Youtube Trending Videos Analysis

Data Mining Project

https://github.com/MRDVGroupProjectCSPB4502

Vidya Giri
Post-Bacc CS
CU Boulder
Houston, Texas
United States
vigi7924@colorado.edu

Rouzbeh Salehi Pour
Post-Bacc CS
CU Boulder
Vancouver, BC
Canada
rosa7328@colorado.edu

Michael Taylor
Post-Bacc CS
CU Boulder
Boulder, Colorado
United States
mita6978@colorado.edu

Daniel Shackelford
Post-Bacc CS
CU Boulder
Boulder, Colorado
United States
dash6877@colorado.edu

1. PROBLEM STATEMENT/MOTIVATION

Youtube is an American video streaming platform that serves as a key source of news, entertainment, and reference in the digital age: allowing all ranges of global users to broadcast their videos to the world and view videos that are relevant to their interests. *Variety* magazine states that in order, "to compile its trending-video rankings, YouTube uses a proprietary algorithm that factors in total views, likes, comments and searches."[1] This project intends to identify what determines the performance of a trending YouTube video through the analysis of string and numeric data.

In our project, we have divided the broad topic of characterizing trending videos into three main groups.

Firstly, we will perform sentiment analysis on video performance as it pertains to punctuation (exclamation points, ellipses, and emoji) and sentence case in video titles.

Secondly, we will perform analysis on categories and tags with techniques such as clustering. We are interested in seeing the effect of categories/tags on the trending duration. Additionally, we want to see if videos of the same category trend together in similar time periods.

Third of all, we will analyze factors that correlate with higher viewership such as the video duration, like-to-dislike ratio, video title length, video description length, and date/time posted.

2. LITERATURE SURVEY

There are ten documented projects on Kaggle that use the YouTube trending video dataset. Many of them are redundant exploratory data analysis projects. The three projects that appear to be the most relevant to our project is the "Sentiment Analysis USA (Step by step| Emojis|Tags)" project, "Youtube US Trending Videos EDA" project and the "Word Cloud of South Korea YouTube" project.

2.1 Sentiment Analysis USA (Step by step| Emojis|Tags) Project

The project first tried to identify if there was a relationship between view count, likes and dislikes. A linear relationship between the three values was identified. In both scenarios likes and dislikes lead to increased video view counts, but videos performed better if they were getting likes rather than dislikes. The project also identified which video tags were the best performers and displayed the words in a word

[1] Spangler, T. (2020, December 1). *YouTube Reveals 2020 Top Trending 2020 Videos*. Variety. https://variety.com/2020/digital/news/youtube-top-trending-dave-chappelle-1234842592/.

cloud. Lastly, the project performed emoji analysis, but it is not clear what the creator of the project was trying to accomplish with the emoji analysis.

2.2 Youtube US Trending Videos EDA Project

This project attempts to identify what attributes of videos have a stronger correlation with becoming a high-performing trending video. The project author examined likes, dislikes, comment count, genre, etc along with if the popularity of the video channel the video was published on had an impact.

2.3 Word Cloud of South Korea YouTube Project

This project is mostly written in Korean text which makes it difficult to understand precisely what the creator of the project is trying to achieve. Based on the source code, it appears that the project creator is using words that are nouns gathered from the video titles, channel titles, video tags or video descriptions to compile a list of words to be displayed in a word cloud.

3. PROPOSED WORK

Because the data set is constantly appended to, the first step of work with the data is establishing a baseline from which to work from. This is to ensure that integration of the solution across all team members is successful. Prior to the final release of the solution, an attempt to pull in all data received during analysis efforts can be made to have the most up to date solution.

Once the dataset is defined, a general skim of the data must be completed to ensure data accuracy and consistency across all of the data. Redundant and duplicate data values will need to be scrubbed, along with any rows that do not contain the relevant data needed. It is already known that there are many rows that contain null values. In scenarios where the null values can be filled in with placeholders, such as in the video description category, or where insertion of different values, such as the central tendency of the data set, those values will be filled in to preserve as much pertinent information as possible in our data set. In situations where filling in those values does not make sense, the data will be removed. Data will also need to be scrubbed due to the language barrier. The team will unfortunately not have the requisite language skills to sort through data in other languages, and translation of that data would introduce too many unknowns into the set. All data presented not in english will be removed. Finally, outliers to the set will need to be removed based on standard deviation and overall trend analysis.

In analysis, the data will need to be normalized so that clustering techniques can be used in a number of the tests. In investigating sentiment analysis as it pertains to punctuation, the pertinent information in the dataset will need to be split and parsed so that counts of different punctuation can be captured. In category and tag analysis as it relates to trending duration, data will need to be separated into specific categories, and separate metrics will need to be created for each group. Trending duration denotes time analysis and each category will need to be analyzed over the entirety of its data timeline. In investigating monetization length coinciding with higher viewer count, a correlation matrix will need to be created and the entirety of the data set will need to be sectioned about the monetization time length parameter.

4. DATA SET

The dataset is YouTube Trending Video Dataset (updated daily) from Rishav Sharma on Kaggle.[2]

[2] Sharma, R. (2020). *YouTube Trending Video Dataset (updated daily)*. Kaggle.

The dataset includes the daily trending YouTube videos from 11 countries with up to 200 listed videos from each day. Some attributes that are included are the video title, channel title, date published, video tags, views, likes, dislikes, description, thumbnail, and comment count. The dataset size is 610,322 rows with 17 attributes.

5. EVALUATION METHODS

Our evaluation methods will be dependent on the models and methods used to obtain the results. In the case of evaluating supervised learning algorithms, validation curves can assess if a model is overfitting at a specific parameter value. Confusion matrices can provide a detailed look at how well a model predicts each individual input class. For binary classification models, the values from a confusion matrix can be used to construct an ROC curve to measure the separability of the results. The overall performance of a model can be summarized by its F1 score, which is the harmonic mean of its recall and precision scores, allowing for direct comparisons between different models.

When evaluating unsupervised learning algorithms such as clustering algorithms, the elbow method can identify the optimal number of clusters to generate. Furthermore, silhouette analysis can be performed to measure the separability between the clusters. Since the questions we are asking all refer to data that exists in the dataset (i.e. is labeled), all data points within each cluster can be directly compared to their respective labels. This means that the percentage of each label contained per cluster can be calculated to assess the performance of the clustering algorithm and potentially find new patterns within the data.

In all cases, visualizations will be used in both the initial assessment of the raw data and the final results of the models. These can include histograms to visually assess the distribution of the data, correlation heatmaps and pair plots to compare features between each other, scatterplots to visualize clusters (use PCA for models with more than two features), etc.

6. TOOLS

The bulk of work for this project will be done in Python3. During the data cleaning and preprocessing, tools like Pandas and Numpy will be used heavily to create efficient and accurate data frames to feed into subsequent data analysis techniques.

Analysis of our data will use scikit-learn (sklearn) as its main additional library, along with the aforementioned tools. This tool was chosen because of the team's prior work experience, and overall ease of use and abundant documentation of the library.

Seaborn and MatplotLib will be the main tools used for creation of charts and data visualizations throughout the project lifespan. MatplotLib and Seaborn have been chosen for this application because of their seamless integration with Python, ease of use, and team aptitude.

7. MILESTONES

Project Part	Estimated Completion
2: Proposal Paper	7/4/2021
3: Progress Report	7/11/2021
4: Project Final Report	8/1/2021
5: Project Code and Descriptions	7/25/2021
6: Project Presentation	8/8/2021