TRENDING YOUTUBE VIDEO STATISTICS

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

TABLE OF CONTENTS

01

BUSINESS PROBLEM

02

DATASET & DATA CLEANING

03

EXPLORATORY DATA ANALYSIS

04

ANALYTICAL FINDINGS

- SENTIMENT ANALYSIS
- CLUSTERING ANALYSIS

05

CONCLUSION & FINDINGS

BOSTON UNIVERSITY

BACKGROUND

YouTube is an online video-sharing platform owned by Google, accessible from various devices & platforms such as computers, phones, gaming consoles, and smart TVs.

- 0.13 stickiness factor
- 315.12 million daily active users
- 2.3 billion monthly active users
- $19.7 billion in revenue in 2020.

BOSTON UNIVERSITY

BUSINESS PROBLEM

BUSINESS GOALS

- Commemorate some of the most impactful videos
- Create a series of playlists that capture the daily trending videos
- Playlists will consist of videos that are similar in nature for users' interests

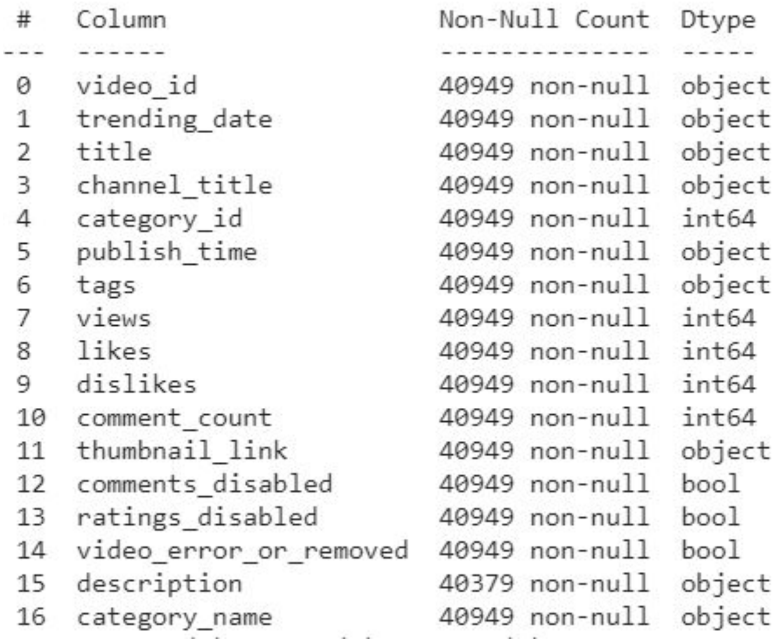
PROJECT DESCRIPTION

- Determine the number of playlists to create
- Generate relevant information for each playlist
- Personalized playlists can be created in future
- Serve as a pilot for individual playlist customization

BOSTON
UNIVERSITY

DATA SET

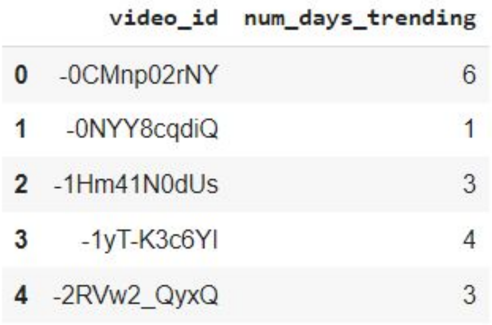
- The dataset ‘Trending YouTube Video Statistics’ was downloaded from Kaggle, originally scraped from the YouTube library
- Several months of data on daily trending YouTube videos in **2017 and 2018**
- This project will focus on the data outlining trending videos in the **US**
- Consisted of a csv file with **16 columns** and **40k rows**, and json file containing category names



```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   video_id                40949 non-null  object
 1   trending_date           40949 non-null  object
 2   title                   40949 non-null  object
 3   channel_title           40949 non-null  object
 4   category_id             40949 non-null  int64
 5   publish_time            40949 non-null  object
 6   tags                    40949 non-null  object
 7   views                   40949 non-null  int64
 8   likes                   40949 non-null  int64
 9   dislikes                40949 non-null  int64
 10  comment_count           40949 non-null  int64
 11  thumbnail_link          40949 non-null  object
 12  comments_disabled       40949 non-null  bool
 13  ratings_disabled        40949 non-null  bool
 14  video_error_or_removed  40949 non-null  bool
 15  description             40379 non-null  object
 16  category_name           40949 non-null  object
```

