BA820: Final Project Paper

Team B2: Phyllis Cao, Zihan Cui, Aash Gohil, Chunxiaqiu(Tommy) Yang, Zixing Li

1. Business Problem

Background/Overview

YouTube is an online video-sharing platform owned by Google, used by multimillion-dollar corporations and individual content creators. Its content is accessible from various platforms such as computers, phones, gaming consoles, and smart TVs. It is one of the stickiest digital products with a stickiness factor of 0.13 with 315.12 million daily active users and 2.3 billion monthly active users and generated $19.7 billion in revenue in 2020.

Project Description

With its undoubtedly significant influence on today's pop culture, YouTube would like to commemorate some of the most impactful videos on its platform. The company is looking to create a series of playlists that capture the daily trending videos to achieve the goal through the past few months. Each playlist should consist of videos that are similar in nature to cater to users with various interests. This project will determine the number of playlists YouTube will need to create and generate relevant information for each playlist to help guide users to choose their playlist of interest. The motivation for this project stemmed from analyzing digital image products like apple photos, which create annual album collages for its users based on their photo library and ML algorithms. In the future, YouTube can also create personalized playlists based on each person's video history, and this project can serve as a pilot for individual playlist customization.

2. Dataset and data clean

https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv

We downloaded the Trending YouTube Video Statistics dataset from Kaggle, and the data was originally extracted from the YouTube library. This dataset includes several months of data on daily trending YouTube videos in 2017 and 2018. While there is data on several countries, this project will focus on the data outlining trending videos in the US.

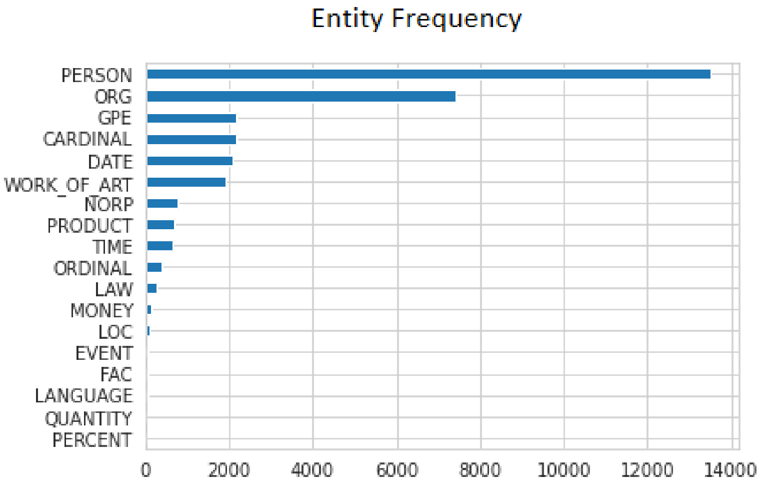
The dataset consists of 40,949 records of YouTube videos (6,351 unique values) and 16 features with a mixture of data types such as string (text), categorical, and numeric variables. Since a video could be trending over multiple days, there are numerous entries for each video id. The data was condensed and reduced from the original number of rows to just the unique video ids, with the last day of trending being considered as the unique value, and an additional feature was introduced, ‘number of days trending’ to capture this information.

Relevant attributes (columns) to focus on in our analysis are:

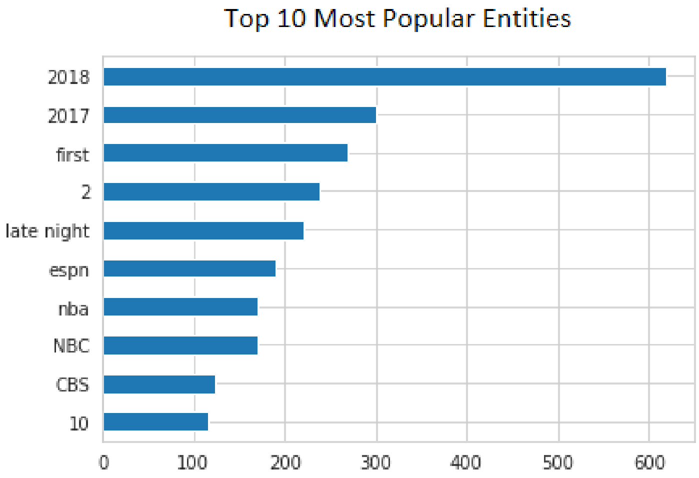
- Video_id – unique ID for each video
- Tags – tags for videos (characteristic)
- Views - number of views received
- Description – short description for the videos
- Likes - number of likes
- Dislikes - number of dislikes
- Comment_count – number of comments
- Comment_disabled - boolean for comment is disabled or not
- Rating_disabled – boolean for rating is disabled or not
- Num_days_trending - Number of days the video was trending for

