Word-Processing with the YouTube Algorithm

Introduction/Goal:

The purpose of this project is to utilize R and run a word-processing program to simulate how YouTube runs its video recommending process as well as identify which keywords may be associated with certain genres/categories of videos to better display titles for videos to both new and old content creators to gather more views and expand their channel. By utilizing certain keywords within the title of your videos and based on keywords that are used in videos you watch, you should receive recommendations with similar words/categories. This process doesn’t seem to have been taken on by anyone and is uniquely focusing on text/word processing whereas there has been an instance which utilized running YouTubes API algorithm with videos instead of text.

Review:

Lovejoy uses this page to describe his process of experimenting with creating his own algorithm to determine which videos are valuable to watch and be suggested. He uses code within python to execute the metrics he collected via the YouTube API. Following this, he then uses metrics such as views per video, subscribers, comment count, etc. to filter out the top-quality videos based on searching with a specific word or set of words. He finds his algorithm is correctly identifying videos that he would enjoy and begins to run it through different mediums such as AWS Lambda. Potential next steps for Lovejoy include text/term searching which would lower the dependency on how videos will be picked up in the search. According to Lovejoy, the code is slow so having something optimized better would create a more efficient experience. Based on the results he has, there are no observable visualizations that would help to filter out results of terms or keywords and the project falls short in observing this.

In conclusion, there hasn’t been much done in trying to observe the best way to filter out the most popular/best quality videos in a similar way to how YouTube runs its algorithm and with being able to visualize the results based on certain keywords and terms for specific categories and genres. Based on research, we would be trying a new approach to an interesting problem which has been touched by few.

Dataset:

One of the toughest challenges in prepping for this project was finding usable data to help us in understanding how we can try and simulate the YouTube algorithm. Of course, with this idea we had to make sure we had videos, lots of videos. The dataset we ended up utilizing actually has a very good set of variables which would assist us in processing the text used in the title of the video to associate it with a particular category type, for example, we have Automotive, Food/Culinary, Gaming, and Film & Movies. In total we have around 40+ categories of videos which will be separated and then used to create separate datasets where only videos with a specific category are shown and almost 41000 videos which we will be sorting through.

Within the dataset, we notice that there are multiple columns listed, for example we have the number of views, comments, likes, and dislikes. In addition, we also have the title and the tags as well, which we can use to help breakdown the text processing for the YouTube algorithm. Our focus will be the YouTube video title column, which is the primary variable of interest and secondly, will be the video category. Using these 2 variables, we should be able to create some interesting visualizations which show the common word usage for all the different categories of videos. Some other visuals we can observe is the breakdown of video likes/views based on the video category type. This will allow us to gain an understanding as to which category seems to have the most presence when surfing YouTube.

However, even with all this data, we do have many variables which are not going to provide any function to our problem, such as the trending date of the video, comments disabled, and the posted time of the video. These are just a few of the variables within the dataset which we will be removing to better work with what we have planned.

Methods:

