HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

PROJECT REPORT
DATA SCIENCE

Topic: YouTube Gaming Trend Analysis

Class code: 136319
Lecturer: Thân Quang Khoát
Group: 16

20200028	Nguyễn Hùng Anh	anh.nh200028@sis.hust.edu.vn
20204909	Nguyễn Trung Hiếu	hieu.nt204909@sis.hust.edu.vn
20204934	Kiều Anh Văn	van.ka204934@sis.hust.edu.vn
20204895	Phan Thái Việt	viet.pt204895@sis.hust.edu.vn
20204935	Đoàn Ngọc Vinh	vinh.dn204935@sis.hust.edu.vn

Hanoi, 2022

CONTENTS

ABSTRACT

Founded in 2005, Youtube has grown to become the second largest search engine in the world (behind Google) that processes more than 3 billion searches per month. In this project we explore the statistics of around 10 most successful Gaming Youtube channels.

Keyword: *youtube*, *gaming*, *trend*.

CHAPTER 1

Introduction

Introduction Social media platforms play a vital role in business, entertainment, marketing, education, media, and communication. YouTube has become the most used platform for sharing videos in society due to its unique behavior. YouTube allows any person to create an account under any category of choosing to upload videos to be viewed by many millions of other people. This has become a trend among the entertainment industry hence it can easily reach to the users and gain popularity for the video materials hosted online. Many YouTube channel keepers are taking different actions to make the video popular.

It is, however, generally a myth how the YouTube algorithm works, what makes a video get views and be recommended over another. In fact, YouTube has one of the largest scale and most sophisticated industrial recommendation systems in existence. For content creators, it is a challenge to understand why a video gets video and others do not. There are many ”myths” around the success of a YouTube video, for example if the video has more likes or comments, or if the video is of a certain duration. It is also worth experimenting and looking for ”trends” in the topics that YouTube channels are covering in a certain niche.

We focus our attention on the analysis of the top 9 gaming channel video data and verify different common ”myths” about what makes a video do well on Youtube, for example the correlation between the number of likes and comments and the number of views, the significance of title length and video length. We also explore the trending topic using NLP techniques for better insights.

Report structure The rest of this report is organized as follows. We give a brief introduction of YouTube in Chapter 2. Chapter 3 is a discussion of our particular problem, the nature of the dataset, data processing and enriching. In Chapter 4, we discuss the results revealed from experiments on the dataset. The model implementation and evalutaion will be considered in Chapter 5. Chapter 6 summarises our conclusion on this project.

CHAPTER 2

Literature review

2.1 ABOUT YOUTUBE

YouTube is a video sharing and social media platform where users can upload, share, and view videos. It was created in 2005 and was later acquired by Google. It is now one of the largest and most popular websites on the internet, allowing users to watch, like, comment and share videos on a variety of topics, including music, entertainment, education, and more. YouTube allows users to create their own channels, where they can upload videos and interact with their audience through comments, likes, and other features. It has also become a platform for influencers and content creators to build their brand, connect with fans, and earn money through advertising and sponsorships. In addition to individual creators, media companies, non-profits, and other organizations also use the platform to reach a large, global audience. YouTube becomes a hugely popular and influential platform, with over 2 billion monthly active users.

2.1.1 YouTube Gaming

YouTube gaming is a section of the YouTube platform that is specifically dedicated to gaming content. It provides a hub for gamers and game enthusiasts to discover and watch videos about their favorite games, as well as connect with other players and content creators in the gaming community. YouTube gaming features a wide range of content, including gameplay footage, game trailers, walkthroughs, reviews, and more. The platform also offers tools for content creators to monetize their gaming content, such as advertising revenue and sponsorships. Additionally, YouTube gaming offers live streaming capabilities, allowing players to broadcast their gameplay and interact with their audience in real-time. With millions of users and billions of views, YouTube gaming has become an important platform for the gaming community and a powerful tool for gamers and content creators to share their passion and connect with others.

CHAPTER 3

Data

3.1 DATA COLLECTING

The dataset used in this project is collected using the YouTube API. The gaming channels are chosen manually based on their popularity, which include the total number of views, likes and subscribers. There are total of 9 channels selected, all of which are english-speaking gaming channels. To collect the channel data, we obtained the channel ID manually from the channels URLs. The data contain the channels name, total subscribers, views and videos as well as their channel playlist ID.

