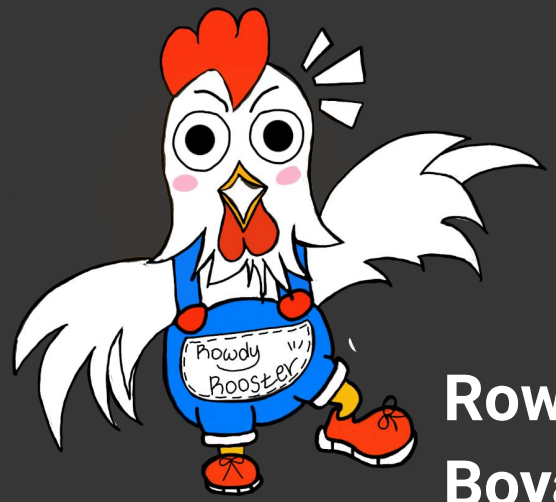
Trending Video Analysis

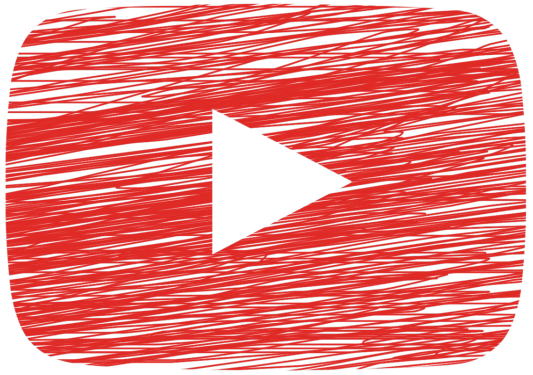


Rowdy Rooster Team
Boya Li , David Moon, EunJeong Heo

Contents

Introduction

Why YouTube is important?



> As of 2020, there are **2+ billion** logged-in monthly users out there. The number of channels grew more than **23%**, and the number of channels that earn **$10K USD** per year or more grew by **50%**. **70% of viewers** bought from a brand after seeing it on YouTube.

YouTube is the second most visited website in the world!

Introduction

10 Crucial YouTube Ranking Factors

(www.searchenginepeople.com)

- Channel Keywords
- Video Title
- Video Description
- Video Tags
- Video Quality
- User Experience Metrics
- Watch Time
- View Count
- Thumbnails
- Closed Captions & Subtitles

Question to answer

What factors lead YouTube videos to a trending video?

Objective

- Identify hot topics and key variables that contribute to a video's success.
- Analyze trending video data over the previous two years to discover possible variations between variables across separate English-speaking countries/ regions.

Project/Dataset Summary

Hypothesis and Findings

- There are key differences in trending video metrics among the three English speaking regions

- Title length, tag length, published time, and category show similar trends across the three countries

- The variables might have the similar trend for the trending videos among those three English speaking countries

Dataset Summary

Dataset used

YouTube Trending Video Dataset (updated daily)

https://www.kaggle.com/rsrishav/youtube-trending-video-dataset

Key Features

- 3 countries selected (US, GB, CA)
- 200 trending videos a day over 1.5 years
- Over 100k videos per country
- 11 variables included

Metadata

Metadata	
video_id	view_count
title	likes
publishedAt	dislikes
channelId	comment_count
channelTitle	comments_disabled
categoryId	ratings_disabled
trending_date	description
tags	thumbnail_link

Data Cleanup & Exploration

Dataset Exploration

title	publishedAt	channelTitle	categoryId	trending_date	tags	view_count	likes	dislikes	comment_count	
I ASKED HER TO BE MY GIRLFRIEND...	2020-08-11T19:20:14Z	Brawadis	22	2020-08-12T00:00:00Z	brawadis|prank|basketball|skits|ghost|funny vi...	1514614	156908	5855	35313	http
Apex Legends | Stories from the Outlands – "Th...	2020-08-11T17:00:10Z	Apex Legends	20	2020-08-12T00:00:00Z	Apex Legends|Apex Legends characters|new Apex ...	2381688	146739	2794	16549	ov
I left youtube for a month and THIS is what ha...	2020-08-11T16:34:06Z	jacksepticeye	24	2020-08-12T00:00:00Z	jacksepticeye|funny|funny meme|memes|jacksepti...	2038853	353787	2628	40221	
XXL 2020 Freshman Class Revealed - Official An...	2020-08-11T16:38:55Z	XXL	10	2020-08-12T00:00:00Z	xxl freshman|xxl freshmen|2020 xxl freshman|20...	496771	23251	1856	7647	Su http
Ultimate DIY Home Movie Theater for The LaBran...	2020-08-11T15:10:05Z	Mr. Kate	26	2020-08-12T00:00:00Z	The LaBrant Family|DIY|Interior Design|Makeove...	1123889	45802	964	2196	