3. Exploratory Data Analysis

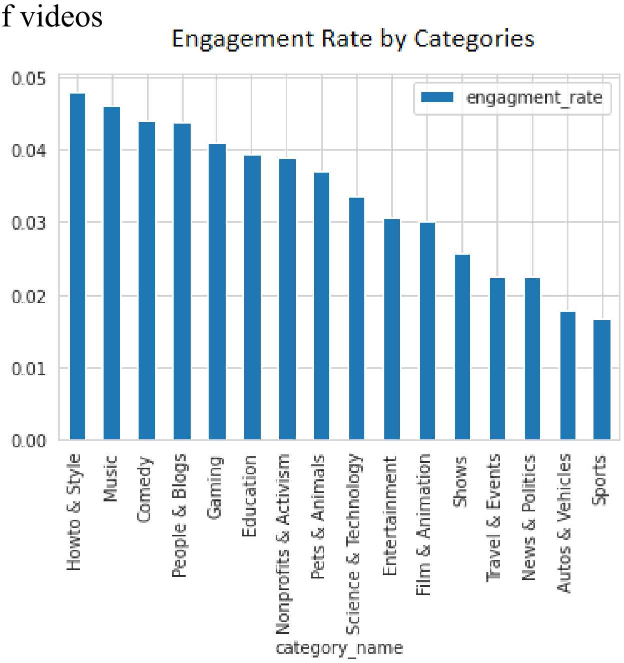
In addition to the exploratory analysis in the previous deliverable where we studied the relationships between several variables, we explored the text values in the tags column using Spacy NER. Because the tags used by the YouTuber can often reveal the themes of their channels and videos, we wanted to better understand the information contained within these tags. PERSON and ORG have the highest frequency among the 32,516 entities discovered in the data set.



To further understand the values of these tags, we also summarized the most frequent values within tag entities, apart from the most popular tags outlining the year of the videos and some numeric numbers. Sports-related tags such as “espn " and “nba” appear popular. Other frequently used tags include news channels such as “NBC” and “CBS.” Given this analysis, we speculate that sports and news are common themes among many of these videos in the data set.



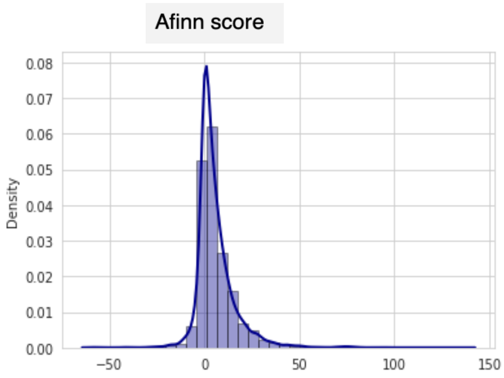
Previously, we calculated the engagement rate of videos and visualized the average engagement rate for each category as shown in this bar chart. Despite the high frequency of sports and news tags, we observed that these categories have some of the lowest engagement rates. One possible explanation could be that sports and news are very popular topics and users tend to use platforms like YouTube to stream sports and news events, which may lead to the view count being significantly higher than other categories. Because the engagement rate is calculated by dividing the sum of comment count, likes, and dislikes with view counts, the high view counts of videos in these categories could have offset the engagement rate.



4. Analytical Findings - Sentiment Analysis

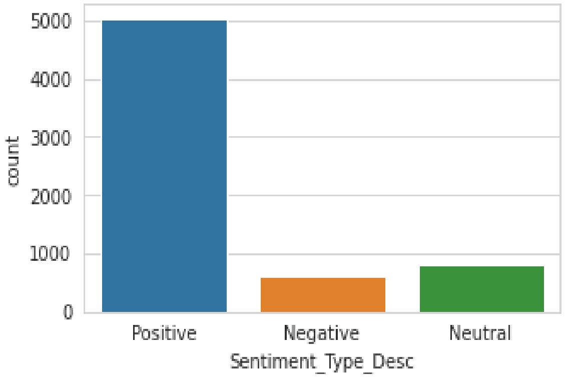
AFINN Score

For the description of the US video dataset, we generated the density and histogram to see the distribution of the AFINN score. As shown on the dark blue bar chart, we can see that most of the scores are concentrated in between 0 to 10, which means that most of the descriptions are positive attitudes.