DATA CLEANING AND ML PREPROCESSING

- Initial Data had **40k** rows, but only **6k** unique videos
- Videos can be trending over multiple days, hence many rows for each video
- Latest entry taken for video id based on date
- New feature created ‘number of data trending’ to capture this datapoint
- **Regex** used to clean text columns of unnecessary characters



	video_id	num_days_trending
0	-0CMnp02rNY	6
1	-0NYY8cqdiQ	1
2	-1Hm41N0dUs	3
3	-1yT-K3c6YI	4
4	-2RVw2_QyxQ	3

spaCy

UNIFORM MANIFOLD
UMAP
APPROXIMATION & PROJECTION

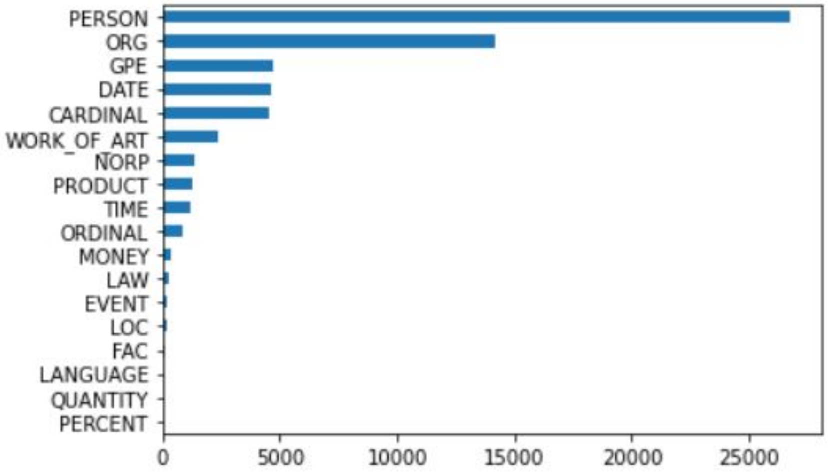
- Combined the tags and description column.
- Utilized Spacy to create word vectors and similarities.
- Feature space : 300-word vectors +numeric variables (eg: likes, dislikes, comments).
- UMAP to reduce the word embeddings and numeric data to 2 dimensions.

BOSTON
UNIVERSITY

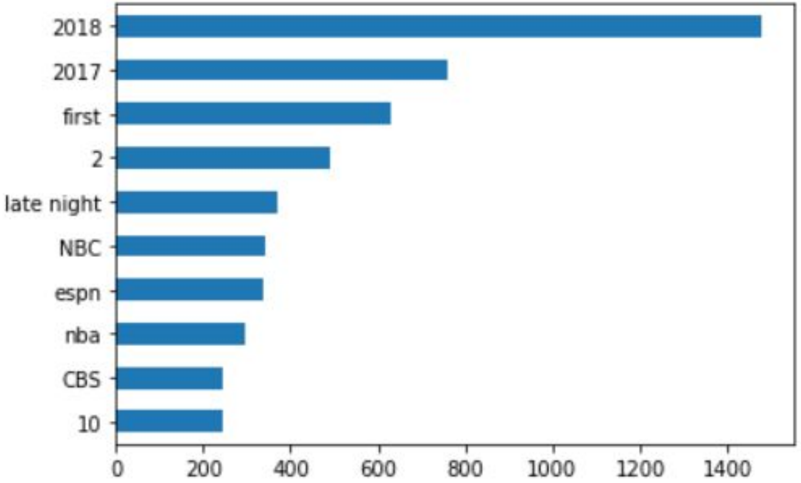
EXPLORATORY DATA ANALYSIS

- PERSON and ORG are the most frequently used entities in the tags of those trending videos
- 2018 and 2017 (year) are the most used entity values, but more insights can be drawn from other frequent values such as NBC, espn, nba, and CBS. These values suggest that sports and news are common topics among trending videos

