# Building a Content-Based Youtube Video Recommender System

Capstone Project 2

Anna Apa

# Goal:

Build a content-based recommender system that recommends similar Youtube videos based on the video's features

# Data

The following JSON and CSV files were ownloaded from Kaggle.

- GB_category_id.json
- GBvideos.csv
- US_category_id.csv
- USvideos.json

The datasets provide information on the top trending videos in Great Britain and the United States from 11/14/17 to 06/14/18. Each day, there are around 150-200 videos included on that list for each country.

# GBvideos.csv & USvideos.csv

| |
|---|
| video_id |
| trending_date |
| title |
| channel_title |
| category_id |
| publish_time |
| tags |
| views |
| likes |
| dislikes |
| comment_count |
| thumbnail_link |
| comments_disabled |
| ratings_disabled |
| video_error_or_removed |
| description |

- Both csv files were converted to Pandas DataFrames. Both DFs contain the columns shown on the left
- Null values: the Great Britain data has some null values, but since this is just 0.23% of the data, we simply drop it
- The trending_date column contains the date when that video was on the trending list. It initially had an object type. To convert this to a datatime type, we first changed its current format from year/day/month to year/month/day

# GB_category_id.json & US_category_id.json

Both JSON files were read as Pandas DataFrames and had the following format:

| | kind | etag | items |
|---|---|---|---|
| 0 | youtube#videoCategoryListResponse | "m2yskBQFythfE4irbTIeOgYYfBU/1v2mrzYSYG6onNLt2... | {'kind': 'youtube#videoCategory', 'etag': '"m2... |
| 1 | youtube#videoCategoryListResponse | "m2yskBQFythfE4irbTIeOgYYfBU/1v2mrzYSYG6onNLt2... | {'kind': 'youtube#videoCategory', 'etag': '"m2... |

Since the GBvideos and USvideos DFs from the previous slide only contain the category_id of each video, we need to get corresponding category name for those category IDs. We are only interested in the “items” column of this DF.

# Feature Engineering (continued)

We extract the information found on the “items” column and create the dataframe on the bottom. We then merge this with existing main dataframes.


```
{'kind': 'youtube#videoCategory',
 'etag': '"m2yskBQFythfE4irbTIeOgYYfBU/Xy1mB4_yLrHy_BmKmPBggty2mZQ"',
 'id': '1',
 'snippet': {'channelId': 'UCBR8-60-B28hp2BmDPdntcQ',
  'title': 'Film & Animation',
  'assignable': True}}
```


| | category_id | category |
|---|---|---|
| 0 | 1 | Film & Animation |
| 1 | 2 | Autos & Vehicles |
| 2 | 10 | Music |
| 3 | 15 | Pets & Animals |
| 4 | 17 | Sports |

# Feature Engineering (continued)

We count how many tags were used in each video and add it on a new column

| | tags | tag_count |
|---|---|---|
| 0 | SHANtell martin | 1 |
| 1 | last week tonight trump presidency\|"last week ... | 4 |
| 2 | racist superman\|"rudy"\|"mancuso"\|"king"\|"bach"... | 23 |
| 3 | rhett and link\|"gmm"\|"good mythical morning"\|"... | 27 |
| 4 | ryan\|"higa"\|"higatv"\|"nigahiga"\|"i dare you"\|"... | 14 |

