CSC 422 Project Proposal: Effect of Titles on Trending YouTube Videos

Sagnik Nayak, Vinhson Phan, Rhea John
NC State University, March 2021

1 DATASET DESCRIPTION

Trending Youtube Video Statistics:
https://www.kaggle.com/datasnaek/youtube-new
This dataset contains a list of top trending videos on Youtube with the following attributes: video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

2 PROJECT IDEA

We want to use the Youtube Trending data set to study what, and how different attributes of a video title affect the ranking of a video in the trending section. We will use *Natural Language Processing* techniques to analyze various parts of the string title. Attributes like capitalization, sentiment, title length, etc. An example approach to our strategy would be to look at words that were capitalized in the title and find patterns for the same word. Hence, infer how that word affects the Youtube Trending rank vs videos without that word, and videos with that word without the capitalization. Our study aims to find how different titles affect the way the Youtube algorithm behaves to certain videos and therefore how it makes it to the trending page which directly affects the viewership of the video.

Our data set contains information on trending videos with attributes like title, likes, dislikes, tags, etc. We will pre-process this data set to cater to our needs of a title, ranking on the trending page, and the amount of time on the trending page.

3 SOFTWARE WE WILL WRITE

For our project, we will write a pre-processing script, an NLP learning model for the video title, analysis code, and visualization script. The pre-processing script will set up our data to best suit our learning model needs i.e. filter out video titles, like dislikes, video rank, etc. and remove irrelevant attributes, account for missing data, and feature scale any attribues if necessary. There could be some dimensionality reduction as well. We are yet to decide on the NLP learning model and what algorithms we will use. We'll internally research the exosting technology by reading papers and looking up necessary information to better understand the NLP field of study. Our visualization script will use existing libraries and pandas and matplotlib to visualize our findings like most popular words, capitalization trends, and frequency plots of best performing title words and phrases. We will use Jupyter for our code as well as common ML frameworks like numpy, scikit-learn, NLTK, etc.

4 WORK DIVISION

Due to our unfamiliarity with many of the technologies, we will be on zoom to pair-program together and manage the work through a project management plan like Kanban/Trello. We plan to meet Tuesday/Thursday after class every week up until the due date of the project.
Tasks to consider for the project are :-

- Data import and processing strategization
- Experimental design for the project
- Trade study of learning algorithms relevant to our project
- Implementation of selected learning algorithm using existing libraries and self implementation as necessary
- Testing and validation of the learning model
- Visualization of study
- Inferences from the study
- Research paper content coompilation
- Concluding housekeeping tasks

5 MIDTERM REPORT DELIVERABLES

For the Midterm Report, we plan to finish pre-processing the data, analyze one aspect of the title (capitalization, sentiment, etc) and its effect on popularity, and visualization for this one aspect.

REFERENCES

- How Long Will Your Videos Remain Popular? Empirical Study of the Impact of Video Features on YouTube Trending Using Deep Learning Methodologies
  https://bit.ly/38VMdF1

- Trending Pattern Identification of YouTube Gaming Channels Using Sentiment Analysis
  https://rb.gy/6sj3ek

- A First Step Towards Understanding Popularity in Youtube
  https://ieeexplore.ieee.org/abstract/document/5466701
  This paper discusses its findings on how users exhibit a behavior of liking, commenting, and adding a video to their favorites every 400 times a video is viewed.

- Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques
  https://rb.gy/6sj3ek