YOUTUBE TRENDS DATASET:

======================================

=========

1. CONTENT

YOUTUBE (THE WORLD-FAMOUS VIDEO SHARING WEBSITE) MAINTAINS A LIST OF THE TOP TRENDING VIDEOS ON THE PLATFORM.

ACCORDING TO VARIETY MAGAZINE, “TO DETERMINE THE YEAR’S TOP-TRENDING VIDEOS, YOUTUBE USES A COMBINATION OF FACTORS INCLUDING MEASURING USERS INTERACTIONS (NUMBER OF VIEWS, SHARES, COMMENTS AND LIKES). NOTE THAT THEY’RE NOT THE MOST-VIEWED VIDEOS OVERALL FOR THE CALENDAR YEAR”.

TOP PERFORMERS ON THE YOUTUBE TRENDING LIST ARE MUSIC VIDEOS (SUCH AS THE FAMOUSLY VIRILE “GANGAM STYLE”), CELEBRITY AND/OR REALITY TV PERFORMANCES, AND THE RANDOM DUDE-WITH-A-CAMERA VIRAL VIDEOS THAT YOUTUBE IS WELL-KNOWN FOR.

2. ABOUT THIS FILE

Geography: Worldwide.

Time period: 2017-2018

Features of analysis: TREND RANK

This dataset includes several months (and counting) of data on daily trending YouTube videos.

Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day. Includes data from RU, MX, KR, JP and IN regions (Russia, Mexico, South Korea, Japan and India respectively) over the same time period.

Each region’s data is in a separate file:

Variables: 16 variables

AS WE HAVE 16 COLUMNS IN THE DATA, THE FIRST10 SAMPLE COLUMNS ARE:

THE DATASET INCLUDES DATA GATHERED FROM VIDEOS ON YOUTUBE THAT ARE CONTAINED WITHIN THE TRENDING CATEGORY EACH DAY:

- video_id (Common id field to both comment and video csv files)
- title
- channel_title
- category_id (Can be looked up using the included JSON files)
- tags (Separated by | character, [none] is displayed if there are no tags)
- views
- likes
- dislikes
- thumbnail_link
- date (Formatted like so: [day].[month])
- comment_text
- likes
- replies

3. PROBLEM STATEMENT

PUTTING YOUTUBE_VIDEO_TRENDS IN CONTEXT

Possible uses for this dataset could include:

- ANALYSING WHAT FACTORS AFFECT HOW POPULAR A YOUTUBE VIDEO WILL BE.

- A DASHBOARD FOR VISUALIZING VIDEO TRENDING ANALYSIS

- PREDICTION ALGORITHMS FOR REGRESSION & CLASSIFICATION

- CAN BE EXTENDED TO SENTIMENT ANALYSIS OF TRENDING VIDEO’S CONTENT DESCRIPTION

- AND OTHER MULTIPLE INFERENCES CAN BUILD A CUMMULATIVE PROBLEM STATEMENT

TIME ORIENTED:

- What are the lengths of trending video titles? Is this length related to the video becoming trendy?
- How are views, likes, dislikes, comment count, title length, and other attributes correlate with (relate to) each other? How are they connected?
- Which YouTube channels have the largest number of trending videos?
- Which video category (e.g. Entertainment, Gaming, Comedy, etc.) has the largest number of trending videos?
- When were trending videos published? On which days of the week? at which times of the day?
- Whats the most frequent type of video?
- Best time to publish videos?

INTRO LEVEL NLP ORIENTED:

- Whats the most frequent names in title, description, tags?
- How many trending videos contain a fully-capitalized word in their titles?
- What are the most common words in trending video titles?
- Most common words in video titles?

4. ROUGH APPROACH

- Extensive EDA and research from YouTube’s regional wise data.
- Time Series analysis based on the varying video trends.
- Beginner level NLP implementation
- Usage of K-means clustering in order to cluster the video trends based on geographic location, number of views, genre type,…..
- Categorising YouTube videos based on their comments and statistics.
- Feature Extraction to derive quantifiable attributes of YouTube Trends
- Analysing what factors affect how popular a YouTube video could be.

- Statistical analysis demonstrated through TIME_SERIES_ANALYSIS.
- Training ML algorithms:
  - REGRESSION PROBLEMS: to predict the no_of_days_to_trend
  - CLASSIFICATION PROBLEMS: to classify the intensity of trends(L/M/H) through derived quantifying metrics
- Sentiment analysis in a variety of forms
  - To group the data into segments based on the sense of comments
  - To segregate the pattern of trends based on the genre extracted from video descriptions

ADDITIONAL:

- NLP can be used for WORD-CLOUD
- TIME SERIES also can be used for PLOTTING
- TREE MAPS generation for explainability purposes

5. THEORETICAL PUBLICATIONS:

THE GTD HAS BEEN LEVERAGED EXTENSIVELY IN SCHOLARLY PUBLICATIONS, REPORTS, AND MEDIA ARTICLES.

- *HTTPS://VARIETY.COM/2017/DIGITAL/NEWS/YOUTUBE-2017-TOP-TRENDING-VIDEOS-MUSIC-VIDEOS-1202631416/GLOBAL TERRORISM DATABASE CODEBOOK*

- *HTTPS://WWW.YOUTUBE.COM/FEED/TRENDING*