YouTube Trending Dataset Analysis

Sam Logsdon, Maryam Bokhari, Jeong Rae Park

***Abstract*—The Youtube Trending Videos Dataset [5] is a snapshot of the trending videos on Youtube almost every day from In this report, we examine the dataset for fascinating trends. Particularly, we looked at the age of trending videos, the popularity of different categories, shared trending videos between countries, and videos that trended globally in all countries. We also discuss problems with the dataset and how we approached them.**

I. Introduction

To be ranked on the ytrending videos list, Youtube takes many factors into account, including age of the video, view accumulation speed, where the views are coming from, and more. The list is further filtered to prevent inappropriate content from being added to the list, including using staff as a ”final filter” in ”the US and other locales”. [6] In examining the trending dataset, we discovered a great deal of diversity and some surprising trends, which are illustrated later in this report.

II. Dataset Description

A. File Layout

The dataset consists of ten CSV files containing video data per country and ten JSON files containing category information per country. The data is a total size of 514 MB.

B. Data Scheme

Each csv file contains trending video data for a particular country collected from November 14th, 2017 to June 14th, 2018. Each row represents an instance of a trending video on a particular day, and contains information such as likes, dislikes, category_id, and views for the video.

III. Preprocessing

The data for each country was loaded into a single sqlite database table with an additional row identifying the country, which was extracted from the title of the file. Unicode errors were replaced with backslash encodings using python's file open function. Boolean columns were coerced from string to boolean types, and the trending_date column was parsed into a datetime type.

Similarly, category information was loaded into a separate table. One thing to note is that while there is a category file for each country, the only difference is that certain categories are unavailable in all countries. Otherwise, the ID numbers map to the same categories for all countries, so we were able to flatten the category data into a single table without preserving country information.

There were a number of videos with a video_id of *”#NAME?”*, an error associated with Excel. We were able to correct for this using the thumbnail_link, which contained the video_id. All of the identifiers from this set started with the ’’ symbol, which makes it likely that Excel attempted to convert them to formulas.

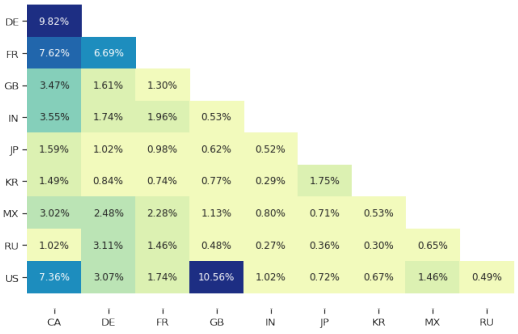

Fig. 1. Percentage of Videos shared among countries

IV. Analysis

A. Tools

For our analysis we primarily used the pandas [4] library, with matplotlib [3] and seaborn for plotting. We additionally used scikit-learn [2] for it's datamining capabilities.

B. Shared videos between countries

To perform this analysis, an SQL query was constructed which self-joined the video table on the basis of *video_id*, excluding entries where a row was joined to itself. Using this result set, a pandas dataframe was constructed with

We found that there is generally only a small overlap in trending videos between each country, with Germany and Canada topping out at a little over 9%. When comparing the average ratio of videos countries shared with each other, Canada ranked the highest at 3.8%. Figure 1 shows you the heatmap of shared videos between countries, and Table 1 gives you the average ratio for a given country.

C. Average Time To Trend

In order to figure out the average time it takes for a video to go from published to trending, We sorted the data by trending date and kept only the first instances of a video's trending_date. We then subtracted the normalized published timestamp from this first trending date for each video, and calculated the average age per trending_date in the dataset. What became clear initially was that there was a large amount of variance in this set, with videos as old as 2006 appearing

TABLE I
AVERAGE % OF VIDEOS SHARED WITH OTHER COUNTRIES

country	avg % shared
CA	0.038293
DE	0.030026
FR	0.024103
US	0.021806
IN	0.019080
MX	0.018557
GB	0.016199
RU	0.015276
JP	0.013658
KR	0.013325

on the trending list in 2017. This variance drops dramatically in March of 2018, presumably as Google changed their algorithms to exclusively trend younger videos.

