CPT_S 315 Final Report Abstract

Washington State University
1st Ethan Villalovoz, B.S. Computer Science, 11751324
2nd Logan Sutton, B.S. Applied Mathematics, 11798384
3rd Berkeley Conkling, B.S Computer Science, 11813150
4th Roy Zabetski, B.S. Computer Science, 11771891
5th Kyle Hawkins, B.S. Software Engineering, 11759096
6th Chance Bradford, B.S. Computer Science, 11720208
7th Matthew Bruggeman, B.S. Computer Science, 11763228
8th Wenjie Wang, B.S. Computer Science, 11361265
9th Silvestre Pamatz-Rangel, B.S. Computer Science, 11731487

Introduction:

YouTube is one of the largest online entertainment industries. Anyone with a Google account can upload short and long-form videos to the platform. People who have found success on the platform, often called content creators, post videos and hope to make it on YouTube's trending page. What do all of these trending videos have in common? What statistics make the YouTube algorithm classify a video as trending? Our research will use association rule mining, linear regression, and classification to reveal the commonalities between YouTube's trending videos. As a result, we want to allow content creators to understand what makes a video get on the YouTube trending page.

Our agenda:

We will use three algorithms to analyze YouTube statistics datasets and conclude which attributes of a YouTube video make it popular among the general audience. Additionally, given the data on trending videos, we want to predict whether or not a video has the potential to become trending.

Our methods:

The dataset we are using is named "Trending YouTube Video Statistics and Comments" (https://www.kaggle.com/datasets/datasnaek/youtube) [1]. It includes data gathered from videos on YouTube that are within the trending category each day.

We plan on using three data mining algorithms:

1. **Association Rule Mining (Apriori)**: Association rule mining techniques such as the Apriori algorithm [2] can be applied to uncover frequent item sets or patterns among categorical variables like video tags or categories. This helps identify associations between different attributes and understand viewer preferences.
2. **Regression Analysis (Linear Regression)**: Linear regression [3] can be used to model the relationship between numerical variables such as views, likes, and comments. It helps understand how changes in one variable affect another, thus providing insights into factors influencing video popularity.
3. **Classification (Logistic Regression):** Logistic Regression [4] is a statistical algorithm used for binary classification. It models the probability of a binary outcome by fitting a logistic function to the observed data. We will use this algorithm to predict whether a video will trend based on its features (e.g., likes, dislikes, views, comments, etc.).

There are also several key questions we asked ourselves when deciding which data to mine and collect:

- What are the key factors contributing to a video trending on YouTube?
- Are there any patterns in the characteristics of videos that tend to trend on YouTube (i.e., length, category, language)?
- How does viewer engagement (likes, dislikes, comments) correlate with video performance (views, trending duration)? Can we use this to predict the video's view count?
- Are there any notable differences in viewer engagement across different video categories (e.g., music, gaming, entertainment)?
- Can sentiment analysis of comments help predict the success of a YouTube video?
- Are there any temporal patterns in video trends, such as certain times or days of the week when videos are more likely to trend?
- How does the title (or keywords within the title) impact a video's view count?
- Are there any correlations between viewers' geographic locations and the types of videos that trend in those regions?
- Can we identify influential creators or channels based on their video performance metrics and viewer engagement?
- What emotional conclusions can we draw from performing sentiment analysis in various forms? Do viewers feel happy, amused, or other emotions?

Answering these questions will allow us to collect meaningful data for content creators.

What we expect to find:

We expect to find a correlation between datasets that may reveal answers to the questions we set out to solve, such as:

- What aspects distinguish a popular video
- Association rules between different video attributes such as likes, dislikes, comments, and views
- Correlations between specific video categories and user engagement metrics
- Possible patterns in viewer behavior
- Prediction models to estimate the number of views based on stats like (video title, description, number of comments)
- Identifying characteristics that lead to a video receiving a large number of likes, comments, or views
- Impact of video length, creator upload frequency, and content type for user engagement metrics

By utilizing the three data mining algorithms (Apriori, Linear Regression, Classification), we can gain valuable insights to help content creators, Advertisers, and YouTube data Analysts learn what factors contribute to the success of trending videos. This knowledge can then be utilized to optimize content strategies by these findings.

References:

[1] “Trending YouTube Video Statistics and Comments,” Kaggle, Oct. 25, 2017.
https://www.kaggle.com/datasets/datasnaek/youtube

[2] Wei, Dr. Honghao “Mining Frequent Patterns, Association and Correlations: Basic Concepts and Methods.” Class lecture, CPT_S 315 Introduction to Data Mining, Washington State University, Pullman, Washington, February 5, 2024.

[3] GfG, “Linear Regression in Machine learning,” GeeksforGeeks, Mar. 14, 2024. https://www.geeksforgeeks.org/ml-linear-regression/

[4] GfG, “Logistic regression in machine learning,” GeeksforGeeks, Jan. 30, 2024. https://www.geeksforgeeks.org/understanding-logistic-regression/