Data-Miners: Final Report

By Jacob Erickson, YeJee (Jenny) Lee, & Clara Mae Wells

Problem Statement & Goals

For our project we chose to focus on trending YouTube videos and what leads to a successful video. The primary customers that we have in mind for our project are the content creators who upload videos to YouTube with the hope of reaching the largest possible audience and receiving positive user engagement (likes, comments, views). Specifically, our project seeks to serve these content creators by helping them both understand what leads to "successful" YouTube videos, and what they can do to reach a larger audience for their own videos. For the purposes of our project, we focused particular emphasis on the text based features of a YouTube video (such as its title and tags) since a content creator has significant control in setting these for each video. Additionally, we hypothesized that certain words would more likely lead to a higher chance of success for a YouTube video than others.

Further, we considered the needs of content creators from multiple perspectives, and we incorporated these aspects into our project. For example, a content creator might like to have an idea of how many views they will get based on the words that they use in their video title, tags, and channel title. A content creator might also like to know what message the tags that they use are sending to an audience and how these words are associated with different genres. Our project considered these aspects, and several others, each with the goal of giving content creators a detailed view of how they could fit their own videos into the YouTube trending video architecture, and maximize user engagement.

The Dataset & Questions Along the Way

The dataset that we chose to tackle our problem statement consists of 10 separate files each representing one geography. The geographies in our dataset span North America, Asia, and the EU. Each dataset contains the same features, including what we will call "engagement metrics" such as views, likes, dislikes and comment count, as well as descriptive features such as video title, description, tags, and channel title. For the simplifying purposes of our project we focused primarily on three countries, each of which are english-speaking; US, GB, and CA.

This data set enabled us to perform myriad operations utilizing a variety of data science tools and concepts. The following is a brief and non-exhaustive overview of the data mining techniques we deployed in our project:

- Predictive Regressions

- Mutual information regressions
- Association rules
- Sentiment analysis
- Text processing
- Clustering
- Classification

Using these techniques we sought to address our problem statement, but we also aimed to answer several specific questions about trending YouTube videos:

- Can we predict the user engagement metrics (likes, comments, views) of a video based on the words used in the title, tags, and channel title?
- Can we predict the category that a video belongs to based on the words used in the title, description, and channel title?
- What percentage of videos in each country have positive, negative, or neutral sentiments?
- Which words lead to higher user engagement?
- What is the correlation between likes, dislikes, views, and comment count?
- How have the metrics of trending videos changed over time?
- Which words make up the frequent itemsets of trending videos' titles, and what kinds of lift, support, and confidence do we see with them?

Each of these questions poses unique challenges and each are relevant to the field of data science. Each question required us to utilize techniques that we learned over the course of this semester. These questions also make-up a mixture of predictive and descriptive analytics. Practically speaking these questions all address highly relevant aspects of commercial success on the YouTube platform for a content creator. For example, a company that understands which words lead to higher user engagement can use this to reach a wider audience. Knowing how trends are changing over time on YouTube, and understanding country specific idiosyncrasies such as which countries have higher positive sentiment titles may also be actionable for a business leader who wants to target their content to viewers in each country. Ultimately content creators are seeking to expand, grow and reach new audiences. YouTube is a widely adopted platform, and understanding more about successful YouTube videos is a useful endeavor for any business or content creator.

Over the following pages we will detail how we addressed these questions and our problem statement in our final product, and we will share some of the interesting insights that we learned along the way. We will also provide additional detail on the design and benefits of our service. We will conclude with several ‘next steps’ which could be applied to this dataset in future work.

Correlation Analysis

One of our first steps was to determine the correlation between numeric user engagement features. For example, what is the correlation between the number of views that a video gets and the number of likes or dislikes that it gets? While this doesn’t tell us that one variable causes the other, it does give us some insight into the value of certain user engagement metrics. Please see below for the correlation results.

	category_id	views	likes	dislikes	comment_count
category_id	1.000000	-0.176825	-0.191450	-0.033775	-0.074698
views	-0.176825	1.000000	0.791670	0.405290	0.485986
likes	-0.191450	0.791670	1.000000	0.448010	0.763192
dislikes	-0.033775	0.405290	0.448010	1.000000	0.745064
comment_count	-0.074698	0.485986	0.763192	0.745064	1.000000