Data Cleanup & Exploration

Data Cleaning

Simple Data Cleaning

```
# Lowercase title and tags columns
df['title'] = df['title'].str.lower()
df['tags'] = df['tags'].str.lower()
df['description'] = df['description'].str.lower()

# Splitting tag and title contents for easier parsing
df['title content'] = df['title'].str.split()
df['tag content'] = df['tags'].str.split("|")
df['description content'] = df['description'].str.split()

# Getting the total word count of video title (title length)
df['total count title'] = df['title'].str.split().str.len()

# Getting the total tag count of video tags (tag length)
df['total count tag'] = df['tags'].str.split("|").str.len()
```

tag content	title content	description content	total count title	total count tag
[brawadis, prank, basketball, skits, ghost, fu...	[i, asked, her, to, be, my, girlfriend...]	[subscribe, to, brawadis, ▶, http://bit.ly/sub...	7	15
[apex legends, apex legends characters, new ap...	[apex, legends, |, stories, from, the, outland...	[while, running, her, own, modding, shop,, ram...	10	25
[jacksepticeye, funny, funny meme, memes, jack...	[i, left, youtube, for, a, month, and, this, i...	[i, left, youtube, for, a, month, and, this, i...	11	30
[xxl freshman, xxl freshmen, 2020 xxl freshman...	[xxl, 2020, freshman, class, revealed, -, offi...	[subscribe, to, xxl, →, http://bit.ly/subscrib...	8	23

- Convert all characters to lowercase and split title and tags for easier parsing
- Count splitted title/ tags words into items in a list then get total count

Data Cleanup & Exploration

Data Cleaning

```
# Clean 'publishedAt' column
# Remove the dates, mins, and seconds in 'publshedAt' column
df['publishedAt'] = df['publishedAt'].str[10:]
df['publishedAt'] = df['publishedAt'].str[:3]

# Divide into three countries
df_us = df[df['country'] == 'US']
df_gb = df[df['country'] == 'GB']
df_ca = df[df['country'] == 'CA']

# Create dictionaries of published time counts
time_counts_us = df_us['publishedAt'].value_counts().to_dict()
time_counts_gb = df_gb['publishedAt'].value_counts().to_dict()
time_counts_ca = df_ca['publishedAt'].value_counts().to_dict()

# Create dataframes of published time counts
df_ca_time = pd.DataFrame(list(time_counts_ca.items()),
                          columns = ['time','count']).sort_values(by=['time'])
df_gb_time = pd.DataFrame(list(time_counts_gb.items()),
                          columns = ['time','count']).sort_values(by=['time'])
df_us_time = pd.DataFrame(list(time_counts_us.items()),
                          columns = ['time','count']).sort_values(by=['time'])
```

```
2020-08-11T19:20:14Z
2020-08-11T17:00:10Z
2020-08-11T16:34:06Z
2020-08-11T16:38:55Z
2020-08-11T15:10:05Z
2020-08-11T20:00:04Z
2020-08-12T00:17:41Z
2020-08-11T17:15:11Z
2020-08-10T22:26:59Z
2020-08-11T23:00:10Z
2020-08-11T20:24:34Z
2020-08-11T17:00:31Z
2020-08-11T17:13:53Z
2020-08-11T19:00:10Z
2020-08-10T22:33:48Z
2020-08-11T22:00:05Z
2020-08-10T18:41:19Z
2020-08-11T23:00:06Z
2020-08-11T12:04:40Z
2020-08-11T15:00:13Z
```

	time	count
11	T00	1806
14	T01	1495
17	T02	1273
16	T03	1297
12	T04	1802
13	T05	1560
22	T06	588
23	T07	527
21	T08	761
18	T09	1137
20	T10	865
19	T11	1024
15	T12	1425
9	T13	2297
7	T14	2818
2	T15	4207

- Convert ‘publishedAt’ and ‘categoryId’ columns into simple readable format
- Create dataframes out of counts afterwards

Data Cleanup & Exploration

Aa Textual Analysis with NLTK Libraries

Textual Analysis

```
##Extracting hot topics with NLTK
text = df_us_20['title'].str.lower().replace('|', ' ').str.cat(sep=' ')


stop_words = set(stopwords.words('english'))

word_tokens = word_tokenize(text)

filtered_sentence = []

for w in word_tokens:
    if w not in stop_words:
        filtered_sentence.append(w)
```

```
# Stemming with NLTK
Stem_words = []
ps =PorterStemmer()
for w in filtered_sentence:
    rootWord=ps.stem(w)
    Stem_words.append(rootWord)
```

```
# remove unnecessay words
stopwords = ["'s", "?", "..." , "ft." , "2" ,"x" , "1", "n't", "-", "3", "5", "4",
             "2021","2020","trailer", "de", "official", "season", "video", "official", "season",
             "episode","la", "le", "je", "part", "je", "des","world","day", "10","e", "avec", "'",
             "à", "music", "none", "new","lil", "like", "songs", "song","thee","love","bad","g",
             "mix", "100","6"]

for word in list(filtered_sentence):  # iterating on a copy since removing will mess things up
    if word in stopwords:
        filtered_sentence.remove(word)
```

- **Tokenization**
  : is the process of breaking up a given text into units called tokens.