Entity Frequency



Most Frequent Entity Values

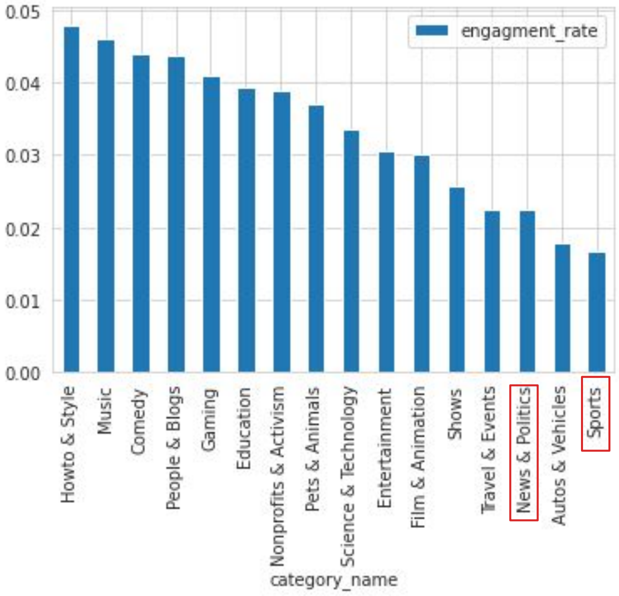


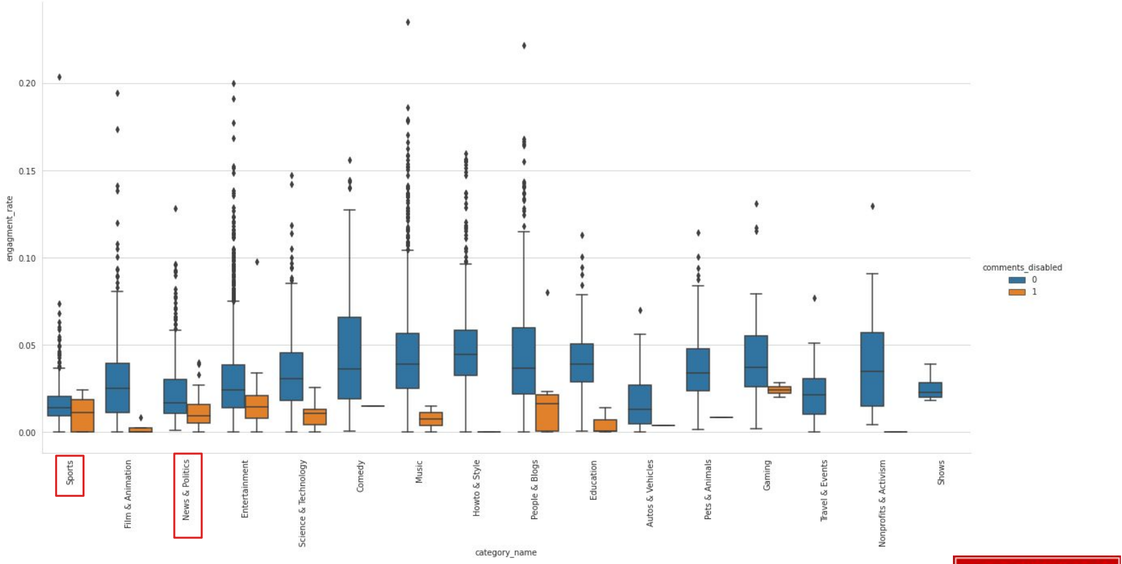
BOSTON UNIVERSITY

EXPLORATORY DATA ANALYSIS

- Despite the high frequency in tags, news and sports appear to be some of the least engaging videos
- Engagement rate = (likes + dislikes + comment count)/views

Engagement Rate of Each Category

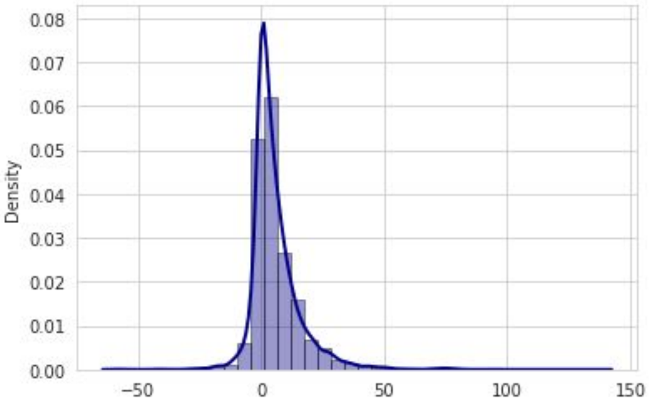




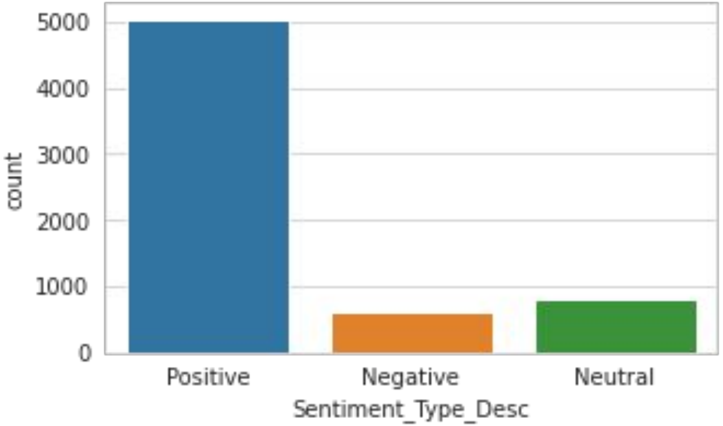
SENTIMENT ANALYSIS

- According to the Afinn score and textblob, most of the description are positive attitude
- Reason might be positive descriptions can get more viewers and increase watch time