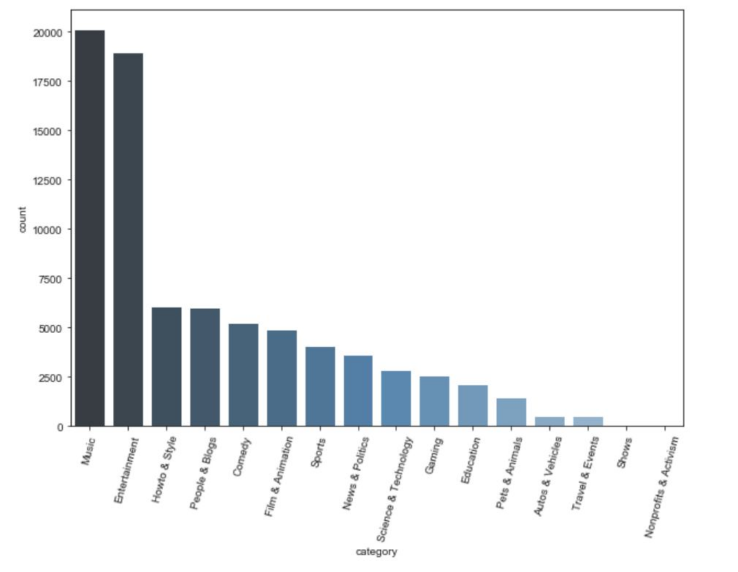
# Main DataFrames

1) We merge our US and GB dataframes horizontally

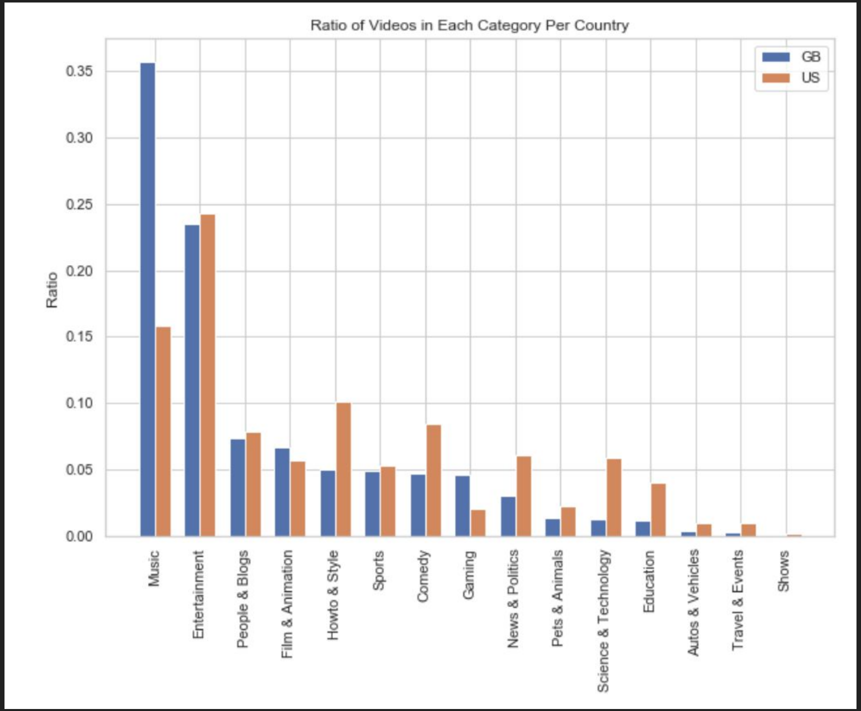
2) Since videos can appear be on the trending list for more than just a day, we get the final stats for each video-- such as its number of views, comments, likes and dislikes

# Exploratory Data Analysis

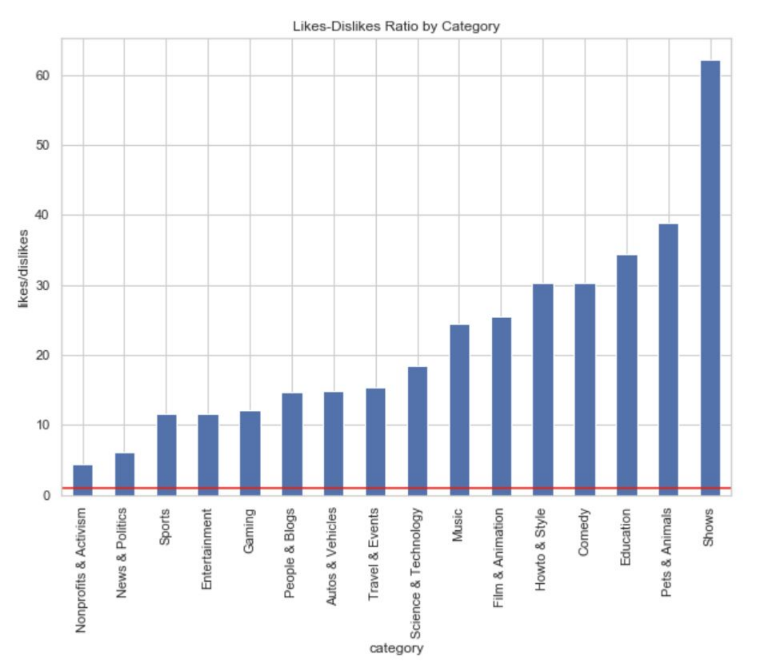
Which categories have the most trending videos?