- **Stop words**
  :are a set of commonly used words in any language.

- **Stemming**
  :is the process of reducing inflected words to their word stem, base or root form—generally a written word form

Data Cleanup & Exploration

Aa Textual Analysis – World Cloud

WordCloud

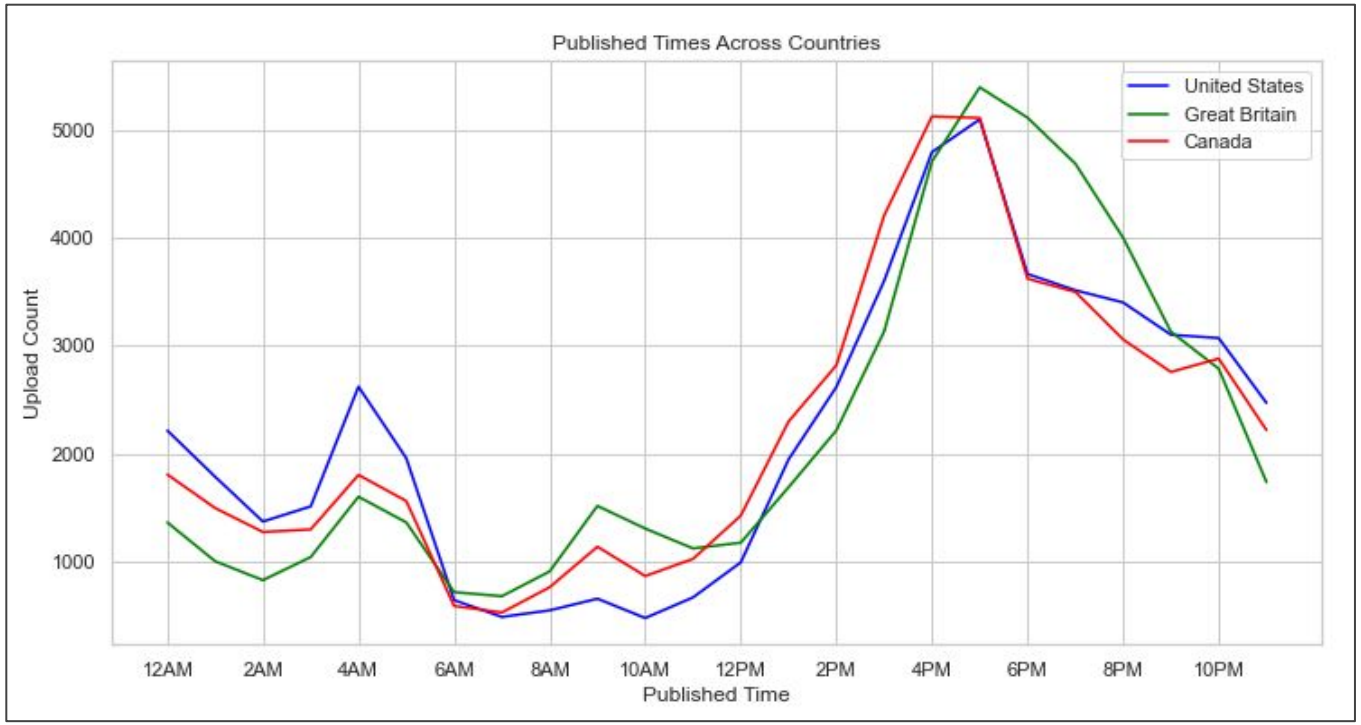
```
#wordcloud
word_could_dict= Counter(filtered_sentence)

wordcloud = WordCloud(width = 1000, height = 500, background_color ='black',
                      stopwords = stopwords,
                      min_font_size = 10).generate_from_frequencies(word_could_dict)


plt.figure(figsize=(15,8))
plt.imshow(wordcloud)
plt.axis("off")
# plt.show()
plt.savefig('us20_title_wordcloud.png', bbox_inches='tight')
plt.close()
```

