Data Nutrition

Introduction:

This project will analyze YouTube video data from 2017 to 2018 to analyze people's video preferences in the pre-COVID era. The data set includes the status of each YouTube video's status from 2017 October 13th to 2018 December 31th, which is just two years before the Covid. By using this data set, we can analyze what kind of videos are sought after by people before Covid-19. The data sets tallied the videos' likes, dislikes, comment count, like-view ratio, and video views as the dependent variables and their category as the independent variable. We decided to use the data from the US to analyze. From the data set, we can have an overall insight that how's popular videos in YouTube trending before Covid-19.

—----------

Created by: Mitchell J

Content: An article with csv file.

Source: https://www.kaggle.com/datasets/datasnaek/youtube-new/

Reference source:

https://www.kaggle.com/datasets/themrityunjaypathak/most-subscribed-1000-youtube-ch

—----------

Use Cases

Potential real-world applications of the dataset

- What category of videos should a new YouTuber in the US create to be popular?
- What categories of video are becoming more popular in the US?
- What categories of video did people more likely to notice before COVID in the US?
- If people want to run a channel on YouTube before Covid, what kind of videos may be better for them?
- With the data focusing on video trends before Covid, what can we suppose about how Covid-19 affects people's life from the trends of video on YouTube?

—----------

Description:

TELL US ABOUT THIS DATASET

This data set includes the status of each YouTubers' status constructed by Mitchell J and clean by

Yukang Zhao, who is our team member.

IS THERE AN INTENDED PURPOSE FOR THE DATASET? WHAT DOMAIN WAS IT DESIGNED FOR?

This data is expected to be used to analyze the popularity of Youtube videos in the United States, with the aim of predicting future video popularity. This data can also be used to analyze people's preferences for Youtube videos before COVID.

ARE THERE ANY ADDITIONAL RESTRICTIONS UNDER WHICH THE DATASET IS MADE AVAILABLE? ARE THERE ANY LEGISLATION, CODES OF PRACTICE OR GUIDANCE FROM THE JURISDICTION OR DOMAIN IN WHICH THE DATA WAS COLLECTED ABOUT THE USE OF THIS DATA? ARE THERE TASKS FOR WHICH THE DATASET CANNOT BE USED? PLEASE PROVIDE A DESCRIPTION OR GUIDANCE.

In order to engage with this issue, we used diversified methods to gain a certain in-depth understanding on this problem. Fortunately, we came to a quite positive conclusion, which means no, and dispelled concerns about future related research.

ARE THERE TASKS FOR WHICH THE DATASET WOULD BE CAUTIONED AGAINST BEING USED? IF SO, PLEASE PROVIDE A DESCRIPTION.

Fortunately, for the subject of this study, this unfortunate situation did not occur in this study.

HAS THE DATASET BEEN USED FOR ANY TASKS ALREADY? IF SO, PLEASE PROVIDE A DESCRIPTION AND LINKS TO PAPERS OR SYSTEMS USING THE DATASET

Unfortunately, this data is not currently being used by other researchers, leaving it as an undiscovered treasure. We look forward to the day when this data set can exert its due value and make corresponding contributions to cultural development.

WILL THE DATASET BE UPDATED (E.G., TO CORRECT LABELING ERRORS, ADD NEW INSTANCES, DELETE INSTANCES)? IF SO, PLEASE DESCRIBE HOW OFTEN, BY WHOM, AND HOW UPDATES WILL BE COMMUNICATED TO USERS (E.G., MAILING LIST, GITHUB)?

Data is not updated yearly because it is static data.

IS THERE A MECHANISM THROUGH WHICH AN INDIVIDUAL CAN REQUEST THEIR PERSONAL INFORMATION BE REMOVED?

We pay attention to the protection of personal privacy. This data set will not contain sensitive personal information such as names, so there is no mechanism for deleting personal information - because they will not even find themselves in this data set.

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

It represents the US videos published in 2017 and 2018, which is the reason why we claim that it also can be used for researching the video preference of people in the US before COVID 19.

How many instances are there in total (of each type, if appropriate)?

There are 23778 instances in total in the dataset. We are glad that we can have such a large amount of data to reduce the potential bias.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

This data set doesn't contain all possible instances, it is a sample from a larger set.

Are relationships between individual instances made explicit?

Every individual instance can be generally treated as independent of each other. There's no doubt that current affairs can influence people's focus, but including this in our research is out of our ability, so we decided to treat them as individuals independently.

—----------

Alert

1. The data is from the whole United State, which means it is not referring specifically to a state. Different states may have different situations.
2. All the data are tallied randomly so the results of the data do not represent any racial preferences.
3. Since the data is a sample from a larger set, it is not representative to any specific study filed in digital media.
4. There are videos that may include controversial information or topics, which are Irrelevant to the purpose of our study.