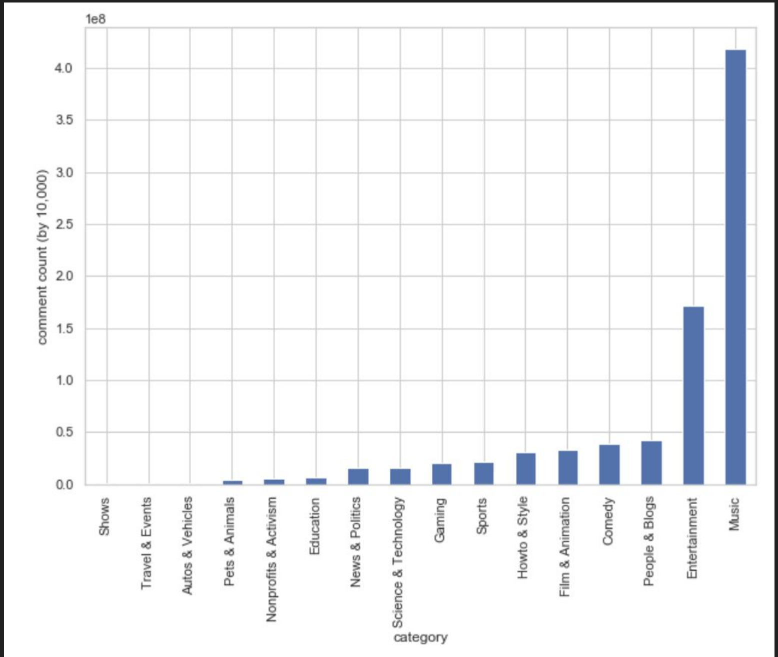
# Which categories are more popular in each country?



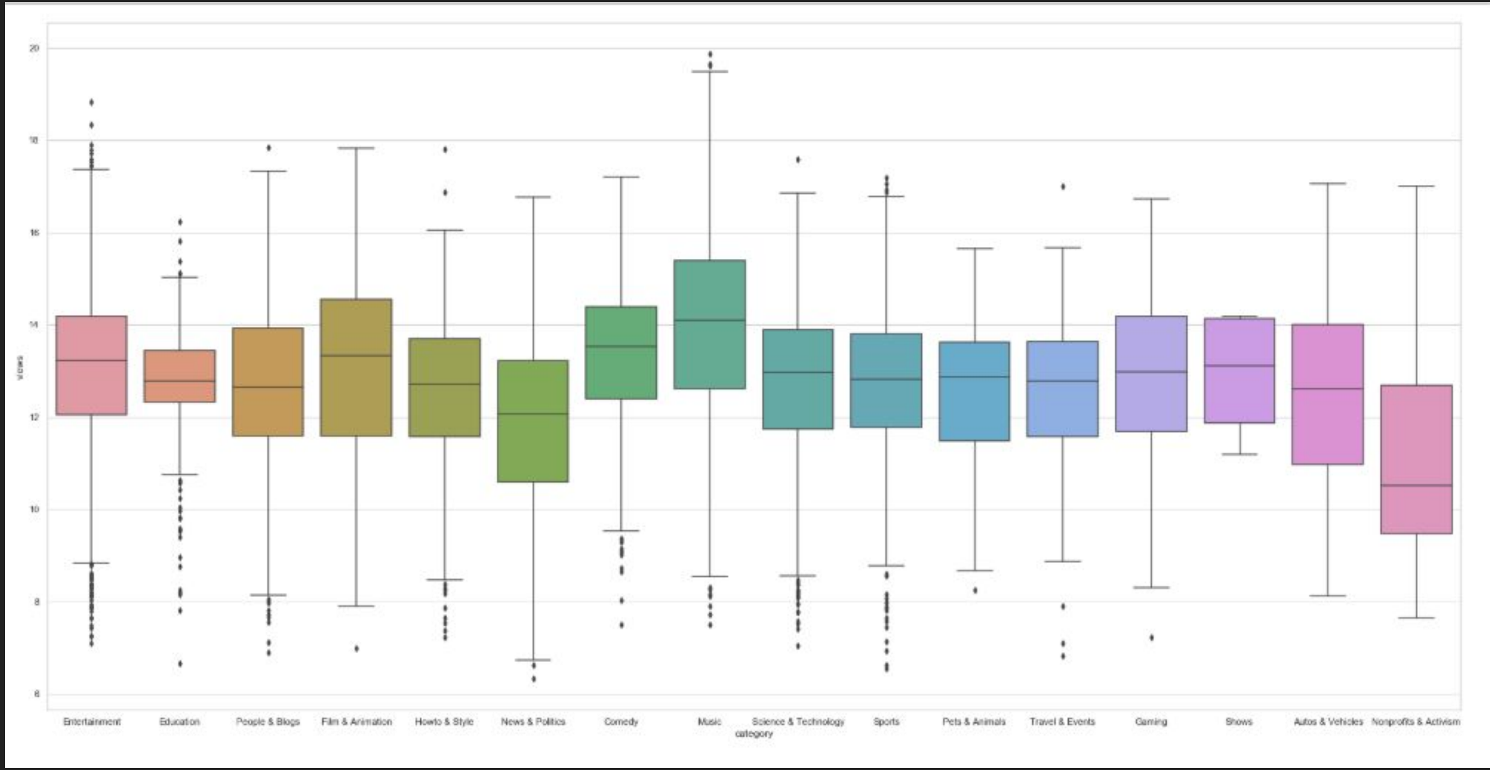
# Which categories have the lowest like/dislike ratio?