Channels name	Subscribers	Views	Total videos
jacksepticeye	29M	15.8B	5046
Markiplier	34.2M	19.4B	5382
VanossGaming	25.8M	15.2B	1662
W2S	16.3M	4.7B	653
Ali-A	18M	5.8B	4001
H2ODelirious	13.3M	4.2B	3252
Syndicate	9.7M	2.1B	3526
DanTDM	26.3M	18.9B	3638
PopularMMOs	17.2M	14.5B	4685
Stats as of 7/1/2023			

Table 3.1 Channels statistics

We also obtain the video statistics for all of the channels. The attributes of the data are:

video_id: ID of video

channelTitle: channel name of that video

title: title of the video

description: description of the video

tags: tags of the video

publishedAt: date and time when video was published

viewCount, likeCount, favouriteCount, commentCount: Number of views, likes, favourites, comments of the video respectively

duration: duration of the video

definition: definition of the video, mostly HD

caption: caption of the video

	video_id	channelTitle	title	description	tags	publishedAt	viewCount	likeCount	favouriteCount	commentCount	duration	definition	caption
0	q05xi6CXTVc	W2S	MY BRO GETS A 193 FUT DRAFT WORLD RECORD - FIF...	FIFA 20 NEW WORLD RECORD FUT DRAFT \nMy Insta:...	[Football, Soccer, football challenge, W2S, wr...	2020-02-19T20:05:52Z	11853432	284006	None	13180	PT14M24S	hd	false
1	xEq8aKDo2DQ	W2S	CRISTIANO RONALDO vs LIONEL MESSI FOOTBALL CHA...	LIONEL MESSI vs CRISTIANO RONALDO FOOTBALL CHA...	[Football, Soccer, football challenge, soccer ...	2020-02-15T18:06:57Z	6050075	184674	None	4865	PT14M43S	hd	false
2	SLgTKVVnU1w	W2S	YOU'LL NEVER SEE A BETTER TOTY PACK OPENING - ...	FIFA 20 TEAM OF THE YEAR GREATEST EVER PACK OP...	[Football, Soccer, football challenge, W2S, wr...	2020-01-21T22:11:25Z	19848347	398562	None	13003	PT14M26S	hd	false
3	M6KhB1XaKAY	W2S	WORLD RECORD TOTY 192 FUT DRAFT!! - FIFA 20	FIFA 20 TEAM OF THE YEAR FUT DRAFT NEW RECORD\...	[Football, football challenge, W2S, wroetoshaw...	2020-01-10T19:24:34Z	7210231	216804	None	6640	PT15M59S	hd	false
4	OYszZYwTfSQ	W2S	FIRST TO SCORE WINS £50,000 CAR CHALLENGE	W2S FAMILY OLYMPICS CHALLENGE \nBOOK A KICKTO...	[Football, Soccer, football challenge, soccer ...	2019-12-27T20:02:24Z	6068660	229532	None	6466	PT13M11S	hd	false

Figure 3.1 Video data sample

3.2 DATA PREPROCESSING

A few preprocessing steps are needed before we could start on the analysis.

Firstly, we checked for empty values for non-null attribute.

Next, we reformatted the date and time columns (”publishedAt” and ”duration”)

Finally, we checked the data type of the columns. Some count columns such as view count and comment count are not in correct data type(string). In this step, we convert these count columns into integer.

3.3 DATA ENRICHING

In addition to the preprocessing step, it is also necessary to enrich the data with some new features that might be useful for understanding the videos' characteristics, which include: + Create published date column with another column showing the day in the week the video was published, which will be useful for later analysis

+ Convert video duration to seconds instead of the current default string format

+ Calculate number of tags for each video

+ Calculate comments and likes per 1000 view ratio

+ Calculate title character length

pushblishDayName	durationSecs	tagsCount	likeRatio	commentRatio	titleLength
Saturday	1385.0	1	73.673405	2.855095	17
Friday	2682.0	1	59.551585	1.967364	21
Thursday	997.0	1	68.400276	3.470858	35
Wednesday	8761.0	1	40.230914	2.763468	18
Tuesday	2223.0	1	60.219038	3.189578	35

Figure 3.2 New video attributes

CHAPTER 4

Exploratory analysis

4.1 RANKING OF CHANNEL IN SCOPE

4.1.1 Subscribers

The average subscribers of the 9 channels in scope are 21.1 million. 8 out of 9 channels have over 10 million subscribers, while Syndicate being the least subscribed with 9.7 million . The most subscribed channel are Markiplier with over 34.2 million subs.