Afinn Score



Textblob

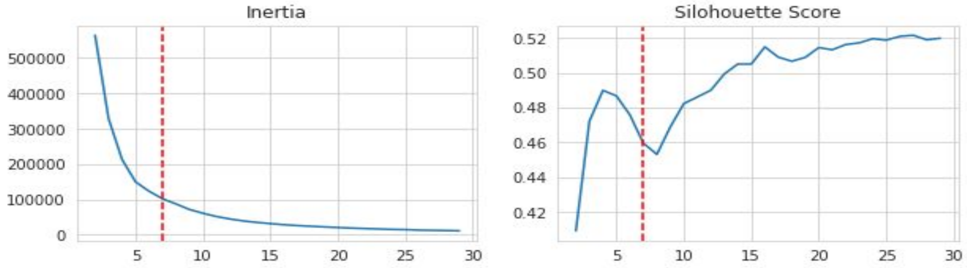


BOSTON UNIVERSITY

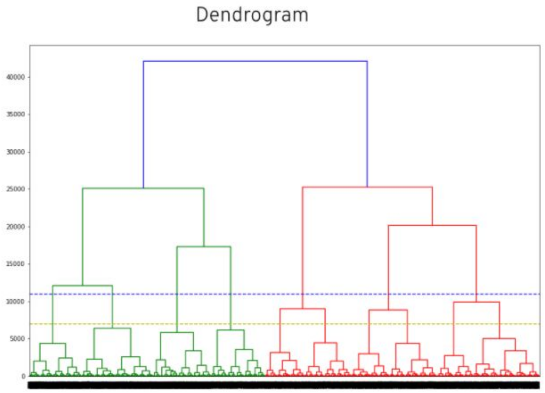
CLUSTERING ANALYSIS

- Both H-clustering and K-means clustering recommended for setting 7 playlists

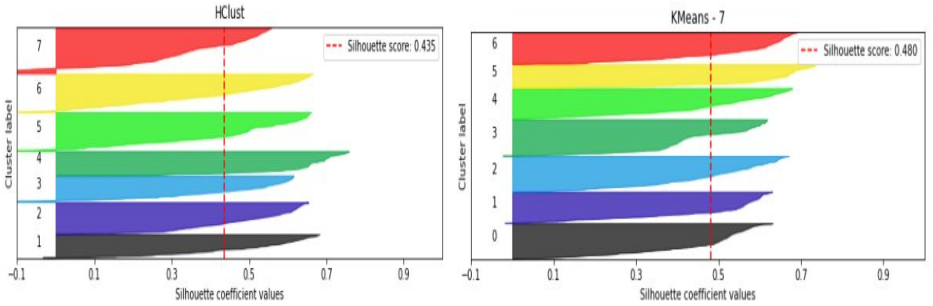
K-means Clustering



Hierarchical Clustering



Comparison using Silo

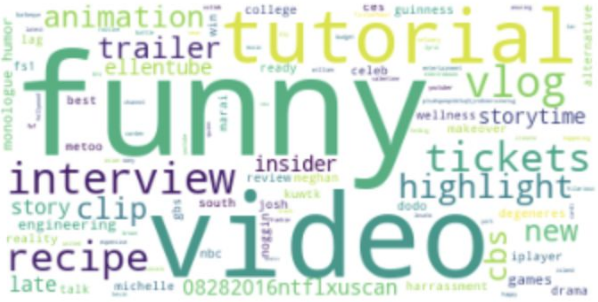


WORD CLOUD

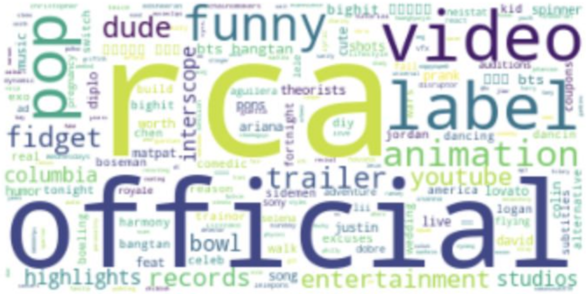
- Identify key text aspects to qualitatively label each playlist/cluster

- 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

Cluster -0



Cluster -1



CONCLUSIONS & RECOMMENDATIONS

- Generate 7 playlists containing the trending videos
- Significant keywords within specific playlists
- Analyzing specific features to provide more personalized playlists