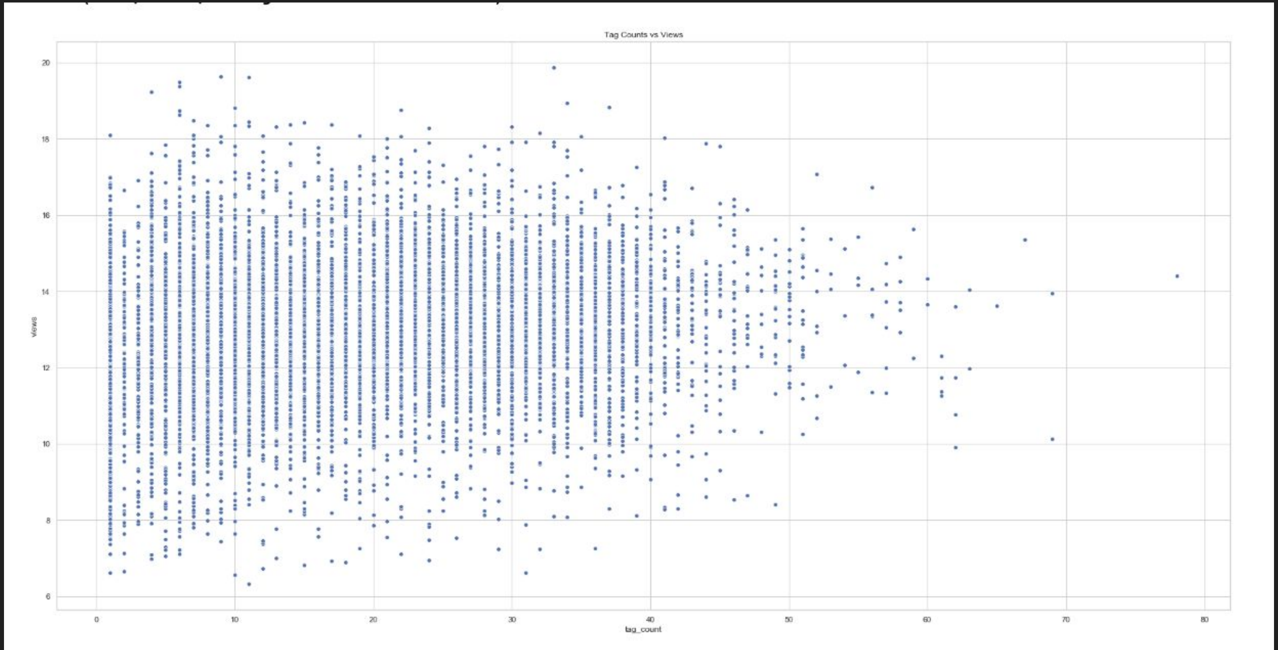
# Which categories usually have more comments?



