Team Members: Aravind Patnam, Jeremy Tan, Timothy Le

Trending Here, Trending There

An analysis on trending and non trending youtube videos.

The questions we explored included a general exploratory data analysis and as well some in depth NLP. Some of the EDA included cleaning and find numeric data, such as the ratio of likes to dislikes between each country, how long a video stayed trending, which are the most popular and unpopular youtube channels, finding the most popular youtube videos, finding which country has the most dynamic and active audience, and other general stats.

Another piece of our project included an in depth NLP analysis which included finding the most popular words in tags/descriptions, conducting a sentiment analysis, and also running a LDA topic inference model on them. By doing this, we were able to predict what sort of videos get to trending status and which do not. These are clearly reflected in the sentiment analysis and LDA visualization in our notebook.

The data sources we used included the datasets we found on Kaggle(https://www.kaggle.com/datasnaek/youtube-new/), and also generating our own using the youtube api. The ones on Kaggle only contained trending videos statistics, so to perform a holistic analysis, we utilized the youtube api and fine tuned our request payload to get very similar videos that were not trending. This allowed us to conduct a full blown analysis. The list of trending datasets are contained in the zip file and have the word "**trending"** in them, and the non-trending ones have the keyword "**not_trending"** in them. We also provided a bank of positive and negative words which were used for the sentiment analysis.

Some of the places that helped us produce our code include the following:

https://realpython.com/python-data-visualization-bokeh/

https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html

https://towardsdatascience.com/the-next-level-of-data-visualization-in-python-dd6e99039d5e

https://towardsdatascience.com/basic-binary-sentiment-analysis-using-nltk-c94ba17ae386

https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

Some of the viz references include:

https://medium.com/@ChannelMeter/youtubes-top-countries-47b0d26dded

https://matplotlib.org/mpl_toolkits/mplot3d/tutorial.html

https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/

Team Members: Aravind Patnam, Jeremy Tan, Timothy Le

Some of the final results from the sentiment analysis and LDA are that most trending videos have a negative sentiment while most non-trending videos have a positive sentiment. These sentiments are also represented via the topics that we found to be the most popular through our LDA model.

Some of the libraries we used include: Pandas, Numpy, Bokeh, Scikit, nltk, requests, spacy, genism, pickle, matplotlib, and pyplot.

Some of the more notable tasks that were accomplished included conducting a sentiment analysis using nltk and also the LDA topic inference model using spacy and genism. Since they were some of the stretch goals for our project, they were pretty hard to implement as they took a lot of time and trial and error. We had to make sure all our parameters for the model training were fine tuned properly. Additionally, another common struggle we all had was cleaning the input data so that we could feed it into the model or the other visualizations. Often times, the inputs required were very specific and getting data into that format was rather time consuming. Furthermore, fine tuning our visualizations was a large learning curve, yet a fruitful one as we were able to produce some stunning visualizations.