Trending Analysis on YouTube Videos

Jiaona Ma

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, and in general, 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 of the trending videos will have impact on the design and evaluation of personalization services such as precise advertising.

In this project, I try to figure out what are the general features of trending videos and 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.

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 purpose, I only use data from 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 employee a number of techniques from the Scikit-Learn, NumPy, Pandas, Matplotlib toolkit in Python to analysis the dataset at hand.

There are number of works conduct on “YouTube’s Trending Video Statistics”, and most of the works are applying Exploratory Data Analysis (EDA) approach on US dataset to conclude what kind of attributes does a trending YouTube video have.

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, therefore, I decide to analysis not only USA data, but also Great Britain, Germany, Canada, France, and Mexico. Moreover, I am also planning to apply Natural Language Processing on video titles to figure out what kind of words have higher frequency in the trending videos since an attractive title is the very first thing that will catch viewers eyes.  Finally use the results to tell how the YouTubers are supposed to refine their videos to obtain more subscribers.

As for the brief plan of action, I do the systematic data preprocessing analysis and data visualization first, employee a number of techniques from the Scikit/Learn, NumPy, Pandas, Matplotlib toolkit in Python, since before conclude 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, visualizing the results extracted from the first step. I will also apply the NLP to analysis the video titles in this step. I will try to build predictive models to predict the trending videos by splitting the existing dataset to the train dataset and test dataset. The models I am planning to use are the linear regression, multilinear regression, logistic regression, k-nearest neighbors, and decision trees. After building up the models, I will do the model selection and evaluation by applying Leave One Out Cross Validation, k-Fold Cross-Validation and The Bootstrap. At the end of the project, evaluate the performance of the models, wrap up all the findings, and make suggestions on how to produce videos with higher probability to appears on the trending lists.

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 (I split the available dataset into the train and test sets before).

For the mid-quarter presentation, I am going to show the overall project goal, work have done so far, such as data preprocessing analysis, and what expect to finish during the next half of the quarter.

For the final presentation, I am going to demonstrate the works finished after the mid-quarter checkpoint first, and maybe point out the changes made for the works that finished before the mid-quarter checkpoint, and finally do a conclusion and discussion on the findings from the project.

**References:**

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:

<https://www.kaggle.com/yanpapadakis/trending-youtube-video-metadata-analysis>

<https://www.kaggle.com/quannguyen135/what-is-trending-on-youtube-eda-with-python/notebook>

<https://www.kaggle.com/mlenzovet/newpublish-time-strange-distribution/>