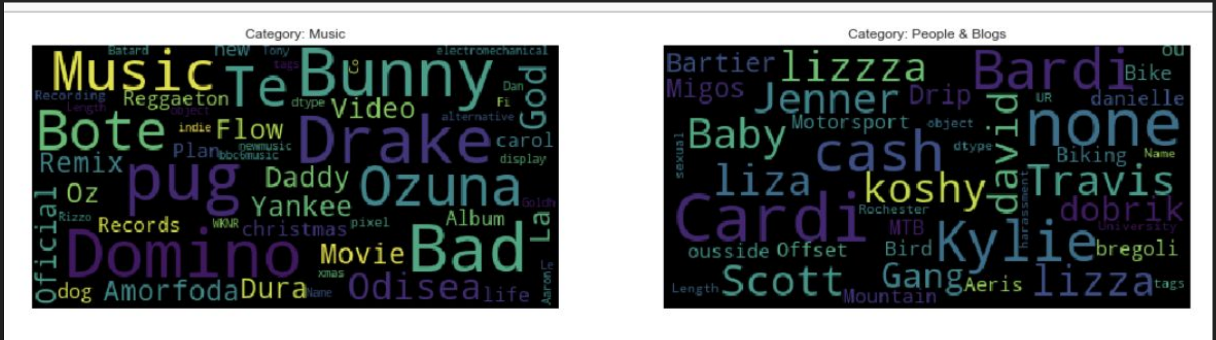
# Which categories had the most views?