Textblob

By using a sentiment analyzer from NLTK, we can examine the polarities of description from all Youtube Trending Videos in the US video dataset. We categorized the Description column into Positive and Negative sentiments using TextBlob. Negative indicates that most of the words have negative sentiment and positive in contrast, and we can see that most descriptions in the right plot are positive. Because YouTubers may use video descriptions to stimulate viewers' interests, it makes sense that the tone of most trending video descriptions is positive.

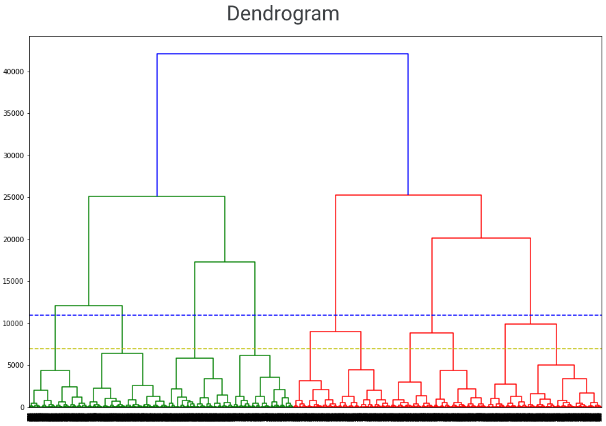


5. Analytical Findings - Clustering Analysis

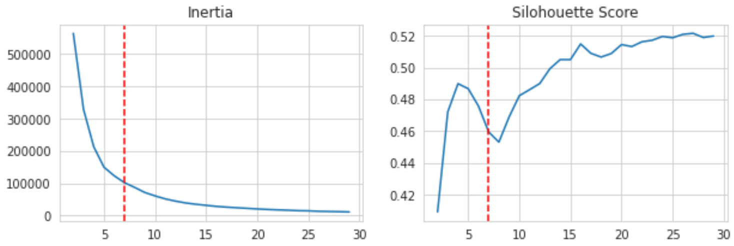
In order to find potential patterns in our text, we combined the tags and description column. We then utilized Spacy to create word vectors and similarities. In addition to the 300-word vectors, we also combined our numeric variables (eg: likes, dislikes, comments). Finally, we utilized UMAP to reduce the word embeddings to 2 dimensions.

Since our goal is to divide the videos into different playlists, we used two clustering methods: hierarchical and k-means clustering. After that, we compared the two methods and picked the more appropriate one.

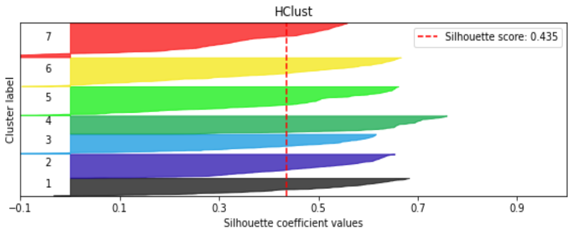
For the hierarchical clustering, we used four different dendrograms to visualize the hierarchical relationship and eventually chose the ward method because the data is noisy and the distribution of the ward method is more even. We found that the playlists could be divided into 7 clusters as shown in the dendrogram. The clusters are divided as shown under the blue line. In addition to the clean representation shown in the dendrogram, we also would like to avoid having too many playlists as it may be distracting for the users. The distances between data points at this level are also reasonable.

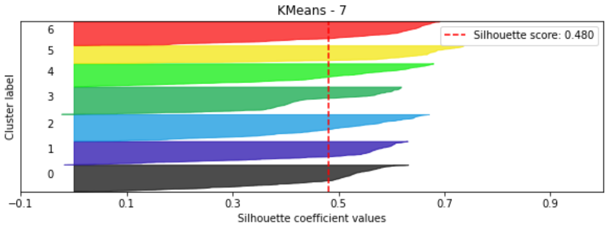


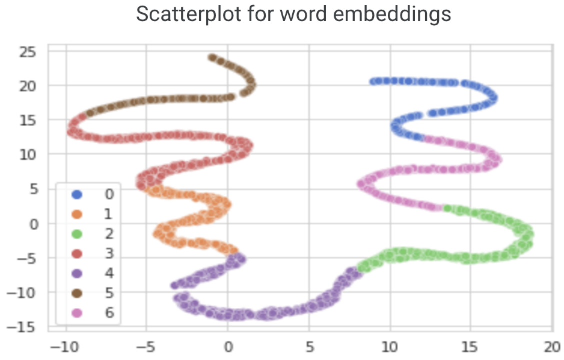
For the K-means method, we calculated the inertia and silhouette score, in which the elbow of the inertia is around 5 to 7, and the peak of the silhouette score is around 6 to 7. We also ran the lower right corner chart for better verification. For a better trade-off between the quantity and quality of clusters, we do not want the silo score to be extremely high or low, so we chose the k value of 7.