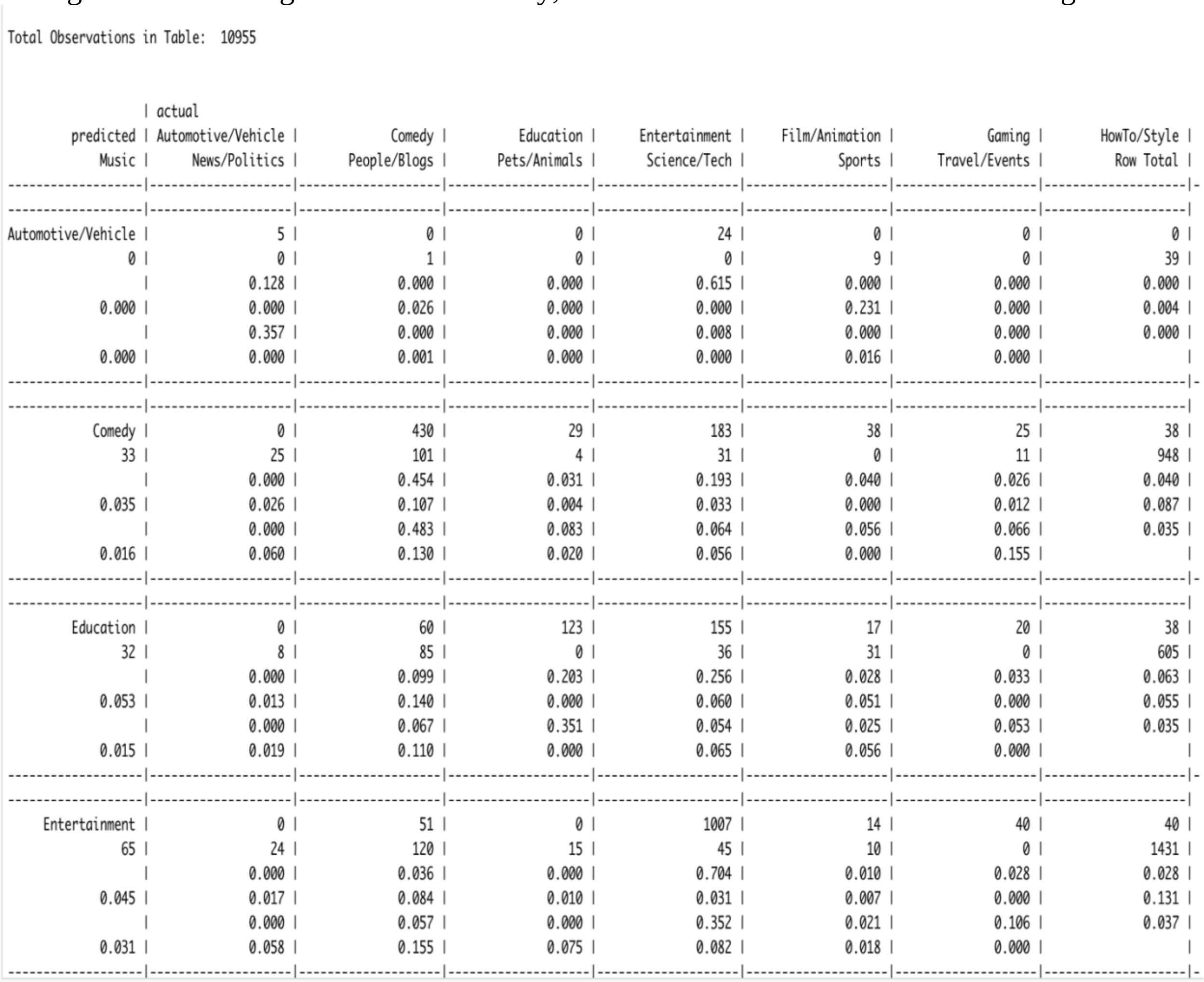
The methods that we plan on implementing for this project include running a Machine Learning text processing algorithm which will pick up the most common nouns, adjectives, and verbs used per category where it will be created into a visualization that displays this information. The algorithm will clean out stop words, and place holder words so that these words won't appear as a commonly used word since they're used in every other sentence/title. In addition, we will be visualizing the count for each word per category as well as seeing which category has stayed the most popular over the timespan of the dataset.

Pros/Cons:

When working with these methods, we have a handful of pros which we can elaborate on. For example, as mentioned above, we have the algorithm removing basic filler words such as "and", "the", "to", "they", etc. This makes the algorithm process the text faster and more efficiently for us to use. We can also utilize the visualization to give us better insight into how the data is distributed and understanding the text used per category of videos to get an idea of how we can expect the titles to be used if we were to create our own YouTube video for a specific category. Giving us an open look into how each category performs will ultimately lead us to understanding how this YouTube algorithm works when we use the platform.