Data Analysis

Published Times Across Countries



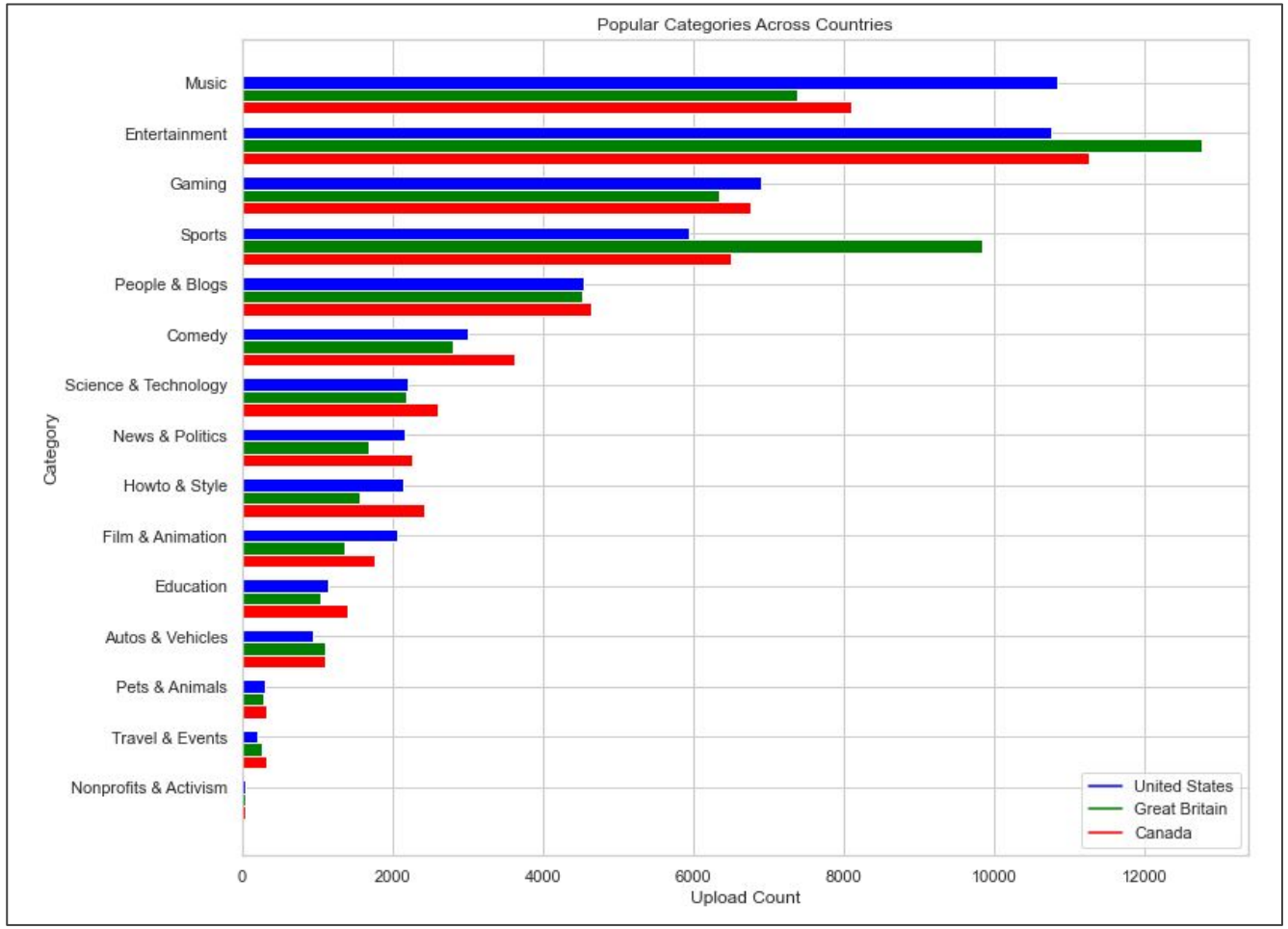
- Similar published times across all countries
- Most videos published between 3:00 PM - 7:00 PM