Next, we need to compare the two clustering methods via silhouette. The following silhouette analysis illustrates that the k-means clustering method is better than the hierarchical clustering method in this case. Firstly, from the distribution point of view, k-means is much more evenly distributed than hierarchical. Secondly, the silhouette plot of k- means shows almost no negative values. Lastly, most of the distribution of k-means exceeds the silo coefficient value. These all indicate that the distribution of clusters made by the k-means method is more evenly and reasonable, which can help make the viewing experience more pleasant for users.



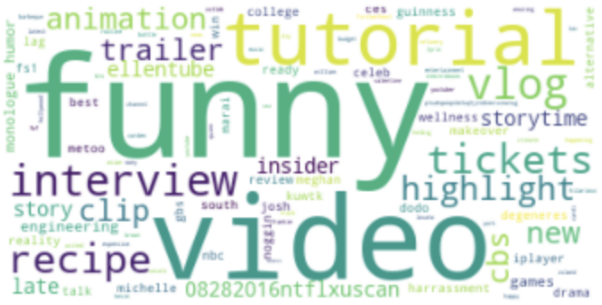


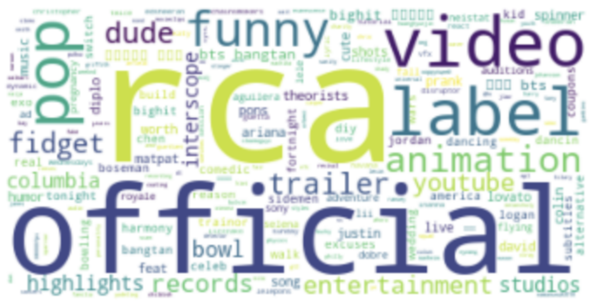


Based on our findings, we decided to proceed with our analysis with 7 clusters. We visualized the word vectors and the numeric variables in a 2-dimension way. Even though the axes are artificial vectors after the UMAP processing, we could still find all clusters are evenly distributed and correspond to connected areas in data space with high density. The areas in data space corresponding to clusters have certain characteristics (such as being convex or linear). We will apply our clusters back to the original dataset and find their insight in the following Word Cloud Analysis.

6. Analytical Findings - Word Cloud (keyBERT) in each cluster

The sample charts below display top words for cluster 0 and cluster 1:





We created the word cloud for each cluster to identify key text aspects, utilizing the tags column. These key aspects can help qualitatively label each playlist/cluster.

We utilized keyBERT to extract keywords in the tags column for each video and then identify the top keywords for each cluster based on the overall frequency. The sample word cloud, as shown above, shows that the most popular words in cluster 0 are “tutorial”, “funny”, “interview”, “vlog”, and “recipe”, which indicates that cluster 0 is more related to People & Blog videos. We also noticed that the most popular words in cluster 1 are “entertainment”, “official”, “trailer”. We can speculate that most official movie trailers are clustered in cluster 1.

While we were able to gain some insightful ideas on the topics of the clusters mentioned above, the word clouds for many other clusters are less intuitive. This may be because we incorporated many features in the clustering analysis, and some clusters are formed based on similarities that are irrelevant to the topics of the videos such as view counts and likes.

6. Conclusions & Recommendations

After some thorough research and analysis, we recommend YouTube generate 7 playlists containing the trending videos given in the data set. This would provide various options for the users who are interested in reviewing the trending videos while allowing YouTube to maintain a reasonable budget by not maximizing the number of playlists. We were also able to observe some significant keywords within a few clusters. However, because the analysis considered both the text and other characteristics of the videos, it is still hard to determine the specific theme of each playlist. In the future, we may consider analyzing the texts and other characteristics separately. Separating the features would allow us to create multiple sets of playlists for the same data set where the playlist sets focus on varying differentiation factors. Further analyzing specific features would also allow us to provide more customization to the playlists.