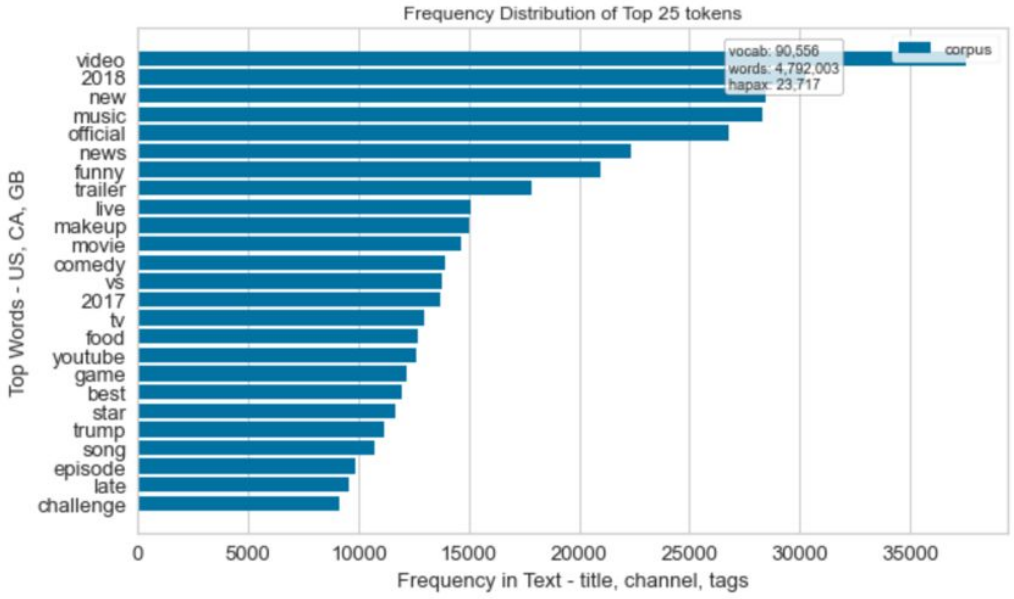
The above table has some interesting takeaways. For starters, one can see that the correlation between the engagement metrics are generally quite high. The correlation between likes and views is 0.79, the correlation between comment_count and likes is 0.76, etc. This even applies to “negative” metrics such as the number of dislikes. While views have a lower correlation with dislikes, it still stands at a relatively strong positive 0.41 correlation. This tells us that in some sense all of the engagement metrics are interrelated. Interestingly, we can also see that there is a slightly negative correlation between views and likes, and the category id. While we wouldn’t really expect a meaningful correlation between category id and views, this does highlight the fact that engagement metrics vary by category.

Text processing

A substantial portion of our project including sentiment analysis, mutual information regression, predictive regression (predicting various engagement metrics), and classification (predicting the

category that a video belongs to) all involved text processing. This is primarily because several of our features were text features as opposed to numeric features. To prepare our dataset for many of the operations we intended to do in our project we had to convert the text features into numeric ones. Our initial thought was to use dummy encoding for each text feature (titles, tags, descriptions, etc). However, given the number of instances we have in each geographical dataframe coupled with the size of the vocabulary in each text feature it quickly became clear that this would result in a massive dataset with a very large amount of columns. Instead of dummy encoding we instead chose to follow the bag-of-words model and we used a count vectorizer from the scikit-learn library for its implementation. In most cases we used a standard count vectorizer with a continuous count of word appearances in each text column, but we did use a binary count vectorizer for association rules. We then used english stop words in each case.

See below for a visual of the kinds of text tokens found in our dataset. Specifically, the below graphic shows the top 25 tokens across the video title, channel title, and tags for all three geographies that we focused on.



Sentiment Analysis

With sentiment analysis, our goal was to see how the sentiment varied across geographies for various descriptive features (such as title). To do this we used the Natural Language Toolkit (NLTK) package in Python, and specifically we used the sentiment.vader module. Using this package we set a threshold, and then computed a compound sentiment score. Using this

compound score we calculated the percentage of positive, negative, or neutral instances in a given column. The results of this analysis revealed a couple of interesting observations. Below are dataframes representing the percentage of titles, tags, and descriptions instances which are positive, negative or neutral based on their compound score for each geography.

```
print(title_df)
```

	Positive	Negative	Neutral
US	19.7%	13.4%	66.9%
GB	18.0%	14.0%	68.0%
CA	16.4%	15.9%	67.7%

```
print(tags_df)
```

	Positive	Negative	Neutral
US	16.3%	8.3%	75.4%
GB	14.6%	9.7%	75.7%
CA	13.5%	9.1%	77.4%

```
print(description_df)
```