Note

slight peak at 4:00 AM

Data Analysis

Published Categories Across Countries



- Similar category rank across all countries

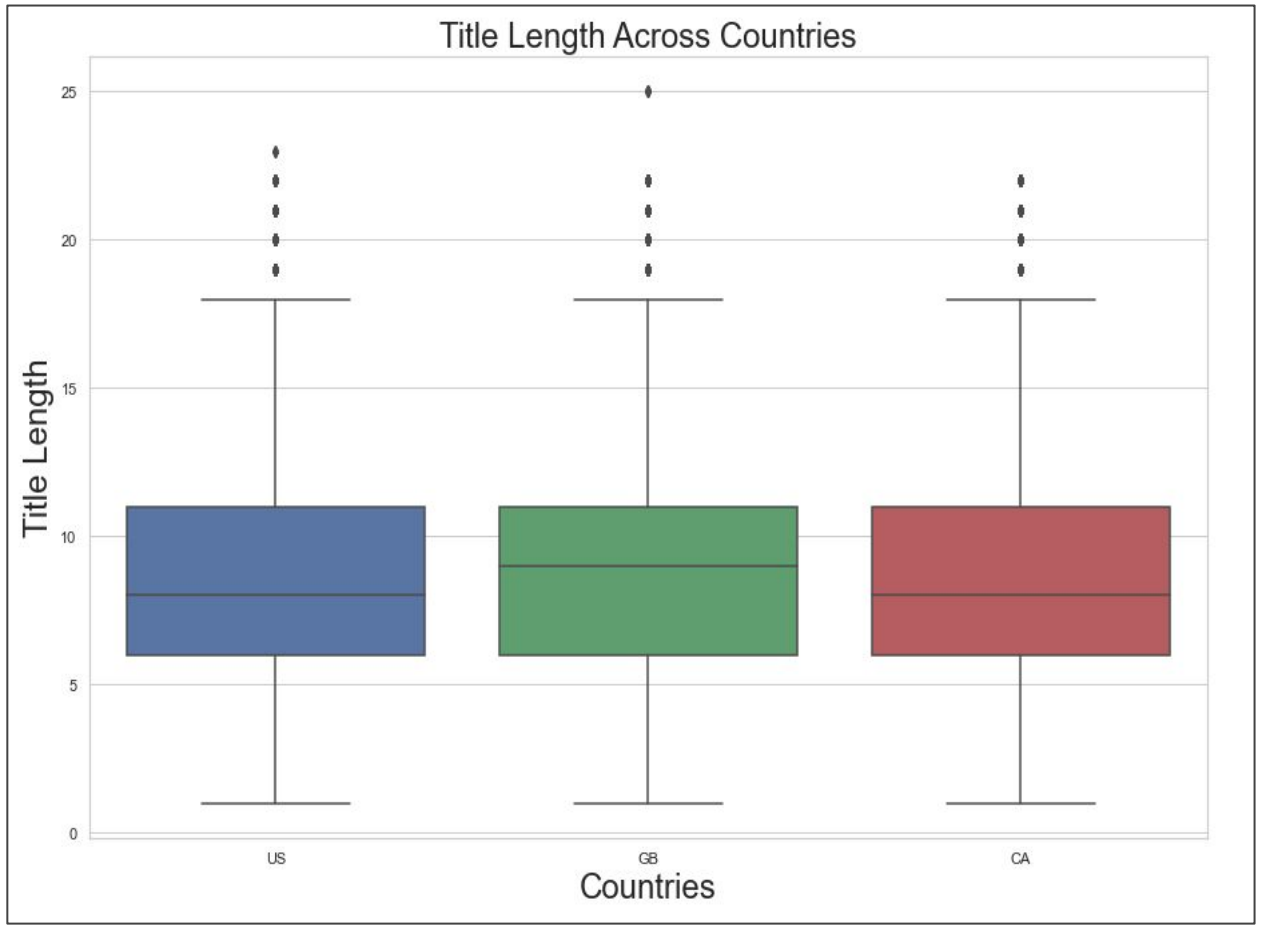
Note

Great Britain shows Sports category as second highest in popularity

- Music, Entertainment, Gaming, and Sports makes up over 50% of all video categories

Data Analysis

Title Length Across Countries



Lower Quartile	6.0
Upper Quartile	11.0
Inter Quartile	5.0

Median of title length		
US	GB	CA
8.0	9.0	8.0

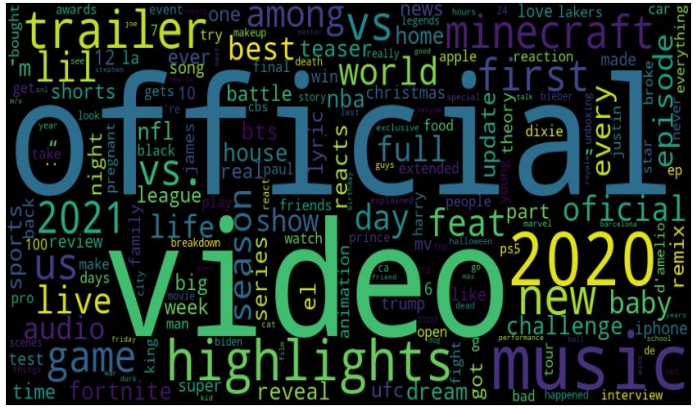
Note

Values below **-1.5** above **18.5** could be outliers

Data Analysis

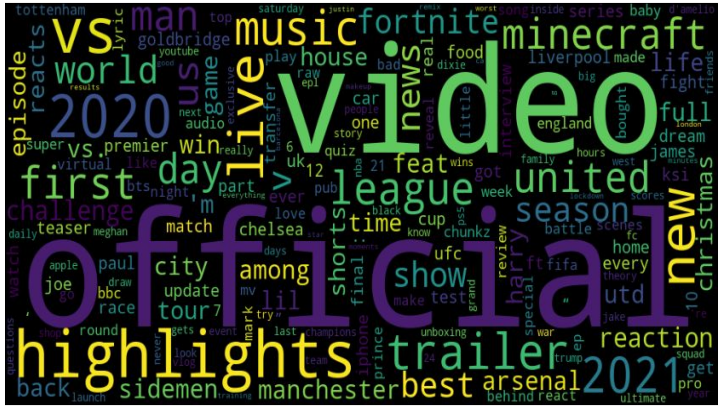
YouTube Trending Video Title Key words

US



US	
official	7241
video	6891
music	2476
2020	2244
highlights	2175
trailer	1883
new	1796
vs	1559
minecraft	1469
first	1364
2021	1351

GB



GB	
official	4799
video	4078
highlights	2726
vs	2626
2020	2273
new	1907
live	1849
trailer	1599
2021	1572
music	1452
league	1314