Method Results:

Based on running the algorithm we did get some decent results when it comes understanding how the YouTube algorithm works. First off, we were able to create multiple visualizations that showcase the word usage breakdown for the categories of videos. By having this, it gives us insight into comparing the different categories with each other based on the usage of specific nouns, adjectives, verbs, etc. Once we are finished comparing the categories, we create a new section to run the ML algorithm through Naïve Bayes. How we go about utilizing this is by creating a new dataset where we have the title of the videos and the respective category of the video as our only variables and begin to create our training and testing Datasets. For our test run, we split the data into a 75/25 distribution where 75% was the training data and the remaining 25% was the test data. After having the algorithm run its course on the training and testing, we go ahead and create the classifier and predictor to determine how accurate our training and testing data is. Utilizing the GModels library, we create a Cross Tab with the following results:

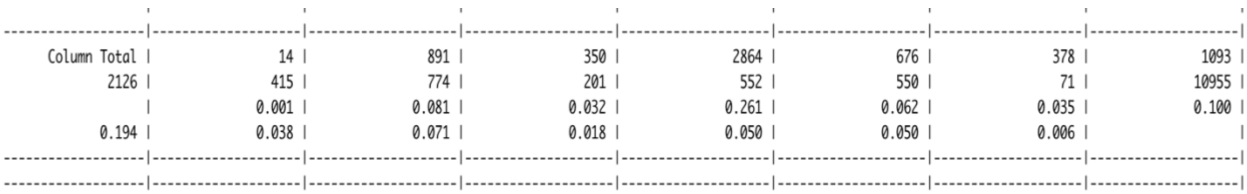


```
Total Observations in Table:  10955


                   | actual 
         predicted | Automotive/Vehicle |             Comedy |          Education |      Entertainment |     Film/Animation |             Gaming |        HowTo/Style |
             Music |      News/Politics |       People/Blogs |       Pets/Animals |       Science/Tech |             Sports |      Travel/Events |          Row Total |
-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
Automotive/Vehicle |                  5 |                  0 |                  0 |                 24 |                  0 |                  0 |                  0 |
                 0 |                  0 |                  1 |                  0 |                  0 |                  9 |                  0 |                 39 |
                   |              0.128 |              0.000 |              0.000 |              0.615 |              0.000 |              0.000 |              0.000 |
             0.000 |              0.000 |              0.026 |              0.000 |              0.000 |              0.231 |              0.000 |              0.004 |
                   |              0.357 |              0.000 |              0.000 |              0.008 |              0.000 |              0.000 |              0.000 |
             0.000 |              0.000 |              0.001 |              0.000 |              0.000 |              0.016 |              0.000 |                    |
-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
            Comedy |                  0 |                430 |                 29 |                183 |                 38 |                 25 |                 38 |
                33 |                 25 |                101 |                  4 |                 31 |                  0 |                 11 |                948 |
                   |              0.000 |              0.454 |              0.031 |              0.193 |              0.040 |              0.026 |              0.040 |
             0.035 |              0.026 |              0.107 |              0.004 |              0.033 |              0.000 |              0.012 |              0.087 |
                   |              0.000 |              0.483 |              0.083 |              0.064 |              0.056 |              0.066 |              0.035 |
             0.016 |              0.060 |              0.130 |              0.020 |              0.056 |              0.000 |              0.155 |                    |
-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
         Education |                  0 |                 60 |                123 |                155 |                 17 |                 20 |                 38 |
                32 |                  8 |                 85 |                  0 |                 36 |                 31 |                  0 |                605 |
                   |              0.000 |              0.099 |              0.203 |              0.256 |              0.028 |              0.033 |              0.063 |
             0.053 |              0.013 |              0.140 |              0.000 |              0.060 |              0.051 |              0.000 |              0.055 |
                   |              0.000 |              0.067 |              0.351 |              0.054 |              0.025 |              0.053 |              0.035 |
             0.015 |              0.019 |              0.110 |              0.000 |              0.065 |              0.056 |              0.000 |                    |
-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
     Entertainment |                  0 |                 51 |                  0 |               1007 |                 14 |                 40 |                 40 |
                65 |                 24 |                120 |                 15 |                 45 |                 10 |                  0 |               1431 |
                   |              0.000 |              0.036 |              0.000 |              0.704 |              0.010 |              0.028 |              0.028 |
             0.045 |              0.017 |              0.084 |              0.010 |              0.031 |              0.007 |              0.000 |              0.131 |
                   |              0.000 |              0.057 |              0.000 |              0.352 |              0.021 |              0.106 |              0.037 |
             0.031 |              0.058 |              0.155 |              0.075 |              0.082 |              0.018 |              0.000 |                    |
-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
```

```
                  .                   .                   .                   .                   .                   .                   .                   .
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
   Film/Animation |                 0 |                49 |                24 |               351 |               437 |                93 |                 0 |
               26 |                 3 |                12 |                 0 |                 0 |                 4 |                 0 |               999 |
                  |             0.000 |             0.049 |             0.024 |             0.351 |             0.437 |             0.093 |             0.000 |
            0.026 |             0.003 |             0.012 |             0.000 |             0.000 |             0.004 |             0.000 |             0.091 |
                  |             0.000 |             0.055 |             0.069 |             0.123 |             0.646 |             0.246 |             0.000 |
            0.012 |             0.007 |             0.016 |             0.000 |             0.000 |             0.007 |             0.000 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
           Gaming |                 0 |                 1 |                 0 |                23 |                 0 |               143 |                 0 |
                0 |                 0 |                 8 |                 6 |                 0 |                11 |                 0 |               192 |
                  |             0.000 |             0.005 |             0.000 |             0.120 |             0.000 |             0.745 |             0.000 |
            0.000 |             0.000 |             0.042 |             0.031 |             0.000 |             0.057 |             0.000 |             0.018 |
                  |             0.000 |             0.001 |             0.000 |             0.008 |             0.000 |             0.378 |             0.000 |
            0.000 |             0.000 |             0.010 |             0.030 |             0.000 |             0.020 |             0.000 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