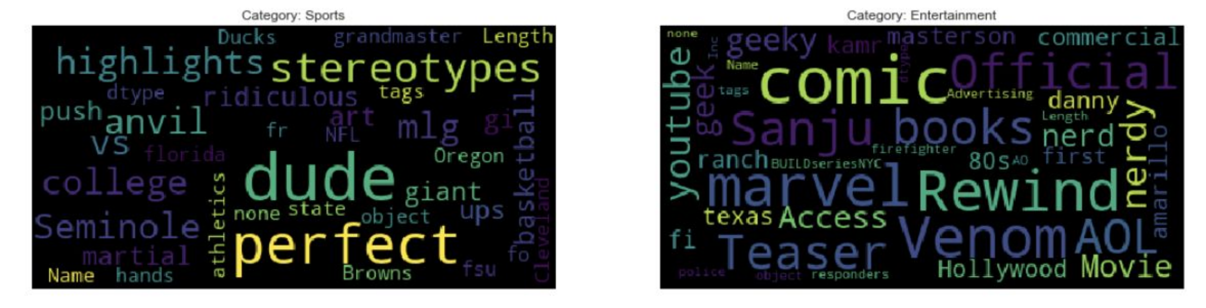


# Do videos with more tag counts have more views?



# Word Cloud plots for each category



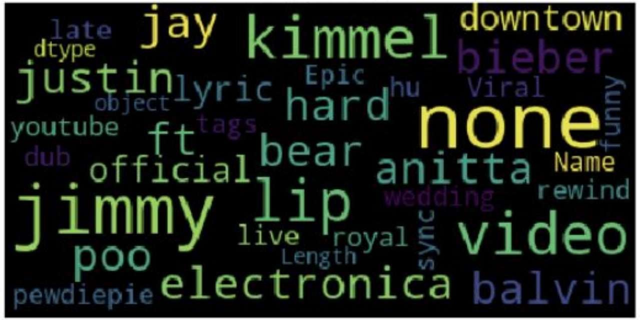



# Word Cloud Plots (continued)

Category: Pets & Animals

Category: Howto & Style

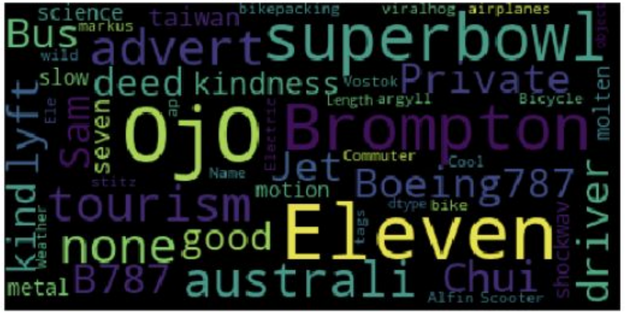
Category: Film & Animation



Category: Comedy

# Word Cloud Plots (continued)

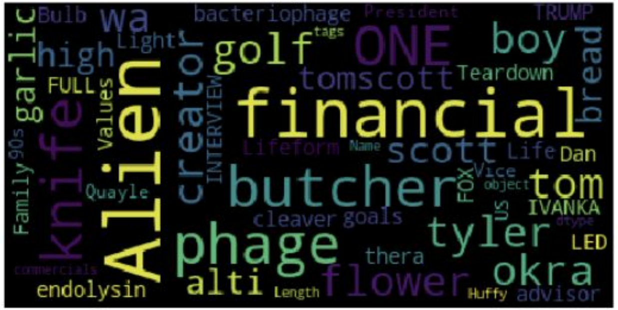
Category: Travel & Events




Category: Gaming


Category: News & Politics


Category: Education

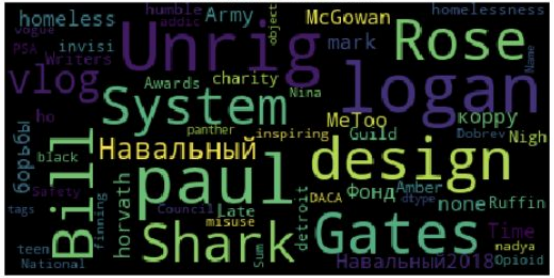


# Word Cloud Plots (continued)

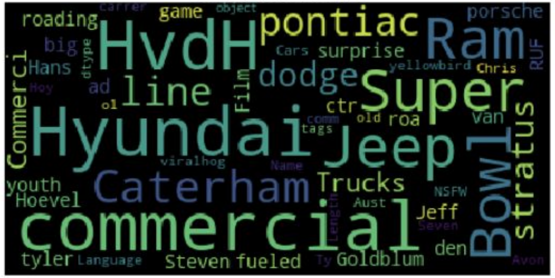
Category: Science & Technology


Category: Shows


Category: Nonprofits & Activism




Category: Autos & Vehicles



# Inferential Statistics

Do features such as the number of views, comments, likes, dislikes and etc. vary significantly among categories? We apply inferential statistics techniques to answer this. In this case, we use F-tests to statistically test the equality of means in each category.

# Inferential Statistics (continued)

For each feature, our null hypothesis is that they are not statistically different from each other across categories; and our alternative hypothesis is that the differences are statistically significant from each other.

# Inferential Statistics (results)

| Feature | F-Stat | P-Value | Statistically Significant*? |
|---|---|---|---|
| Views | 29.91 | 0.0000 | **Yes** |
| Likes | 544.82 | 0.0000 | **Yes** |
| Dislikes | 60.06 | 0.0000 | **Yes** |
| Comment Count | 148.42 | 0.0000 | **Yes** |
| Tag Count | 51.10 | 0.000 | **Yes** |