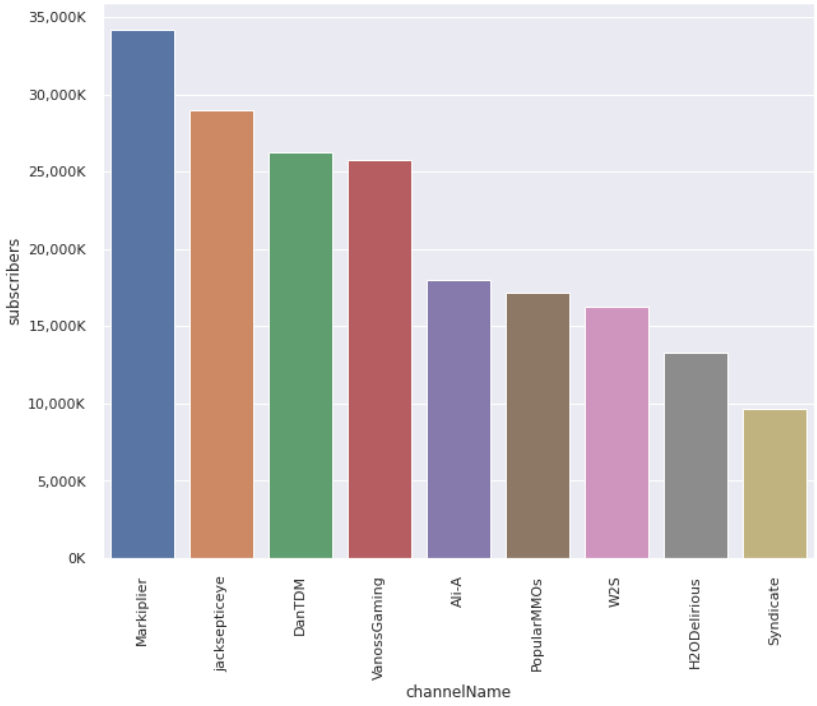

Figure 4.1 Total subscribers of channels

4.1.2 Total views

The rank is fairly similar to the subscriber count rank. Markiplier, DanTDM and jackseptieye remain the three most popular channels considering both subscribers and views. Interestingly, some channels have more subscribers but less views and vice versa. For example, PopularMMOs channel has significantly more views than Ali-A channel, but slightly less subscribers in total.

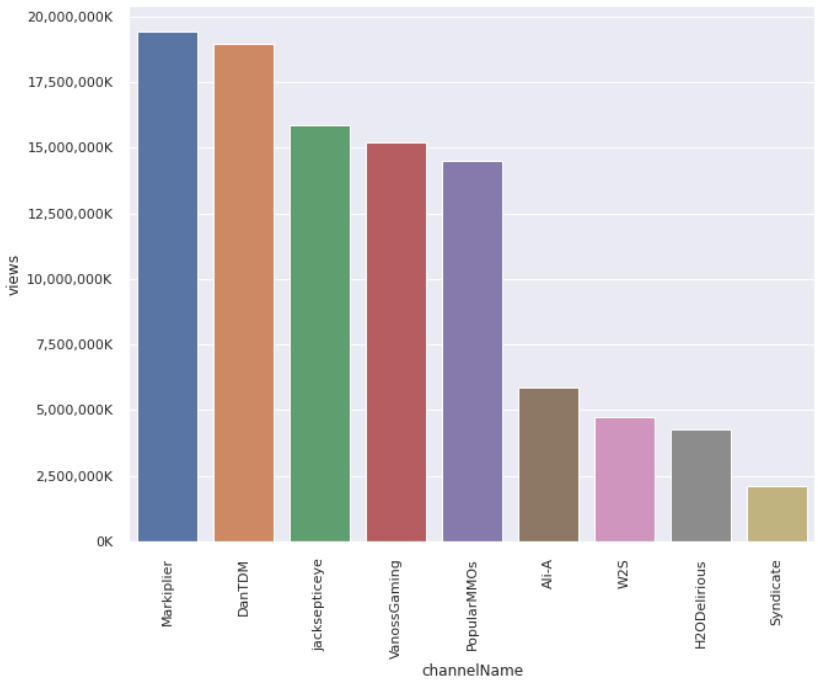

Figure 4.2 Total views of channels

4.2 CORRELATION OF VIDEO STATISTICS AND ITS VIEW

4.2.1 Likes and comments

We analyze the correlation between comments and likes with the number of views received by a video.

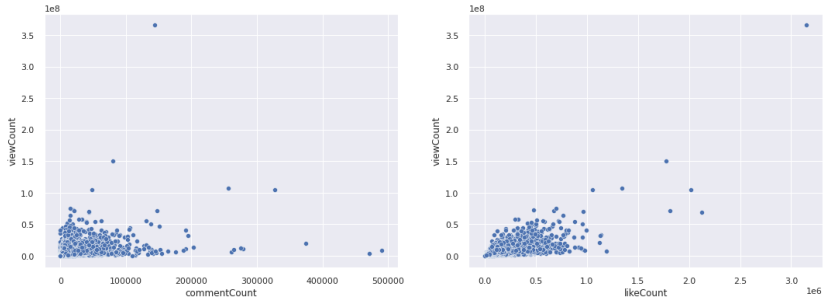

Figure 4.3 Likes and comments by video views

As depicted in the plots, it is evident that there exists a strong connection between the number of views and the number of comments/likes. The number of likes appears to demonstrate a stronger correlation as compared to the number of comments. However, this outcome can be considered predictable as a higher number of views is likely to result in an increased number of comments and likes. In order to account for this factor, the relationships will be re-plotted using the ratios of comments per 1000 views and likes per 1000 views.