      HowTo/Style |                 0 |                53 |                26 |               229 |                 7 |                 0 |               716 |
               49 |                29 |                67 |                 0 |                 5 |                 0 |                 0 |              1181 |
                  |             0.000 |             0.045 |             0.022 |             0.194 |             0.006 |             0.000 |             0.606 |
            0.041 |             0.025 |             0.057 |             0.000 |             0.004 |             0.000 |             0.000 |             0.108 |
                  |             0.000 |             0.059 |             0.074 |             0.080 |             0.010 |             0.000 |             0.655 |
            0.023 |             0.070 |             0.087 |             0.000 |             0.009 |             0.000 |             0.000 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
            Music |                 0 |                44 |                 0 |               147 |                67 |                 6 |                22 |
             1713 |                54 |                44 |                18 |                 8 |                54 |                 0 |              2177 |
                  |             0.000 |             0.020 |             0.000 |             0.068 |             0.031 |             0.003 |             0.010 |
            0.787 |             0.025 |             0.020 |             0.008 |             0.004 |             0.025 |             0.000 |             0.199 |
                  |             0.000 |             0.049 |             0.000 |             0.051 |             0.099 |             0.016 |             0.020 |
            0.806 |             0.130 |             0.057 |             0.090 |             0.014 |             0.098 |             0.000 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
    News/Politics |                 0 |                15 |                 0 |                81 |                 0 |                 0 |                 0 |
               13 |               184 |                 3 |                 5 |                 0 |                 0 |                 0 |               301 |
                  |             0.000 |             0.050 |             0.000 |             0.269 |             0.000 |             0.000 |             0.000 |
            0.043 |             0.611 |             0.010 |             0.017 |             0.000 |             0.000 |             0.000 |             0.027 |
                  |             0.000 |             0.017 |             0.000 |             0.028 |             0.000 |             0.000 |             0.000 |
            0.006 |             0.443 |             0.004 |             0.025 |             0.000 |             0.000 |             0.000 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-


------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
     People/Blogs |                 0 |                74 |                 3 |               376 |                28 |                46 |               136 |
               97 |                42 |               243 |                 8 |                55 |                 0 |                 0 |              1108 |
                  |             0.000 |             0.067 |             0.003 |             0.339 |             0.025 |             0.042 |             0.123 |
            0.088 |             0.038 |             0.219 |             0.007 |             0.050 |             0.000 |             0.000 |             0.101 |
                  |             0.000 |             0.083 |             0.009 |             0.131 |             0.041 |             0.122 |             0.124 |
            0.046 |             0.101 |             0.314 |             0.040 |             0.100 |             0.000 |             0.000 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
     Pets/Animals |                 0 |                 0 |                13 |                33 |                28 |                 0 |                27 |
                0 |                10 |                 0 |                84 |                11 |                14 |                14 |               234 |
                  |             0.000 |             0.000 |             0.056 |             0.141 |             0.120 |             0.000 |             0.115 |
            0.000 |             0.043 |             0.000 |             0.359 |             0.047 |             0.060 |             0.060 |             0.021 |
                  |             0.000 |             0.000 |             0.037 |             0.012 |             0.041 |             0.000 |             0.025 |
            0.000 |             0.024 |             0.000 |             0.418 |             0.020 |             0.025 |             0.197 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
     Science/Tech |                 9 |                76 |               123 |               201 |                40 |                 5 |                76 |
               81 |                36 |                79 |                61 |               352 |                13 |                 0 |              1152 |
                  |             0.008 |             0.066 |             0.107 |             0.174 |             0.035 |             0.004 |             0.066 |
            0.070 |             0.031 |             0.069 |             0.053 |             0.306 |             0.011 |             0.000 |             0.105 |
                  |             0.643 |             0.085 |             0.351 |             0.070 |             0.059 |             0.013 |             0.070 |
            0.038 |             0.087 |             0.102 |             0.303 |             0.638 |             0.024 |             0.000 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
           Sports |                 0 |                38 |                 9 |                39 |                 0 |                 0 |                 0 |
               17 |                 0 |                11 |                 0 |                 9 |               404 |                 0 |               527 |
                  |             0.000 |             0.072 |             0.017 |             0.074 |             0.000 |             0.000 |             0.000 |
            0.032 |             0.000 |             0.021 |             0.000 |             0.017 |             0.767 |             0.000 |             0.048 |
                  |             0.000 |             0.043 |             0.026 |             0.014 |             0.000 |             0.000 |             0.000 |
            0.008 |             0.000 |             0.014 |             0.000 |             0.016 |             0.735 |             0.000 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
    Travel/Events |                 0 |                 0 |                 0 |                15 |                 0 |                 0 |                 0 |
                0 |                 0 |                 0 |                 0 |                 0 |                 0 |                46 |                61 |
                  |             0.000 |             0.000 |             0.000 |             0.246 |             0.000 |             0.000 |             0.000 |
            0.000 |             0.000 |             0.000 |             0.000 |             0.000 |             0.000 |             0.754 |             0.006 |
                  |             0.000 |             0.000 |             0.000 |             0.005 |             0.000 |             0.000 |             0.000 |
            0.000 |             0.000 |             0.000 |             0.000 |             0.000 |             0.000 |             0.648 |                   |
------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-
```