* at a 5% level

# Building the Recommender System

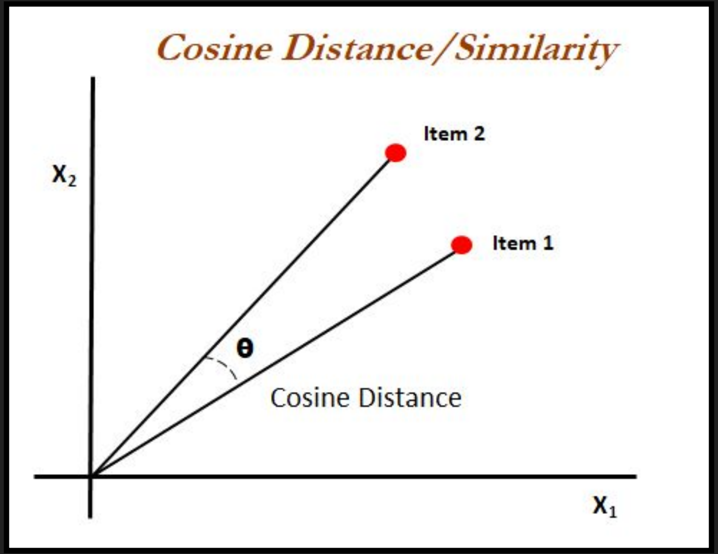
We will use NLP techniques to base our recommender system on the tags used (as well as other features). The more tags used in common, the more likely it is to be recommended to the user.

To do this, we create a TF-IDF model. TF-IDF stands for “Term-Frequency X Inverse Document Frequency”. TF-IDF is "essentially a measure of term importance, and how discriminative a word is in a corpus".

$$(t,\ d) = (t,\ d) \times (t) = n_{td} \log\left( \frac{|D|}{|d\ :\ t \in d|} + 1 \right)$$

# Building the Recommender System(continued)

After building a TF-IDF model, we quantify the similarity between the vectors by calculating their cosine similarities.




$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\,\sqrt{\sum_{i=1}^{n} B_i^2}}$$

# Recommender 1 - Based on Tags

Recommendations for "Drake – God's Plan":

| | title | channel_title | views | category | comment_count |
|---|---|---|---|---|---|
| 162 | Drake - God's Plan (Official Music Video) | OVO Sound | 19586636 | Music | 66540 |
| 1660 | Gods Plan - Drake (William Singe Cover) | William Singe | 2004311 | Music | 3007 |
| 709 | Guillermo – God's Plan | Jimmy Kimmel Live | 4720264 | Entertainment | 3541 |
| 37 | Drake - Nice For What | OVO Sound | 60635812 | Music | 55653 |
| 6654 | Drake's Dad Claps Back At Wendy Williams! He C... | Lailah Lynn | 118828 | Entertainment | 741 |
| 3726 | The epic late-night Fortnite stream featuring ... | ESPN | 636473 | Sports | 1355 |
| 1263 | DRAKE & NINJA PLAY DUOS ON FORTNITE! \| Fortnit... | Twitch Moments | 2663507 | Gaming | 1796 |

# Recommender 2 - Videos from Different Categories

Result from our first recommender system:

```
Recommendations for "The End of the F**king World | Official Trailer [HD] | Netflix":
```

| | title | channel_title | views | category | comment_count |
|---|---|---|---|---|---|
| 2027 | THE CLOVERFIELD PARADOX \| WATCH NOW \| NETFLIX | Netflix | 1575608 | Entertainment | 3032 |
| 459 | The Kissing Booth \| Official Trailer [HD] \| Ne... | Netflix | 7308023 | Entertainment | 5790 |
| 6508 | Derren Brown: The Push I Official Trailer [HD]... | Netflix | 135325 | Entertainment | 395 |
| 5128 | Irreplaceable You \| Official Trailer [HD] \| Ne... | Netflix | 316756 | Entertainment | 506 |
| 6991 | The Moment George Clooney Met Amal \| My Next G... | Netflix | 91193 | Entertainment | 82 |

Recommender 2:

```
Recommendations for "The End of the F**king World | Official Trailer [HD] | Netflix":
```

| | title | views | category | comment_count |
|---|---|---|---|---|
| 4072 | Bright: What Went Wrong? – Wisecrack Edition | 533207 | Education | 4618 |
| 5059 | Reboot: The Guardian Code Official Trailer | 326871 | Film & Animation | 5071 |
| 8085 | 'I have dad moves': Barack Obama discusses dan... | 21700 | News & Politics | 70 |
| 2251 | Ep4 It's on you and I \| BTS: Burn the Stage | 1378098 | Music | 1376 |
| 5194 | HIS & HERS BOUJEE NIGHT OUT! VLOGMAS WEEK 1 | 304917 | Howto & Style | 1204 |