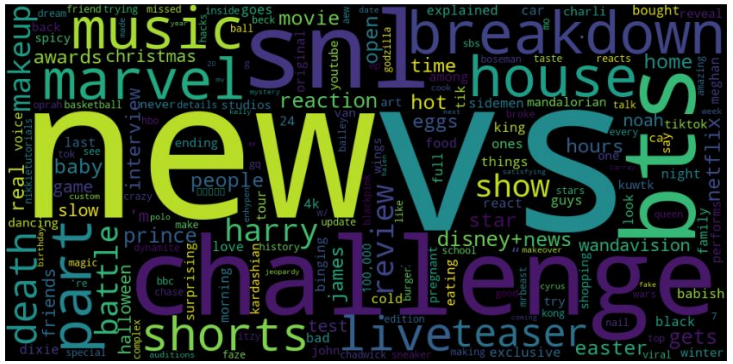
CA

CA	
official	5231
video	4583
highlights	2421
2020	2230
new	1870
music	1725
vs	1724
trailer	1683
2021	1497
minecraft	1447
us	1348

Data Analysis

Entertainment Category - Title

US

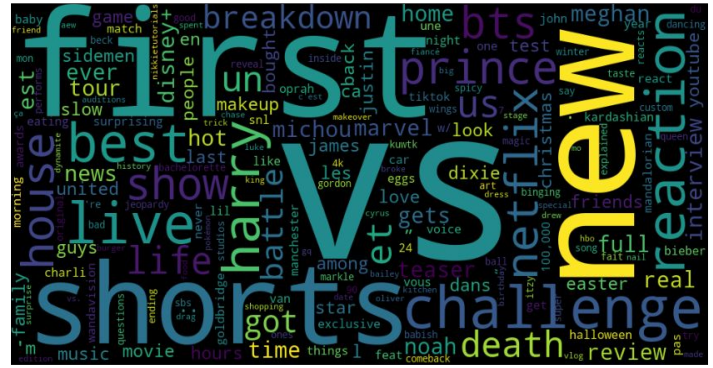


US	
vs	422
new	324
challenge	229
snl	224
bts	201
breakdown	201
music	191
part	188
marvel	187
house	183
towers	170

GB

GB	
live	660
vs	658
united	467
news	415
new	400
goldbridge	387
quiz	367
manchester	363
reaction	348
sidemen	333
pub	332

CA

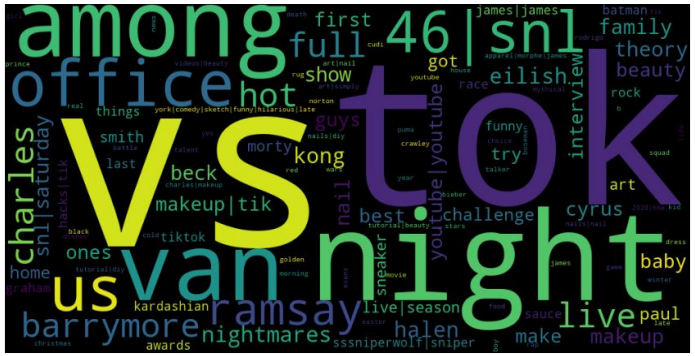


CA	
vs	433
first	306
new	295
shorts	287
challenge	230
live	225
reaction	212
best	197
harry	184
prince	182
netflix	180

Data Analysis

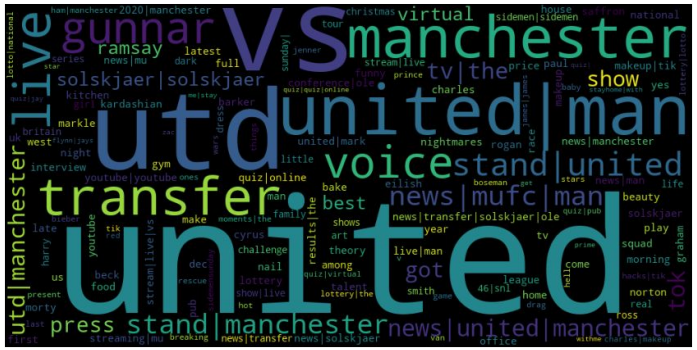
Entertainment Category - Tags

US



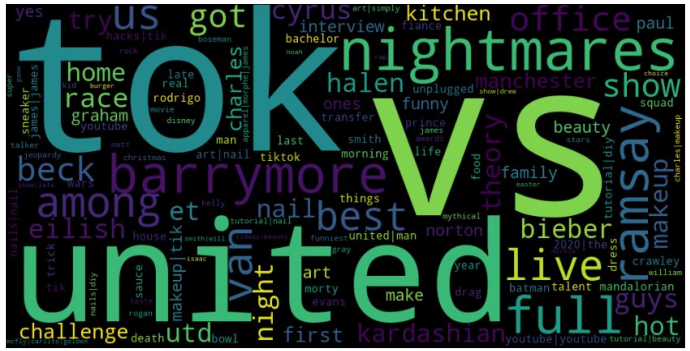
US	
vs	1048
tok	759
night	649
among	446
van	444
office	435
us	412
46|snl	389
ramsay	350
barrymore	343
full	339