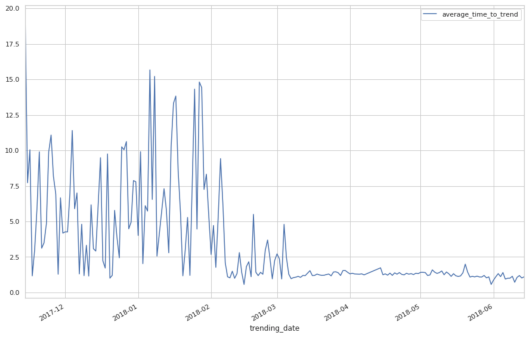

Fig. 2. Change in Average time over years

D. Popular Categories by Country

We are interested in viewing list of categories which is most likely to become trending. Since the videos and the categories are in two different data frames, we merged them together. After that, we took the top categories in each country and counted the number of videos in each categories. The boxplot generates a distribution of all the categories while highlighting the top five with a swarm plot. As shown in figure (b), The "Entertainment" category is the most trending among all countries except "Russia".

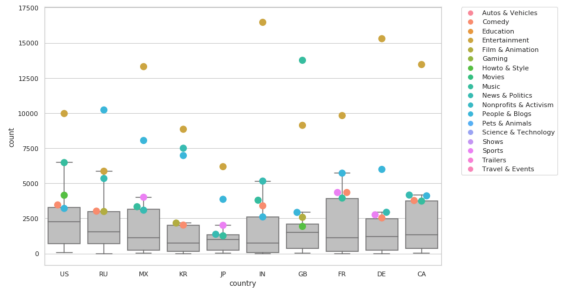

Fig. 3. Top 5 categories in each country

E. Trending collection of Videos in all countries

We are interested in viewing the proportion of popular videos trended at same time in all countries. After loading the data frame, we used an SQL query to group the date based on the video_id and trending date and counted the instances of videos. The videos which had a count number over 10 were trending in all countries at the same time. After generating the table, We noticed that some videos trended more than once. Although, We initially believed that there would be a multitude of videos, however we only had a total of 45 rows where 38 of the videos were unique.

title	channel_title	trending_date
[OFFICIAL VIDEO] HAVANA - PENTATONIX	PTXofficial	2018-02-24T00:00:00
VENOM - Official Trailer (HD)	Sony Pictures Entertainment	2018-04-25T00:00:00
VENOM - Official Teaser Trailer (HD)	Sony Pictures Entertainment	2018-02-09T00:00:00
The Weeknd - Call Out My Name (Official Video)	TheWeekndVEVO	2018-04-14T00:00:00
The ULTIMATE $30,000 Gaming PC Setup	Unbox Therapy	2018-05-14T00:00:00
The Chainsmokers, Drew Love - Somebody (Official Video)	ChainsmokersVEVO	2018-04-21T00:00:00
Taylor Swift - Delicate	TaylorSwiftVEVO	2018-03-13T00:00:00
Taylor Swift - Delicate	TaylorSwiftVEVO	2018-03-14T00:00:00
Selena Gomez - Back To You	SelenaGomezVEVO	2018-06-07T00:00:00
SPIDER-MAN: INTO THE SPIDER-VERSE - Official Trailer (HD)	Sony Pictures Entertainment	2018-06-08T00:00:00
SPIDER-MAN: INTO THE SPIDER-VERSE - Official Trailer (HD)	Sony Pictures Entertainment	2018-06-07T00:00:00
Nicki Minaj - Chun-Li	NickiMinajAtVEVO	2018-05-05T00:00:00

Fig. 4. Sample Popular Video Titles

F. Using Machine Learning to Predict Categories

We attempted to predict categories from descriptions, primarily using Stochastic Gradient Descent due to its efficency with large data sets. Other models such as Kmeans Nearest Neighbor were cost prohibitive to run, and Naive Bayes tended to do worse. Suprisingly, training the data with both the description and the country was no better than the description alone, and adding the title didn't help much either. We also attempted to tune the parameters on the TfidfVectorizer in the feature extraction stage, with ruinous results.

In the end, our best prediction model topped out at a 73% weighted average, and it stands to reason this dataset isn't particularly ripe for machine learning on its' own.

REFERENCES

[1] S. V. D. Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," Computing in Science & Engineering, vol. 13, no. 2, pp. 22–30, 2011.

[2] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011)

[3] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," in Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, May-June 2007. doi: 10.1109/MCSE.2007.55

[4] Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010)

[5] M. J, "Trending YouTube Video Statistics," Kaggle, 15-Jun-2019. [Online]. Available: https://www.kaggle.com/datasnaek/youtube-new. [Accessed: 10-Dec-2019].

[6] "Trending on YouTube - YouTube Help," Google. [Online]. Available: https://support.google.com/youtube/answer/7239739?hl=en. [Accessed: 10-Dec-2019].