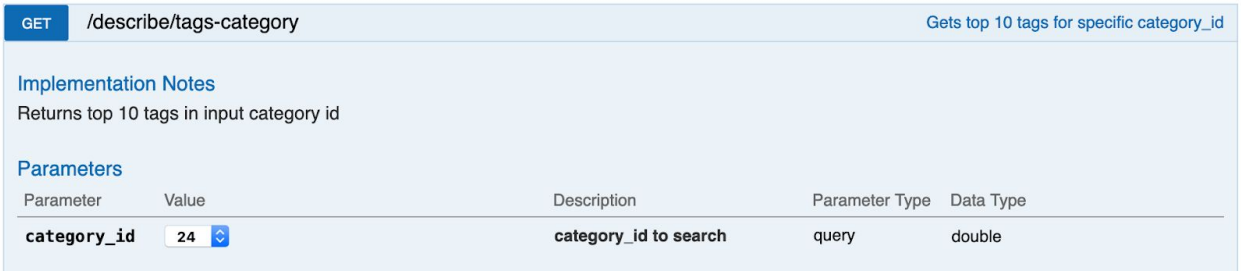
	Positive	Negative	Neutral
US	67.6%	11.3%	21.1%
GB	56.5%	13.1%	30.4%
CA	57.8%	12.6%	29.6%

Notably the tags and titles are overwhelmingly neutral, and the percentages are roughly the same across geographies. However, it flips with the description where the sentiment is overwhelmingly positive, with the US having a significantly higher percentage of positive descriptions than the other geographies. Additionally, the percentages didn't change drastically as we iteratively set the compound score at different levels for determining positivity, negativity, and neutrality of text instances.

Mutual information regressions

As part of the descriptive aspect of our project, we wanted to understand what the top features were when it came to contributing to user engagement metrics. For starters we computed the mutual information of all numeric features, similar to how we handled correlation above. This was mildly interesting but not especially useful. The meat of our mutual information regressions instead came from computing the mutual information between text features and user engagement metrics. For the purposes of our project we completed two different mutual information regression calculations, one solely to learn 'interesting insight' about the data itself, and another for our user-facing API. For the purposes of our API we computed the mutual information amount for tags on the number of views in each category, and we are returning the top 10 tags back to our users. These calculations are relatively quick to perform, and they

provide users with insight into which tags are particularly valuable in each category. In the API, our mutual information regressions look as follows:



Our other mutual information regression computed the value of all words in the title on the number of likes for an entire geography. This analysis took on the order of several hours to run, and thus it was unreasonable to include in our API. Additionally, we were unable to get the country-wide mutual information regression to run on the Canada dataset even after several hours of leaving the program to run and multiple attempts.

Nonetheless, this exploratory aspect of our project revealed some interesting insights about which words contributed the most to likes in each geography. Below is a rounded sampling of what we saw in the US and Great Britain. Notably the top values are quite a bit higher in GB than in the US. It is also notable that while both GB and the US share many top words, there are also substantial differences. This calculation provides some useful context for what words each country finds most engaging.

US - Top 12 Words (Title & Likes)	GB - Top 12 Words (Title & Likes)
1. official - 0.0255	1. video - 0.0500
2. video - 0.0253	2. official - 0.0375
3. bts - 0.0130	3. ft - 0.0318
4. news - 0.0120	4. trump - 0.0193
5. 방탄소년단 (bts) – 0.0098	5. mv - 0.0173
6. 2018 – 0.0097	6. vs - 0.0154
7. highlights - 0.0095	7. new - 0.0151
8. game - 0.0093	8. 2017 - 0.0143
9. mv – 0.0092	9. star - 0.0142
10. dude - 0.0088	10. wars - 0.0134
11. makeup - 0.0081	11. 2018 - 0.0134
12. talk - 0.0079	12. oficial - 0.0130

Predicting User Engagement Metrics (Regressions)

For this portion of our project, we focused on predicting user engagement metrics, including the number of likes, comments, and views based on the words in a video's title, tags and channel

title. We were able to do this by using a count vectorizer on a concatenated column containing the video title, tags, and channel title. Our predictive models take into account the words (and their counts) within this concatenated column when predicting each user engagement metric. We decided to train our models with an 80/20 train/test split, and we created a custom function for doing this split on our dataframe so that we could avoid code replication. We evaluated several different models for our predictive purposes, including linear regression, plus the GradientBoostingRegressor and the BaggingRegressor from sklearn. We ultimately settled on using the linear regression, based on several factors which we will discuss here.