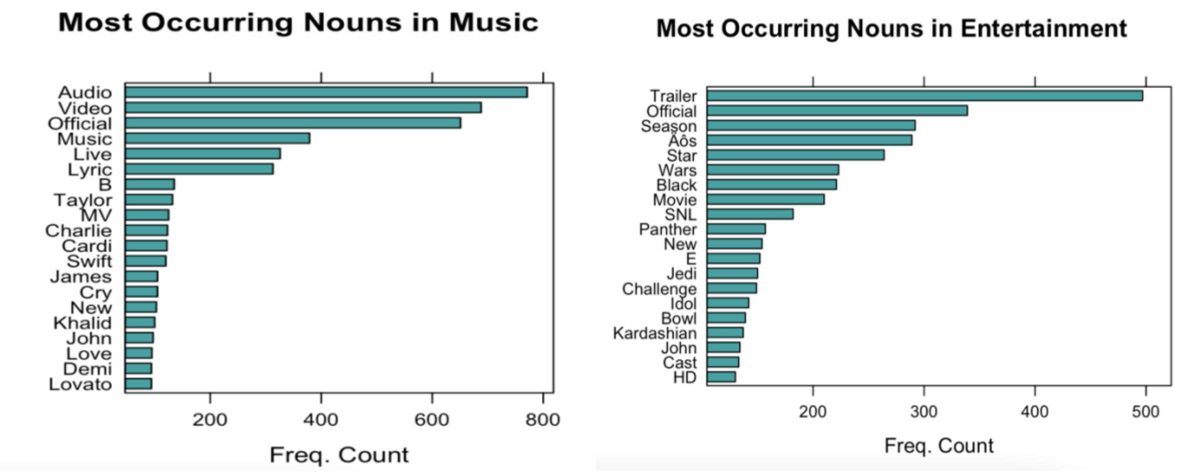


```
                  |                |                |                |                |                |                |                |
------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
     Column Total |             14 |            891 |            350 |           2864 |            676 |            378 |           1093 |
             2126 |            415 |            774 |            201 |            552 |            550 |             71 |          10955 |
                  |          0.001 |          0.081 |          0.032 |          0.261 |          0.062 |          0.035 |          0.100 |
            0.194 |          0.038 |          0.071 |          0.018 |          0.050 |          0.050 |          0.006 |                |
------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
```

