[ECS 175] Report

Group 8

12/07/2020

Introduction (1 page)

Problem Statement

Our project aims to predict whether a given YouTube video is classified as "trending" or "non-trending" by Google. We did this by creating our own dataset through Google's YouTube Data v3 API which contained present and historical information about a video. This problem is a binary classification problem as we are predicting one of two classes. Thus, we implemented prediction with both Neural Network and Logistic Regression models, and compared the two to find the best model overall. In addition, we expanded the problem scope from simple binary classification to multiclass classification by utilizing clustering. Although this does not directly solve our original problem, it gave us avenue to predict a YouTube video's success.

Group Members

Data Gathering Sub-Group:

- **Contribution:** Wrote and manage script to get data from Youtube Data API
- **Members:** Ted Kahl, Rohail Asad

Data Processing Sub-Group:

- **Contribution:** Processed data from API into effective dataset
- **Members:** Phalgun Krishna, Prajwal Singh, Seth Damany

Machine Learning Sub-Group:

- **Contribution:** Created and optimized ML models
- **Members:** Cameron Yuen, Owen Gao, Theresa Nowack, Trevor Carpenter

Documentation and Web App Sub-Group:

- **Contribution:** Created web app and managed documentation
- **Members:** Josh McGinnis, Keith Choung, Nikhil Razdan, Thu Vo

Group Leader:

- **Contribution:** Organized milestones and facilitated communication
- **Members:** Nikhil Razdan

Project Links

- **GitHub:** www.github.com/nsrazdan/ECS-171-Group8
- **YouTube Data v3 API:** developers.google.com/youtube/v3/
- **Original Kaggle Dataset:** www.kaggle.com/datasnaek/youtube-new

Literature Review (1 page)

Primary Source

In looking at other sources with similar topics, we found that various studies defined popularity differently. Our group used YouTube's trending feature as our popularity metric. Another similar study from Stanford used videos' view counts instead. This study, titled *YouTube Videos Prediction: Will this video be popular?*, analyzed the topic of predicting a video's success from the perspective of a YouTube content creator. Instead of a simple binary classification "trending?" target variable like our group utilized, the researchers Li, Kent, and Zhang from Stanford divided the classification into 4 disjoint categories: non-popular, overwhelming praises, overwhelming bad views, and neutral videos. They used multi-class classification to analyse and quantify the success of a video in seeing whether it was reacted to positively or negatively. The researchers used similar attributes as our group, such as view count and duration. However, they also included an important attribute that fits in line with their 'YouTuber' point of view: time gap. This takes into account the possibility that frequent and regular uploads are favored by YouTube's recommendation/trending selection process as opposed to a sporadic upload schedule. By using the YouTube dataset from Kaggle that inspired our project's dataset, this particular research paper bears many similarities with our own.

Li, Kent, and Zhang found that extreme gradient boosting with attributes time gap, category, and description produced the best results with the highest F1 score out of the other methods they tried. Instead of downsampling to account for the imbalanced data, the team of researchers added class weights. They report that the highest indicators of popularity (eg. view count) are a video's category, description and time gap. When concluding their research paper, they note that the issue of overfitting remains a concern. To improve, they suggested adding more attributes such as video thumbnails and subtitles, as well as expanding the dataset for a more balanced set of videos. This particular source provides a clear framework to position our own project. Importantly, we added a large set of non-trending videos in our data to give it enough information to classify trending status.

http://cs229.stanford.edu/proj2019aut/data/assignment$_3$08832$_r$aw/26647615.*pdf*

More Related Works

The Towards Data Science Article by Arvind Srinivasan, YouTube Views Predictor, discussed a model that utilizes very interesting and unique variables our group had not thought of in its prediction model. Although the two projects differ slightly given that the "YouTube Views Predictor" model predicts the number of views a YouTube video will get and ours predicts whether a YouTube video will become trending, many of our features remain the same. However, the creators included other features such as a "Clickbait Score," a NSFW Score," and whether a YouTube video's title contained words related to common or popular YouTube genres to better determine its popularity.

https://towardsdatascience.com/youtube-views-predictor-9ec573090acb

Clustering was a method that the Machine Learning team would like to have implemented in order to discover groups of videos which would allow us to further focus the scope of our predictions of trendability. *Clustering the Unknown - The Youtube Case* by Amit Dvir, Angelos K. Marnerides, Ran Dubin, Nehor Golan did exactly that and took 100,000 video streams to cluster unknown videos based on their title and grouped them with the use of K-means clustering and the help of NLP formulations and Word2Vec. They were able to identify many unique clusters that had their own traits purely based on video title and not by any other traits given by the metadata of youtube videos.

https://www.researchgate.net/publication/332376497$_C$lustering$_t$he$_U$nknown$_-T$he$_Y$outube$_C$ase

Trending Videos: Measurement and Analysis studies Youtube's trending videos in terms of viewership lifecycle and other basic statistics of their content. Researchers also collected a list of non-trending videos in order to do comparative analysis between trending and non-trending videos. To distinguish the difference between trending and non-trending videos, they conducted comparative analysis on (1) the standard video feeds, which provide basic statistics of the videos and (2) video uploaders' profile. Moreover, the study used Granger Causality (GC), which provides deeper insight onto viewership pattern, to derive directional-relationships

among trending-video time-series. The study concluded that there's a distinct difference between the statistical attributes of trending and non-trending videos. GC measurement confirms the directional relationship between trending videos and other videos in the dataset, and among different categories of trending videos.

https://www.researchgate.net/publication/266262149$_{T}rending_{V}ideos_{M}easurement_{a}nd_{A}nalysis$

Dataset Description (0.5 pages)

Proposed solution and experimental results (4-5 pages)

Conclusion and discussion (0.5 pages)

Reference (unlimited pages)