Chujun Ji

MSBA 2019

Professor Hintze

April 17, 2019

**Introduction**

YouTube is currently the biggest video sharing platform and YouTube provide yearly trending video. To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring user interactions (number of views, shares, comments and likes). Note that they’re not the most-viewed videos overall for the calendar year. The dataset I’m using for this project has the information of trending videos in United States including video title, channel title, category, tags, publish time, trending date, views, likes, dislikes and comments.

**Goal**

My first initiative is to predict video views based on this information as insights for YouTube advertisers. However, after attempting and consulting, I changed my goal to find insights from trending video daily view distribution so that YouTube advertiser gets to choose potentiao premium videos to put their advertisements based on the video information.

In order to find insights from similar view distribution video, my pipeline is to cluster the data trend into different categories and then find the mean views, comments, likes and dislikes. After clustering, find common tags and categories inside different clusters.

**Approach**

**A screenshot of a cell phone

Description automatically generated**

This is the visualization of all the video view trend lines, and I realized that the amount of date for each video is different. This is the list of date period available and the number of videos with this date period.

{1: 637, 4: 488, 3: 506, 2: 464, 8: 291, 5: 620, 11: 82, 6: 723, 7: 609, 9: 187, 12: 80, 13: 59, 10: 83, 14: 38, 16: 3, 15: 11}

This outcome indicates that there are many missing data. After researching, I decided to fill in

the missing data with daily average views.

A close up of text on a white background

Description automatically generated

This is the visualization of all video view trend lines after I fill in the data with average daily views. From the graph, it is

obvious that the average data is really lowering the trending views distribution.

After the missing data is filled, I clustered the dataset with K means cluster method. In order to find the best k number, there are many methods that I attempted including plotting the silhouette score of each K, a metric to define if the centroids are close to each data point, changing the missing data into daily max views and daily minimum views. Eventually, I find that cluster of 2

is the best choice.

A close up of a map

Description automatically generatedThis is the graph of centroids of 2 cluster and we can assume that

there is a cluster of high hits videos and a cluster of normal videos.

There are 33 high hits videos and 4848 normal videos. From this, I decided to take a closer look at the high hits videos as they are the ones driving revenues for YouTube and advertisers.

To further analyze the high hits videos, I decided to take a look at the categories and tags.

A close up of a piece of paper

Description automatically generated

This is the distribution of normal videos and high hits videos in different category. It is really hard to find high hits videos as the number is really low.

**A close up of a piece of paper

Description automatically generated**

This is the distribution of categories in high hits videos. Music is the most popular ones in high hits videos contrary to the large amount of entertainment videos in normal video clusters.

After looking at the categories, I searched the popular tags in each categories in high hits and found out that tags are widely spread and there are no tags that appear more than 4 times other than ‘official video’ and ‘trailer’.

**Summary**

After analyzing the data, it appears that music videos are actually the leading videos that have the highest hits and keep getting high hits. Entertainment is the second largest group that create viral videos. In a nutshell, viral videos are really scarce and music videos cost a lot of money. I did not get the best predictors for YouTube advertisers to choose potential premium videos, but with more data and strategy, I might be able to develop a better predictor for advertisers.

**Error Estimates**

* There are a lot of missing data and I filled in a lot of missing data with average, which may cause to influence the result.
* The cluster of 2 is really fragmented with 4848 data of normal trend line and 33 viral videos. I did most of my insights with the viral videos and I did not have a lot of analysis of normal videos.
* There are many columns that I did not utilize.