# Recommender 3 - Based on Description

```
Recommendations for "Drake – God’s Plan":
```

| | title | channel_title | views | category | comment_count |
|---|---|---|---|---|---|
| 156 | Nicki Minaj - Barbie Tingz (Lyric Video) | NickiMinajAtVEVO | 20262996 | Music | 40235 |
| 79 | Nicki Minaj - Chun-Li | NickiMinajAtVEVO | 36759844 | Music | 93136 |
| 1086 | N.E.R.D, Rihanna - Lemon (Drake Remix - Audio)... | NERDVEVO | 3126660 | Music | 2640 |
| 681 | Nicki Minaj - Chun-Li (Live on SNL / 2018) | NickiMinajAtVEVO | 4945185 | Music | 18683 |
| 13 | Post Malone - Psycho ft. Ty Dolla $ign | PostMaloneVEVO | 105629911 | Music | 45784 |

# Recommender 4 - Based on Tags, Description and Channel ID

```
'Drake new music|"Drake Gods Plan"|"Drake God's Plan"|"Scary Hours"|"Drake Charity Giveaway"|"Drake in Miami"'
```

+

```
'God's Plan (Official Video)\\n\\nSong Available Here: https://Drake.lnk.to/ScaryHoursYD\\n\\n \\n\\nDirected by Kare
na Evans\\n\\nExecutive Producers Director X & Taj Critchlow\\n\\nProduced by Fuliane Petikyan\\n\\nFor Popp Rok\\n
\\n \\n\\nMusic video by Drake performing God's Plan.  © 2018 Young Money Entertainment/Cash Money Records\\n\\nhtt
p://vevo.ly/Z6Unb9'
```

+

```
'DrakeVEVO'
```

=

```
'Drake new music|"Drake Gods Plan"|"Drake God's Plan"|"Scary Hours"|"Drake Charity Giveaway"|"Drake in Miami" God's P
lan (Official Video)\\n\\nSong Available Here: https://Drake.lnk.to/ScaryHoursYD\\n\\n \\n\\nDirected by Karena Evans
\\n\\nExecutive Producers Director X & Taj Critchlow\\n\\nProduced by Fuliane Petikyan\\n\\nFor Popp Rok\\n\\n \\n\\n
Music video by Drake performing God's Plan.  © 2018 Young Money Entertainment/Cash Money Records\\n\\nhttp://vevo.ly/
Z6Unb9 DrakeVEVO DrakeVEVO DrakeVEVO '
```

# Recommender 4 - Based on Tags, Description and Channel ID

```
Recommendations for "Drake – God's Plan":
```

| | title | channel_title | views | category | comment_count |
|---|---|---|---|---|---|
| 1086 | N.E.R.D, Rihanna - Lemon (Drake Remix - Audio)... | NERDVEVO | 3126660 | Music | 2640 |
| 7329 | Drake Bell - Rewind | DrakeBellVEVO | 65973 | Music | 836 |
| 156 | Nicki Minaj - Barbie Tingz (Lyric Video) | NickiMinajAtVEVO | 20262996 | Music | 40235 |
| 37 | Drake - Nice For What | OVO Sound | 60635812 | Music | 55653 |
| 79 | Nicki Minaj - Chun-Li | NickiMinajAtVEVO | 36759844 | Music | 93136 |

# Conclusion

In this project, I have built four  content-based recommender systems based on different features.

- For the first two, we focus on the tags used in each video.
- For the second one, I only allow for recommendations outside the category. We see that this does not do very well because the videos don’t seem to be very similar, except for the fact that they had very broad tags (such as “Netflix”) in common.
- For the third one, we focus solely on the description. This can be more helpful for show or movie trailers since actors, directors and plots are often included in the description. Although, there is a higher chance of videos with little to no description will be paired with unrelated videos.
- For our fourth recommender system, we try to address all of these problems by combining the tags used, description, and channel title all in one string.

# Limitations

- The videos in the dataset are only videos that were included in the daily “Top Trending List”  in the United States and Great Britain for six months.
- Data on the videos the user has liked would improve the content-based recommendation model