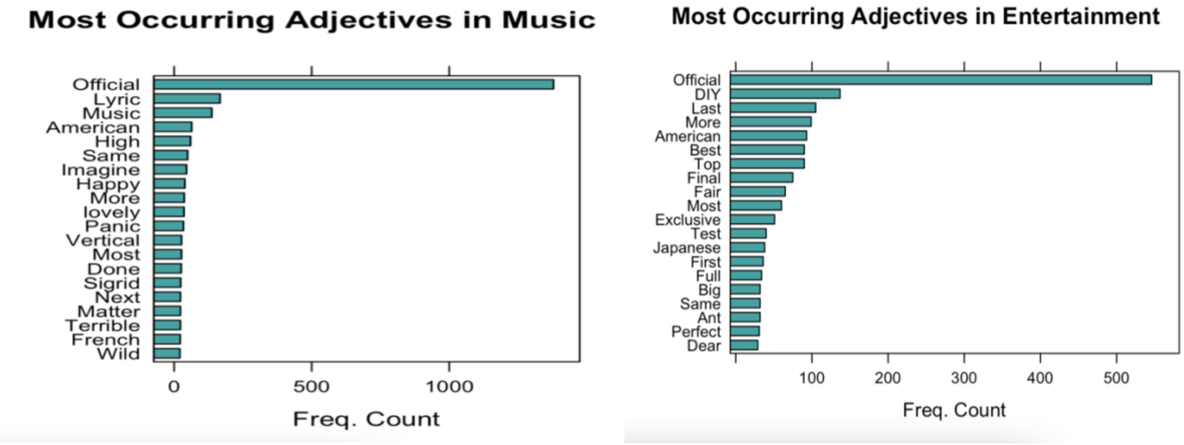
Here we can get an idea how accurate our test results were when running the NaiveBayes and the results are not too far off from being as expected. For example, we can take a look at the Sports category, and we see that about 404 cases were predicted correctly out of the 527 cases that we had for this particular category which is approximately 77% accurate. Another example we see that has a high predict value is within the Music Category, 1713 cases were identified correctly out of 2177, but we also see that about 147 videos (7%) were incorrectly identified as part of the Entertainment category. Seeing this shows us in a few ways how the YouTube Algorithm actually works because not all videos in a viewers feed would be seen as categorized under a single category like Music. However, most of the recommended videos would be part of the primary category whereas the rest would be somewhat related to the primary category, which in this case, Music could also be classified as Entertainment.