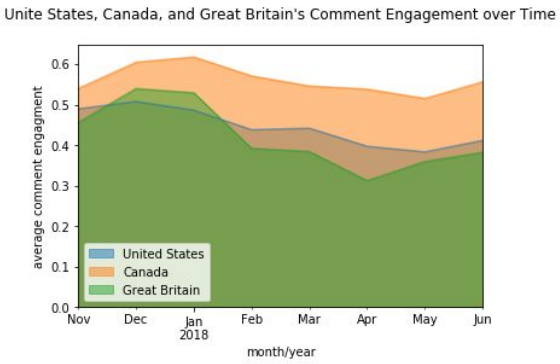
The linear model performed very well on the US and GB dataframes based on the r-squared value we calculated with the score function of our linear_model.Linear Regression object. R-squared values ranged from over 0.75 for views, to over 0.90 for likes and comment count. While these predictions are unlikely to ever be exact, the goal is for content creators to have a general idea of how the words they choose to associate with their video affect the likely view, comment, and likes count. A content creator can test out different words within our API to get different predicted values and using this they have a metric by which to assess performance. The biggest limitation with the linear regression was that it delivered a negative r-squared value for Canada no matter which user engagement metric we attempted to predict. Accordingly this led us to reconsider the linear model, and led to the testing of other models.

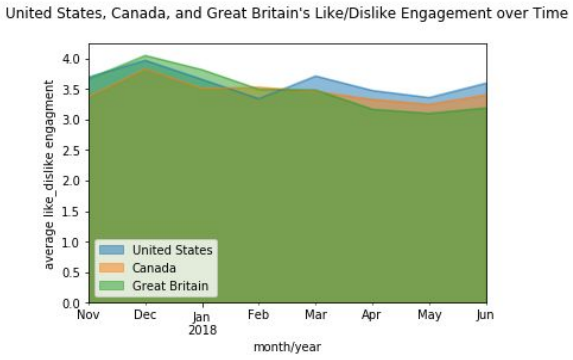
Next, we tried the GradientBoostingRegressor. The GradientBoostingRegressor had a positive r-squared value for Canada, but it performed considerably worse than the linear model on the US and GB. We then tried the BaggingRegressor. This model delivered a far better r-squared value for Canada, and it performed slightly better for the US and GB than the linear model. This initially led us to believe that the Bagging model would be our best choice. However, the Bagging model (and the Gradient model) suffered from a damaging issue; it delivered unreasonably low prediction values. Given that our dataset represents trending videos, these predictions simply didn't make sense when we evaluated them. As one illustrative example, when we predicted the comment count for a US BTS music video (using the text in the video) the linear model predicted 124k comments, while the bagging model predicted only 20k. The actual value in our dataset was 72k. Curiously this general trend of the linear model performing better on "spot-check" examples also held for Canada.

As such, we decided to use the linear model despite its strange performance on the Canada dataset. The results for the US and GB are in this respect more reliable (given the high r-squared values) but we also decided to keep the prediction tool in place for Canada in our API.

Engagement

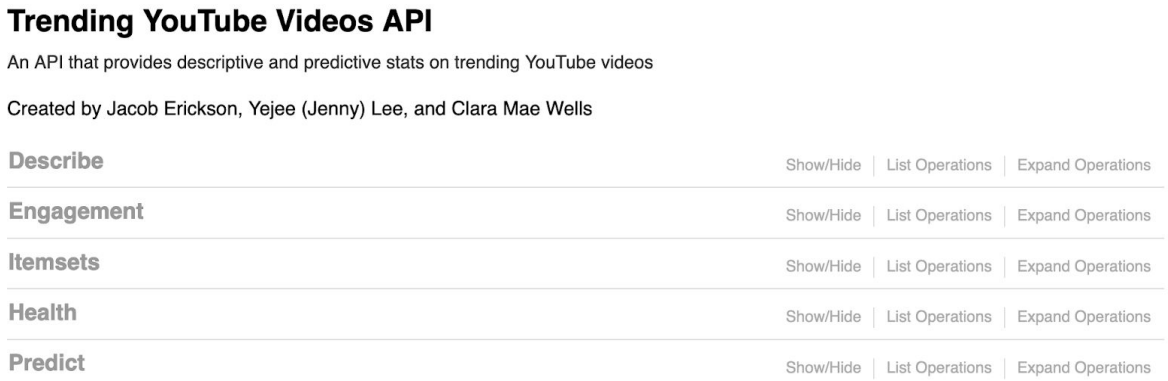
Because the original data source noted that a variety of factors, including user interactions, contribute to determining the trending videos, we added two features to the dataset: comment_engagement % and like/dislike_engagement %. Likes and dislikes are combined because the purpose is to see whether a button was pressed, regardless of the sentiment. Comment engagement is defined as comment count divided by number of views, and like/dislike engagement is defined as the total number of likes and dislikes divided by the number of views. We anticipated a large percentage of engagement for trending videos. However, for all three English-speaking countries, comment engagement remains below 1% for the dataset timeframe; like/dislike engagement is a little higher and remains below 5%.





