**Project Description**

Github Repository Link: <https://github.com/ssharif6/info370-FinalProject>

**Summary**

As described by the official [YouTube Trends blog](http://youtube-trends.blogspot.com/p/about-youtube-trends.html), videos that are featured as ‘trending’ get a significant number of views on both external websites and the YouTube site itself.

The question we are trying to explore is: does tag content impact the popularity of the video on YouTube. That is, would similar videos with varying tags experience varying level of viewership? More specifically, we are defining popularity as a function of total views, number of likes/dislikes, and number of comments.

**Full Description**

Our goal is to determine how to effectively utilize tags in order to maximize the popularity of a YouTube video. In light of recent political events, we have realized that social media has the potential to make a large impact on society. Thus, this projects fits into the broader family of social media research that analyzes how tags (which includes tagging content to represent a certain category, and explicit hashtags in descriptions/status updates) could be applied to many facets of social media outlets. Examples of these outlets are maximizing Twitter/Instagram followers, maximizing Facebook likes, and so forth.

The specific hypothesis we are looking to test is: is there a correlation between tag content and popularity of a video on YouTube? To analyze this, we will be using the [Trending YouTube Video Statistics and Comments](https://www.kaggle.com/datasnaek/youtube) data created by Mitchell J, which can be found on Kaggle. Mitchell j is a student at Glasgow University in the UK, and he collected the data through the YouTube API.

To analyze our hypothesis, we plan to utilize statistical analysis to perform EDA, as well as utilize binomial/logistic regression models to predict trends/analyze our research question. We also plan on using k-nearest neighbor classifiers along with k-folds cross validation to predict whether or not new videos would be considered ‘popular.’

We’re hoping that our audience can learn a lot about this topic, including: whether or not the total amount of tags affect the number of interactions a video will get; whether or not certain tags affect the number of interactions in a video; whether or not certain categories have more trending videos.

**Technical Description**

To present our project and findings, our final web resource will be an html page hosted on Github pages, similar to what we have done on previous assignments. To preface this, we will also be using Jupyter Notebook and Python 3, so exporting an HTML page will be done through the Jupyter interface.

With respect to the data itself, thankfully most of the data is already cleaned and aggregated for us, so there is not much data cleaning required. That being said, we will be spending a lot of time manipulating the data. In order to complete our project, we our group will spend time learning various machine learning models such as RNNs, KNN, and so forth.

For our analysis, we will first analyze our dataset and determine which variables are needed for analysis. Then, for the sake of performance, we will clean our data (if we feel like the data isn’t sufficiently cleaned) and only include columns that is relevant to our analysis. After that, we will perform EDA to gain a better understanding of the breadth and depth of our dataset. During this EDA process, we will explore different relationships between different variables as well as analyze the distributions of the different covariates in our dataset. After the EDA stage, we will look to craft the appropriate statistical models. We will then proceed to use the models to predict the popularity of videos based on tag content. Finally, we will interpret the results and determine if we should accept/reject our null hypothesis, which ultimately answers our research question.

The major challenge we anticipate to run into during this process is how to use machine learning in our specific use case. This mainly stems from our unfamiliarity with how to use utilize different classifiers, and lack of knowledge concerning different models that apply specifically to our dataset. For instance, although it may be a stretch goal, learning about Recurrent Neural Networks would be a big challenge, since we would have to learn the context about when to use it, and how exactly it works.