Word-Processing with the YouTube Algorithm

Introduction:

The purpose of this project is to utilize R and run a word-processing program to simulate how YouTube runs its video recommending process as well as identify which keywords may be associated with certain genres/categories of videos to better display titles for videos to both new and old content creators to gather more views and expand their channel. By utilizing certain keywords within the title of your videos and based on keywords that are used in videos you watch, you should receive recommendations with similar words/categories. This process doesn’t seem to have been taken on by anyone and is uniquely focusing on text/word processing whereas there has been an instance which utilized running YouTubes API algorithm with videos instead of text.

Review:

Lovejoy, Chris. “I Created My Own YouTube Algorithm (to Stop Me Wasting Time).” *Chris Lovejoy*, Chris Lovejoy, 12 Nov. 2021, https://chrislovejoy.me/youtube-algorithm/.

Lovejoy uses this page to describe his process of experimenting with creating his own algorithm to determine which videos are valuable to watch and be suggested. He uses code within python to execute the metrics he collected via the YouTube API. Following this, he then uses metrics such as views per video, subscribers, comment count, etc. to filter out the top-quality videos based on searching with a specific word or set of words. He finds his algorithm is correctly identifying videos that he would enjoy and begins to run it through different mediums such as AWS Lambda. Potential next steps for Lovejoy include text/term searching which would lower the dependency on how videos will be picked up in the search. According to Lovejoy, the code is slow so having something optimized better would create a more efficient experience. Based on the results he has, there are no observable visualizations that would help to filter out results of terms or keywords and the project falls short in observing this.

In conclusion, there hasn’t been much done in trying to observe the best way to filter out the most popular/best quality videos in a similar way to how YouTube runs its algorithm and with being able to visualize the results based on certain keywords and terms for specific categories and genres. Based on research, we would be trying a new approach to an interesting problem which has been touched by few.

Link To Data: <https://www.kaggle.com/datasets/datasnaek/youtube-new?resource=download&select=USvideos.csv>

Dataset:

One of the toughest challenges in prepping for this project was finding usable data to help us in understanding how we can try and simulate the YouTube algorithm. Of course, with this idea we had to make sure we had videos, lots of videos. The dataset we ended up utilizing actually has a very good set of variables which would assist us in processing the text used in the title of the video to associate it with a particular category type, for example, we have Automotive, Food/Culinary, Gaming, and Film & Movies. In total we have around 40+ categories of videos which will be separated and then used to create separate datasets where only videos with a specific category are shown and almost 41000 videos which we will be sorting through.

Within the dataset, we notice that there are multiple columns listed, for example we have the number of views, comments, likes, and dislikes. In addition, we also have the title and the tags as well, which we can use to help breakdown the text processing for the YouTube algorithm. Our focus will be the YouTube video title column, which is the primary variable of interest and secondly, will be the video category. Using these 2 variables, we should be able to create some interesting visualizations which show the common word usage for all the different categories of videos. Some other visuals we can observe is the breakdown of video likes/views based on the video category type. This will allow us to gain an understanding as to which category seems to have the most presence when surfing YouTube.

However, even with all this data, we do have many variables which are not going to provide any function to our problem, such as the trending date of the video, comments disabled, and the posted time of the video. These are just a few of the variables within the dataset which we will be removing to better work with what we have planned.

Methods:

The methods that we plan on implementing for this project include running a Machine Learning text processing algorithm which will pick up the most common nouns, adjectives, and verbs used per category where it will be created into a visualization that displays this information. The algorithm will clean out stop words, and place holder words so that these words won’t appear as a commonly used word since they’re used in every other sentence/title. In addition, we will be visualizing the count for each word per category as well as seeing which category has stayed the most popular over the timespan of the dataset.

Pros/Cons:

When working with these methods, we have a handful of pros which we can elaborate on. For example, as mentioned above, we have the algorithm removing basic filler words such as “and”, “the”, “to”, “they”, etc. This makes the algorithm process the text faster and more efficiently for us to use. We can also utilize the visualization to give us better insight into how the data is distributed and understanding the text used per category of videos to get an idea of how we can expect the titles to be used if we were to create our own YouTube video for a specific category. Giving us an open look into how each category performs will ultimately lead us to understanding how this YouTube algorithm works when we use the platform.