API & User Interface

Given all of the analysis detailed above, it's worth spending a minute on our API and how users interact with it. Our API consists of many functions and files, but a user can interact with it all through the youtube_api.py file. Running this file, and then navigating to the http://0.0.0.0:8080/ui address leads to the following interface:



When a user clicks on any of the drop-down menus they are immediately presented with a variety of different options. As illustrated above, our API layers-in many facets from our project so that users receive a full suite of insights into trending videos. Under the 'Predict' header a user can interact with all of our predictive features. Available options allow a user to predict the category of a video and get suggested tags, as well as to predict the user engagement metrics for a video, and the sentiment of a single video from our dataset (using the video id). The other headers generally lead to descriptive statistics, such as the sentiment by country, the count by category, and a display of engagement metrics.

Our UI design was meant to give users all the tools that we could provide in one single place for convenient access. While we initially had the functions segmented into separate files, we decided that having one hub for users to navigate from would be the optimal user experience. It conveniently displays all of the options available to the user.

Service Benefits & Design

We wanted our service to offer the predictive and descriptive statistics that answer our problem statement and align with the benefits we want to offer content creators, such as enhancing their understanding of viewers and presenting data-informed content ideas. We expect that these offerings will increase content creators' audience reach and the chance that their videos will trend. This is beneficial because creators want to share and spread their messages, and more views and engagement lead to a greater chance of acceptance to the YouTube Partner Program.

Design-wise our predictive statistics take as input the video title, channel title, and video tags. From those we predict the video's view, like, or comment count. Our descriptive statistics present insights as well as hints/guides that could improve view counts. Insights include the number of trending videos in each category, comment or like/dislike engagement for all videos, and comment or like/dislike engagement by category. These insights show which categories are more popular and how users are interacting with each category in the target country. Hints include frequent itemsets (one, two, and three itemsets) and association rules, both on the title column. These itemsets and rules show creators which title words appear together frequently in trending videos, so creators can name their similar video accordingly.

Next Steps

There are several future deliverables for this project. Arguably the biggest weakness of our dataset is that it only includes videos that are trending. Accordingly, the primary next step would be expanding the dataset to include videos that are not trending; this would give us a more filled-in dataset which could be used to draw more complete conclusions. Currently we can only predict views for videos based on the videos and words in our dataset. It's possible that, for instance, trending YouTube videos have certain features and aspects in common which are not shared with YouTube videos that haven't been trending. This would be an interesting area to layer with our project to create a more complete view of YouTube videos.

Another logical next step would be to utilize a system (either hardware or software) for handling massive datasets. One of our limitations in this project was that certain operations took an excessive amount of memory to run, and thus we couldn't fully implement them as desired. This constraint affected several areas of our project, most noticeably frequent itemset generation and mutual information regressions. In terms of frequent itemsets, the right system would allow for a function that takes country and itemset size as input, so that memory capacity does not prevent us from generating and returning frequent itemsets of any desired country in the dataset rather than only one. Itemset hints could be taken further if they were based on a time period (possibly month and year parameters) since trending topics - and therefore frequent itemsets - will change over time, particularly with a dataset that covers a larger range of time than ours. Similarly the mutual information calculation took a very long time to run for the US and GB, and it wouldn't run to completion for Canada. Addressing these shortcomings would allow for a valuable expansion of our project.

A final next step would be to dive-deeper to learn more about why the linear regression model for Canada delivered a negative r-squared despite appearing to deliver "good" predictions when spot-checked against our actual data.

Conclusion

In conclusion, the goal of our project was to create a product for content creators which would help them increase user engagement for their videos on the YouTube platform. Secondarily, we wanted to understand more about what leads to a "successful" video and to pass this same knowledge on to our users (content creators). As highlighted above there are several next steps that could further increase the utility of our project for users. Nonetheless we hope that this project helped to answer some of the questions that we raised and that our service provides users with a valuable product.