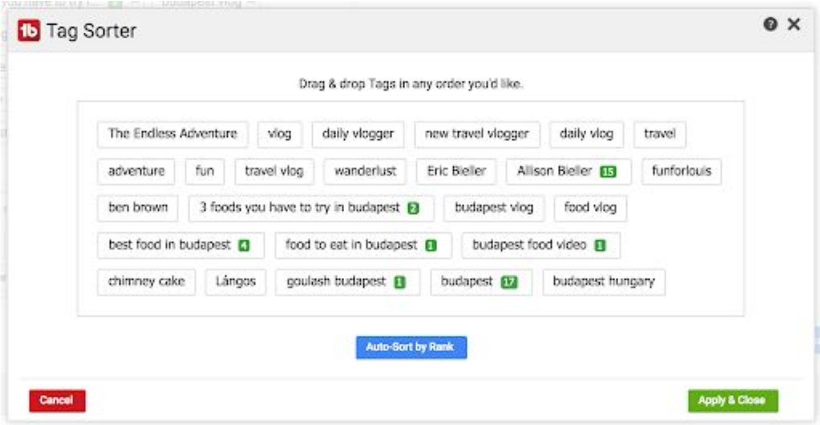
NLP Analysis of Youtube Tags

Asher Khan

What are tags and why are they important?

Youtube tags allows youtube to grasp the video’s content and category, and associates it with similar content. This can amplify your video’s reach!


Tag Sorter

Drag & drop Tags in any order you'd like.

The Endless Adventure | vlog | daily vlogger | new travel vlogger | daily vlog | travel
adventure | fun | travel vlog | wanderlust | Eric Bieller | Allison Bieller 15 | funforlouis
ben brown | 3 foods you have to try in budapest 2 | budapest vlog | food vlog
best food in budapest 4 | food to eat in budapest 1 | budapest food video 1
chimney cake | Lángos | goulash budapest 1 | budapest 17 | budapest hungary

https://blog.hubspot.com/marketing/youtube-tags

Overview

Business Case: I will use different machine learning methods to predict how many views a Youtube video will render based on the tags used and show feature importance.

➔ **Data**

Three data sets were taken from kaggle and joined.

➔ **EDA**

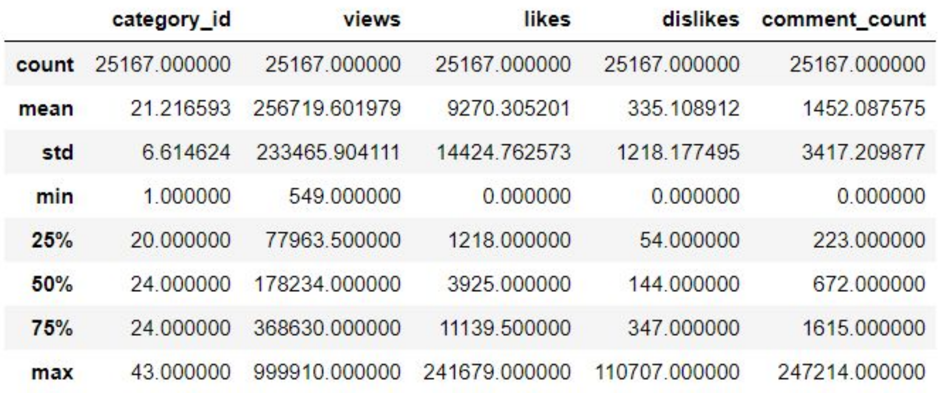
The data was analysed.

➔ **Modeling**

Countvectorizer and TF-IDF methods were used to train the data along with RF regression for feature importance.

Data

- Data sets of trending youtube videos in the U.S, Canada, and Great Britain were joined.
- There were a total of 25,167 videos after cleaning.



	category_id	views	likes	dislikes	comment_count
count	25167.000000	25167.000000	25167.000000	25167.000000	25167.000000
mean	21.216593	256719.601979	9270.305201	335.108912	1452.087575
std	6.614624	233465.904111	14424.762573	1218.177495	3417.209877
min	1.000000	549.000000	0.000000	0.000000	0.000000
25%	20.000000	77963.500000	1218.000000	54.000000	223.000000
50%	24.000000	178234.000000	3925.000000	144.000000	672.000000
75%	24.000000	368630.000000	11139.500000	347.000000	1615.000000
max	43.000000	999910.000000	241679.000000	110707.000000	247214.000000