GB



GB	
united	3785
vs	1740
utd	1511
united|man	1227
manchester	1195
transfer	932
live	913
voice	667
gunnar	667
stand|united	604
news|mufc|man	592

CA



CA	
vs	816
tok	618
united	563
nightmares	492
barrymore	454
ramsay	436
full	425
among	398
live	392
van	389
us	377

Data Analysis

Entertainment Category - Tags

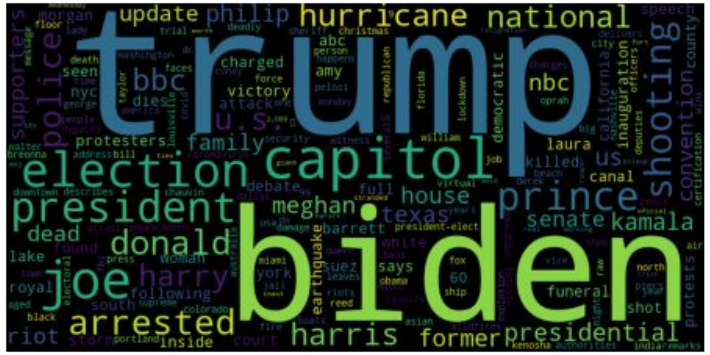
Canada

“Prince Harry Arrives In Canada To Rejoin Meghan Markle And Baby Archie”

Data Analysis

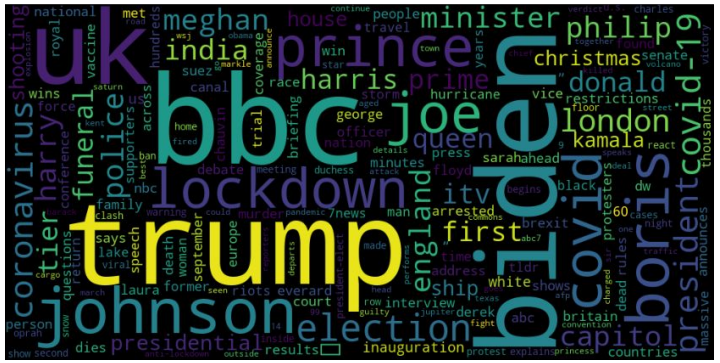
News Category - Title

US



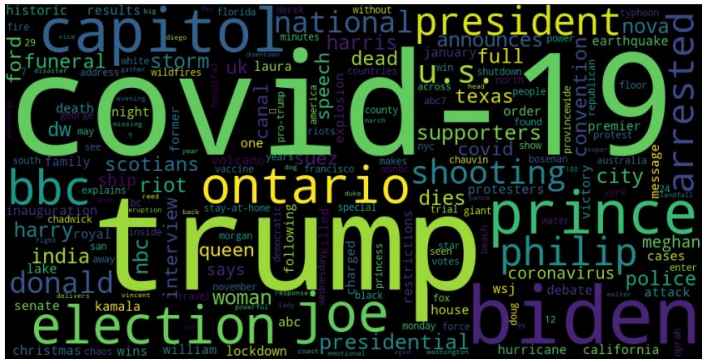
US	
trump	322
biden	234
capitol	184
joe	157
election	131
president	128
prince	112
shooting	99
arrested	89
donald	86
police	75

GB



GB	
bbc	212
trump	190
biden	170
uk	148
johnson	141
boris	126
prince	112
joe	108
lockdown	106
covid	103
election	91

CA

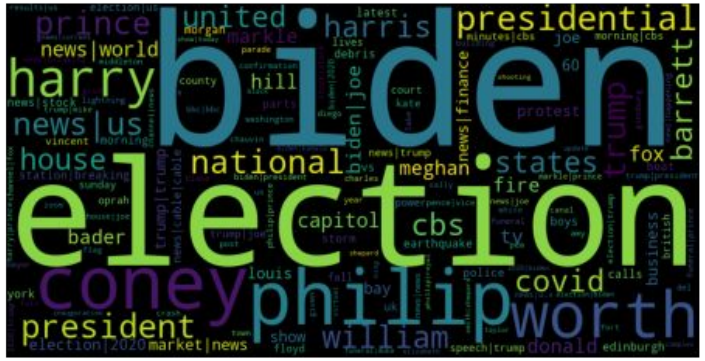


CA	
covid-19	252
trump	225
biden	151
capitol	146
prince	141
joe	118
election	113
ontario	104
bbc	92
president	88
philip	87

