 | Dude Perfect" published in Jan 22, 2018. Therefore, we are further confirmed that the most popular words are usually the consequence of one popular specific video rather than a general topic. Moreover, popular videos tend to have a short lifespan of around 10 days.

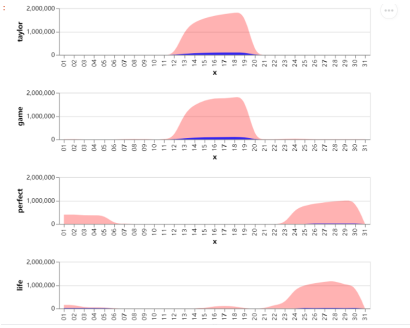

Figure 4: The stack graph of four words in Jan 2018 in US

3.3 Interactive interface

Figure 5 shows our interactive interface of two functions mentioned above. For word cloud, users can input the specific country and month. For stacked function, users can input the specific words of which they want to investigate the details.

Figure 5: interactive interface for word cloud and stacked graph

4 VIDEO CATEGORY ANALYSIS

4.1 Chord Diagram

In this section, we aim to analyze the popularity of video categories and get deeper insight into the relationship among popularity and video categories.

We decided to use Chord Diagram, implemented by D3.js, to visualize the popularity of different video categories. We use "number of views" provided by the Video Trending Dataset as the metric of "popularity", because a video is more popular when more people choose to watch it. Though the range of "number of views" is a discrete number spreading out from 0 to 356,464, we divide the range into 4 integer intervals with equal length and classify them as *Very Popular*, *Popular*, *Normal* and *Not Popular* respectively. These 4 popularity levels are located on the left side of our Chord Diagram. On the right side of our diagram, we only choose the 6 most common video categories to make our visualization tidier. In term of the encoding schema, we use different colors to represent different video categories or popularity levels. The width of curves / ribbons represents the number of videos belonging to a specific video categories and fitting into a specific popularity level.

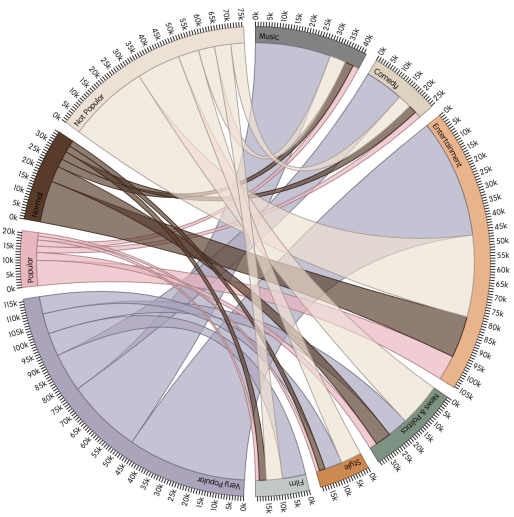

Figure 6: chord diagram for popularity analysis

To make our visualization more interactive and easier to highlight what we want to express, we utilized two techniques. They are *fading* and *tooltip*. Because we are concerned that the crossing ribbons in Chord Diagram will make the audience dazzled and unable to focus on the most important information, we use the *fading* technique to let all the chords fade except the one on which there is a mouse. As demonstrated in Figure 7, this will help emphasize one single two-way relationship which is selected by the users. The second technique we use is *tooltip*, the detailed number encoded in each ribbon will only be presented when users move their mouse on top of a ribbon. More details, such as the name or country statistics of these videos, can also be added in this tooltip.

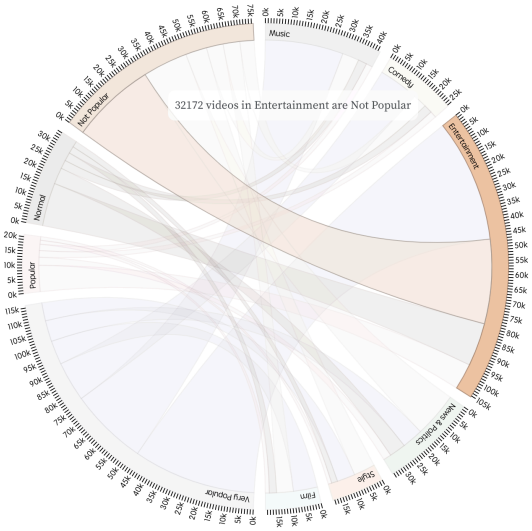

Figure 7: visualization techniques demonstration (fading and tooltip)

This visualization follows two visualization rules learnt from this course, which are *Eyes beat memory* and *Details on demand*. For *Eyes beat memory*, we do not use any transition to require users to memorize the detailed information in order to do comparison, because it is easy to do comparisons between different video categories by simply compare their flow volumes presented with different colors. For *Details on demand*, we use tooltip to hide some detailed information encoded in a specific relationship. When users demand more information, they can choose to move their mouse on the ribbon of their interest to retrieve more details.