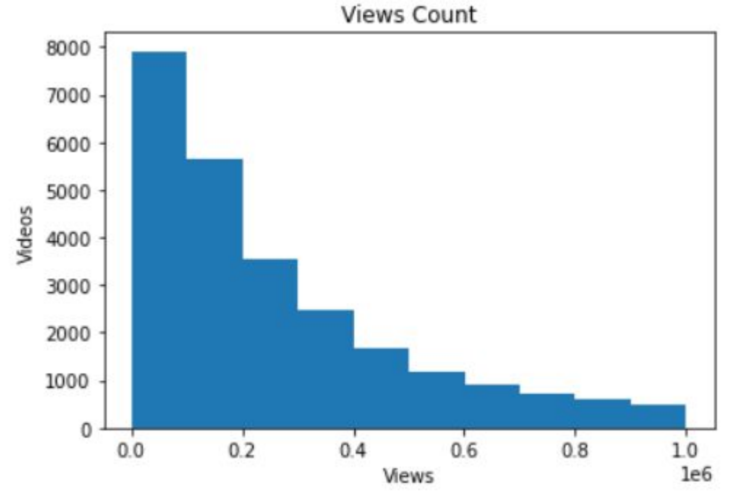
Cleaning

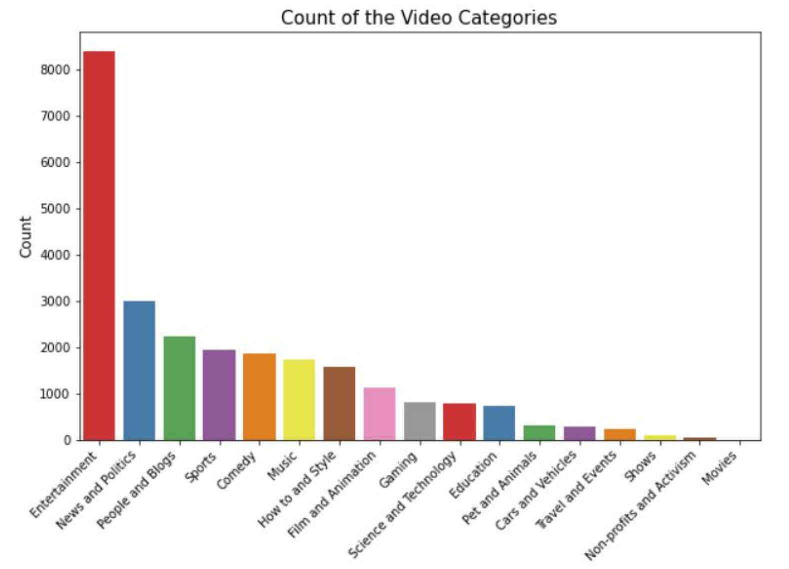
Videos were maxed out at 1 million views to remove outliers.

Videos with no tags were dropped.

EDA

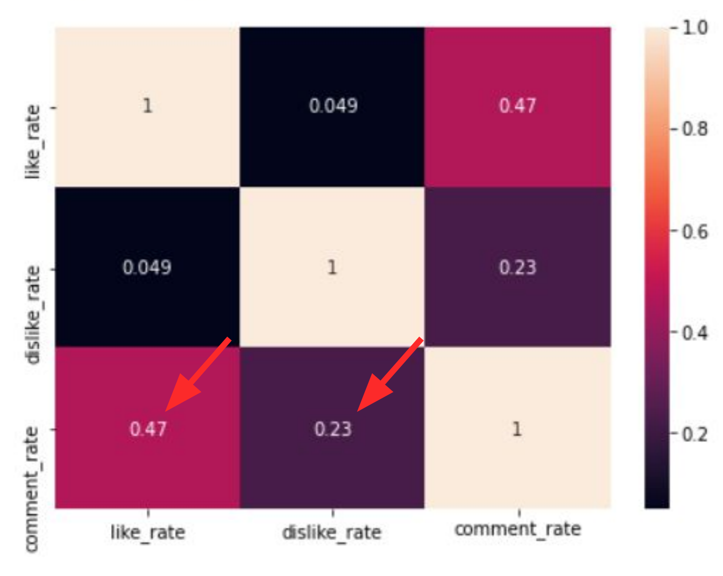
- As expected, the number of videos decreased as the views increased.
- Most videos were in the entertainment category. News and Politics being a far second.

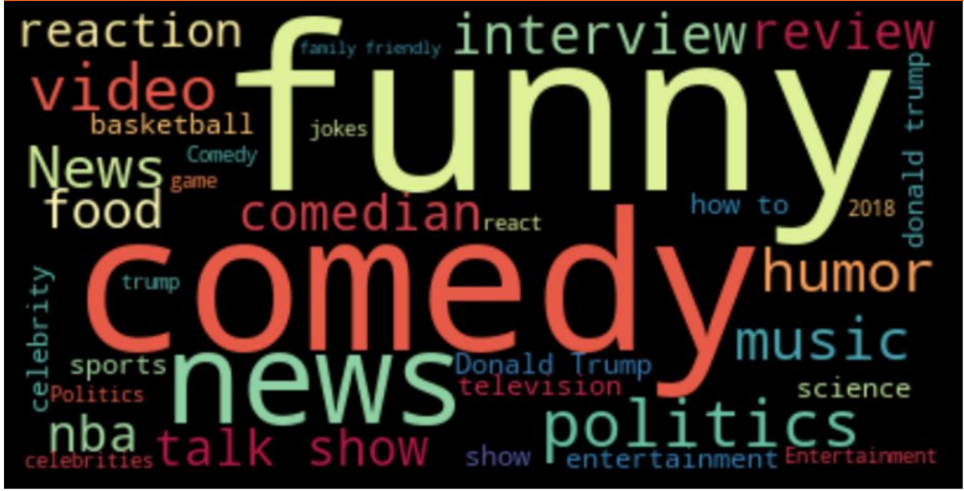




EDA

- 47% chance a person will comment if they like the video.
- 23% chance they will comment if they dislike it.
- Top 3 most common tags:
  - ○ Funny - 2,014
  - ○ Comedy - 1,409
  - ○ News - 1,012