Visualizations:

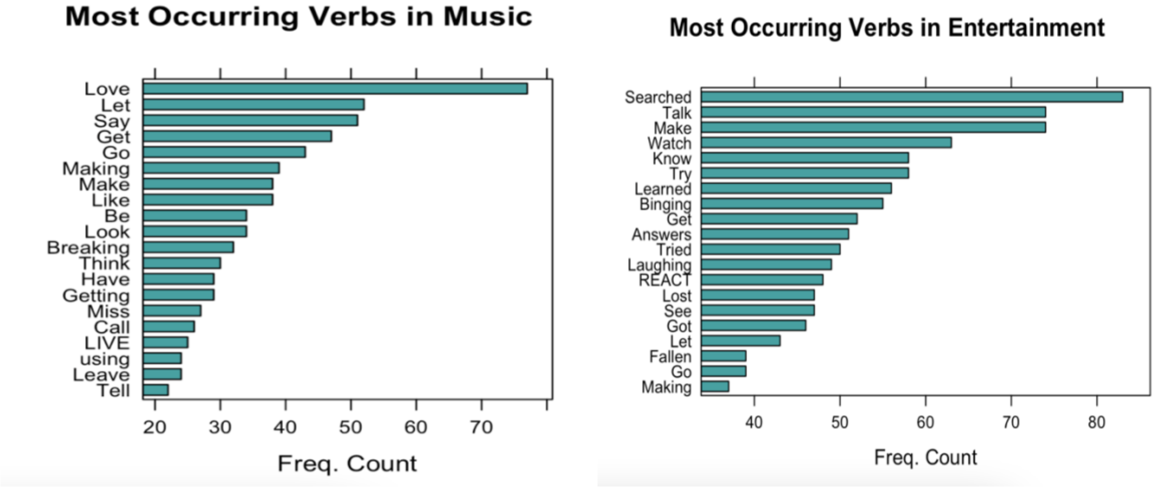
Referring back to our results, we also have to consider that some Categories may have a relationship with each other. One way is to take note of the keywords used in the nouns, verbs, adjectives, etc. Let’s take a look:



Here we see the most occurring nouns for both Music and Entertainment.



Here we see the most occurring Adjectives for both Music and Entertainment.



Here we see the most occurring verbs for both Music and Entertainment.

Based on this, we see that there are a few words that are shared between the two categories, one of the most common repeated words is 'Official' which can also explain why the cross tab predicted some Entertainment category videos over the correct Music category.

<u>Future Changes/Improvements:</u>

As we went ahead and worked with titles for each of the videos that were deemed popular for each of the different YouTube categories, we got an idea of how we can utilize certain key words to get an understanding of how YouTube's algorithm works. By using these key words in our titles for videos, it would allow the algorithm to catch on to our videos to expose them at a much broader scale. The second part to add to this project would be to utilize the tags in the video description which would be similar to how we ran the algorithm for the video titles. By combining the best common keywords in our tags alongside our YouTube title, our videos would have a much higher rate of exposure within YouTube’s algorithm to give us better views and interactions.

Some other visuals that we can add would include the interactions for each different video category and create a chart to show which of the categories we worked with was the most popular so that we can utilize this information to try and create videos related to the most popular categories for maximum exposure in creating YouTube videos/channels.

References:

Lovejoy, Chris. "I Created My Own YouTube Algorithm (to Stop Me Wasting Time)." *Chris Lovejoy*, Chris Lovejoy, 12 Nov. 2021, https://chrislovejoy.me/youtube-algorithm/.

J, Mitchell. "Trending YouTube Video Statistics." *Kaggle,* 3 June 2019, https://www.kaggle.com/datasets/datasnaek/youtube-new?resource=download&select=USvideos.csv.