From this visualization, we find that "Entertainment" contains the largest group of videos, but it takes the biggest part in both "Very popular" and "Not popular", which implies that lots of its videos are low-quality because they are not popular. On the other hand, more than 75% videos of Music are "Very popular", which indicates that Music related videos in YouTube tend to have highest popularity, compared with other video categories.

4.2 Parallel coordinates

In this part, we aim to visualize the inter-relationships among different video categories and their popularity, as well as the user's preference. We use parallel coordinates, implemented by D3.js, to encode three types of data: video category, popularity and user's preference. More specifically, we use "number of views" as a metric for popularity. Meanwhile, we use "number of likes and dislikes" as a metric for user's preference. For example, a video with more likes has higher user preference.

Firstly, we preprocess the dataset. We delete the videos with no record of category, and then normalize "number of views", "number of likes", "number of dislikes" and "number of comments" to the range of 0 to 1.

The four axes in this visualization, from left to right are number of views, likes, dislikes and comments. We use different colors to represent different video categories. Each line represents a trending video in YouTube. For example, a pink line in this visualization represents a video in the category of music. Figure 8 visualizes the YouTube top trending videos in all categories.

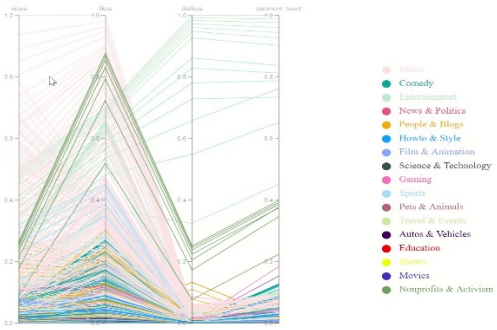

Figure 8: YouTube trending videos in all categories

With a large dataset, the problem of visual cluttering prevents effective revealing of useful patterns for YouTube trending videos. To solve this problem, we add hover effects to highlight a specific category or axis. When users move the cursor to a specific category, first all categories will turn gray, then the specific category will be highlighted. Figure 9 displays the hover effects of this visualization.

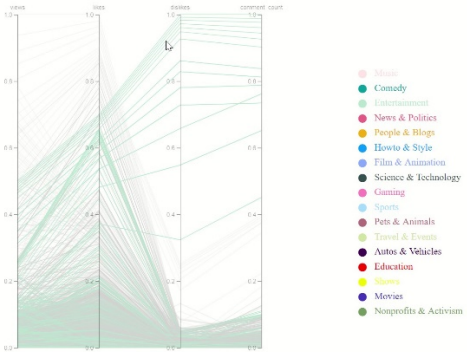

Figure 9: Hover effects of this visualization

This visualization has many advantages. Firstly, this visualization integrates high-dimensional data into one graph. Secondly, it follows the visualization rules *eyes beat memory* and *Zoom in and filter*. We have an overview of YouTube trending videos in all categories. Users can easily make comparisons between different video categories since categories are encoded with different colors. Users can also figure out the detailed information for a specific category. Thirdly,

this visualization solves the problem of visual cluttering by adding hover effects. It can provide effective revealing of meaningful patterns in this dataset.

There are many interesting patterns from the parallel coordinates. For example, music and entertainment are the two most popular categories. But it is obvious that videos in the category of music have higher user preference than entertainment. Although videos in the category of "Nonprofits & Activism" are not popular, they have high user preference. Besides, in US, users prefer videos in the category of Comedy than Gaming.

5 Conclusion

In summary, we demonstrated our knowledge learnt from the course and used D3.js, React, Python and Lodash to achieve visualization technology like interactive map, word cloud, stacked graph, chord diagram, and parallel graph to visualize YouTube trending video data from the perspectives of countries, titles and categories. We found some interesting patterns in our dataset and below briefly summarize what we have found in the three dimensions.

Country Analysis: Users could visualize the difference between countries according to their options and acquire its own findings. For example, Music and Entertainment are both very popular in the U.S. and Canada compare to other countries.

Video Title Analysis: The popularity of words are usually the consequence of one popular specific video rather than a general topic. US and Canada always share the similar trending videos.

Video Category Analysis: Though the Entertainment-type videos dominate YouTube videos in term of number, they contain lots of low-quality videos. On the contrary, because Music-type videos tend to have highest popularity, the overall quality of Music-type is the highest. Videos in the category of music have higher user preference than entertainment.

Our code could be found in this public repository:
https://github.com/fredchen000/COMP4462_2020Final_Project

Due to data privacy issue, we didn't include the dataset in above repository, please download the dataset from Kaggle:
https://www.kaggle.com/datasnaek/youtube-new