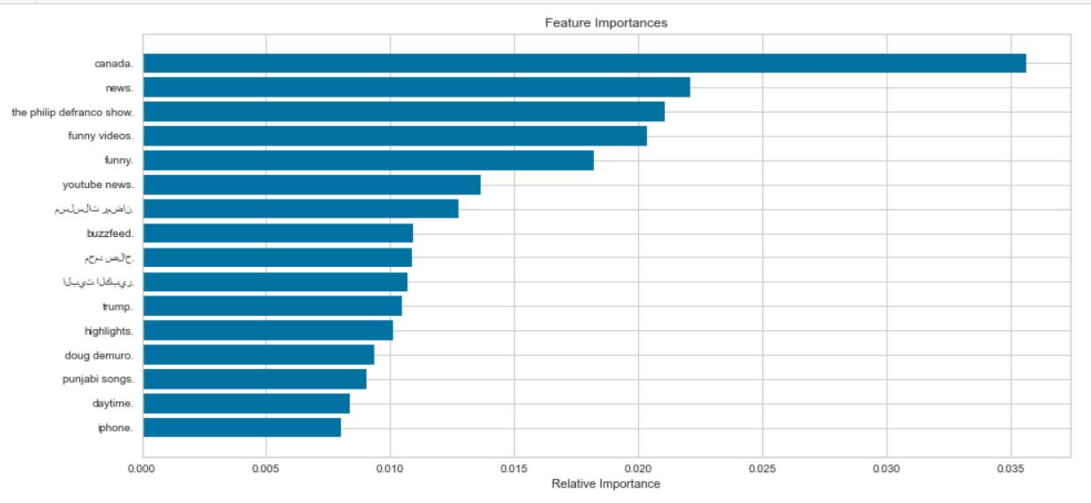
_ Modeling

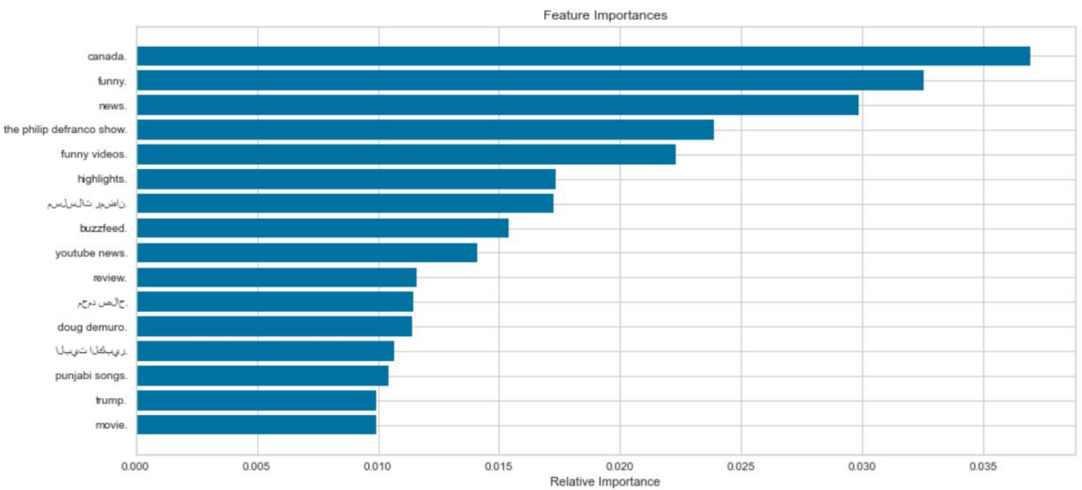
CountVectorization:

- Train:
  - R Squared = 0.344
  - Mean Sqrd Error = 188,782
- Test:
  - R Squared = 0.182
  - Mean Sqrd Error = 212,470

TF-IDF:

- Train:
  - R Squared = 0.341
  - Mean Sqrd Error = 189,268
- Test:
  - R Squared = 0.173
  - Mean Sqrd Error = 213,632





Limitations

Although tags are a factor in affecting views, it is not the only or even the primary.

➔ **What**

Channel
Popularity
Shares
Time Posted
Etc.

➔ **Utility Recommendations**

Use top tags (based on feature importance) for relevant videos to maximize the number of views.

Future Work

- Explore some of the limitations. Specifically using data from a single channel.
- Exploring time the video was posted

Thank You!

Yish

FLATIRON SCHOOL

https://github.com/asherkhan7/Capstone-Project.git