Data Analysis

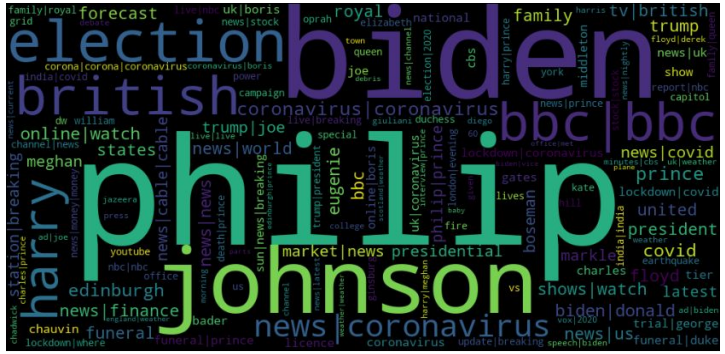
News Category - Tags

US



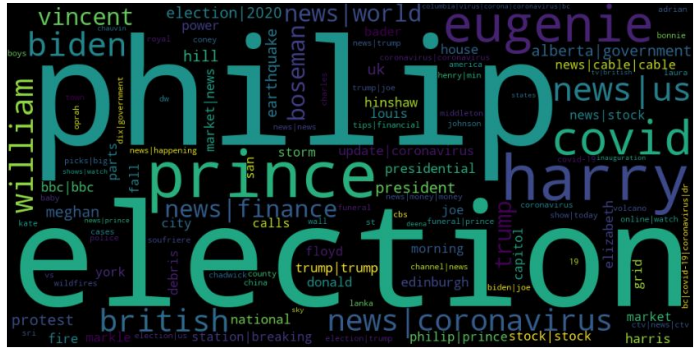
US	
biden	285
election	221
philip	144
coney	135
worth	126
harry	119
presidential	105
trump	98
president	97
national	94
prince	94

GB



GB	
philip	233
biden	153
johnson	141
election	126
british	106
bbc|bbc	102
harry	86
news|coronavirus	85
prince	81
coronavirus|coronavirus	81
covid	74

CA

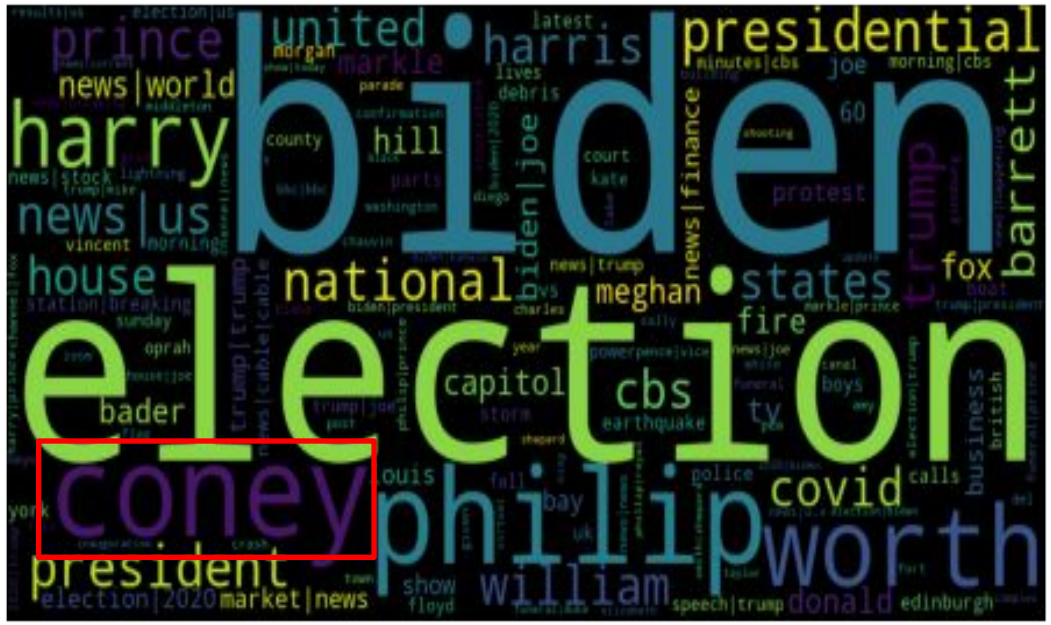


CA	
philip	200
election	194
prince	110
harry	104
eugenie	98
covid	95
news|us	93
biden	91
william	83
british	78
news|coronavirus	74

Data Analysis

News Category - Tags

US



“Video captures fatal shooting inside Detroit Coney Island restaurant”

Conclusion

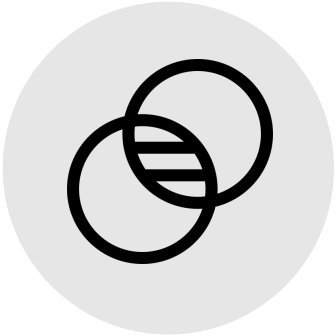
Challenge

- Language barrier
- Size of dataset
- Using single words can bias (N-gram)
- Understanding countries background

Potential Further Research

- Analysis of non-English speaking countries
- Profit analysis of channel, tag, category type and average profit from views
- Description analysis
- Video thumbnail analysis

Conclusion



Similarities in trending videos

(i.g. title length, tag length, publish time, etc.)

Enhance understanding of countries background

65% of people are visual learners-- benefit individual content creators and businesses to enrich video contents

Thank You!

Q&A