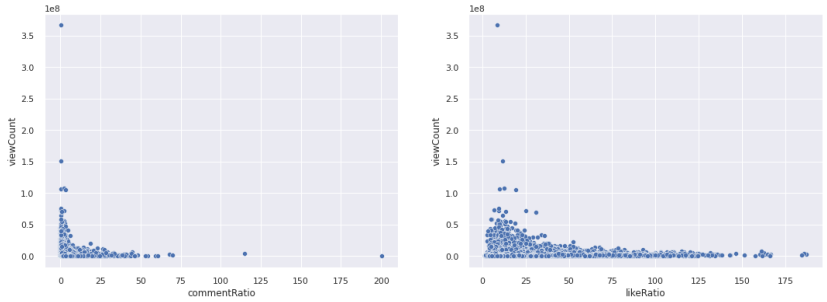

Figure 4.4 Likes and comments per 1000 views by video views

4.2.2 Duration

The length of videos ranged from 300 to 1200 seconds. We remove all videos with length 10,000 seconds or more because of some really long videos (streaming videos).

We plot the duration against **commentCount** and **likeCount**. It can be seen that shorter videos tend to get more likes and comments than longer counterpart.

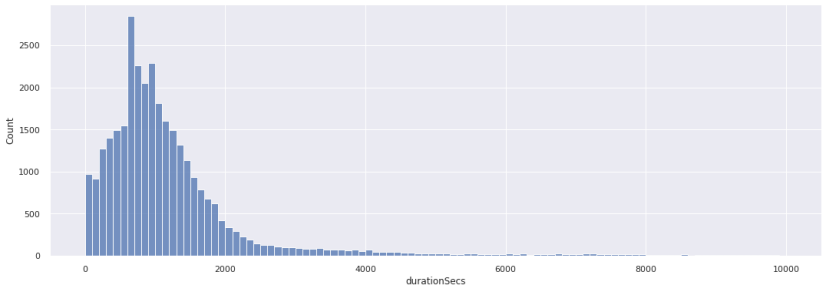

Figure 4.5 Duration of videos

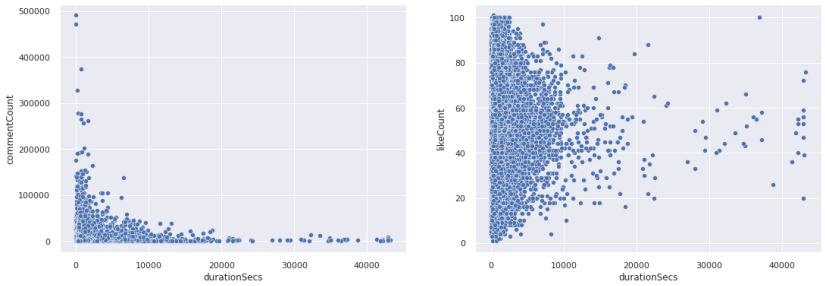

Figure 4.6 Duration by views

4.2.3 Title length

There is no clear relationship between title length and views as seen the scatter plot below, but most-viewed videos tend to have average title length of 30-70 characters.

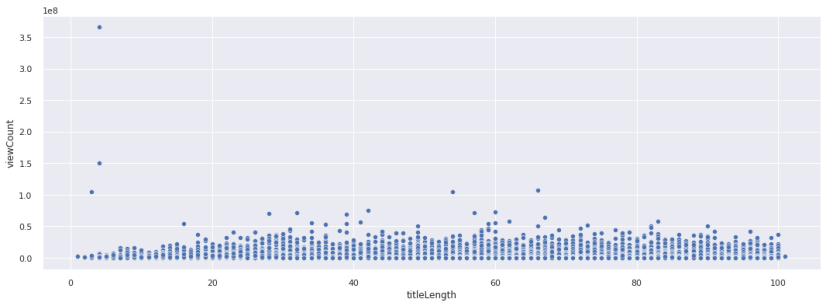

Figure 4.7 Title length by views

4.3 WORDCLOUD FOR WORD IN TITLE

The title of a video are an important statistics which tells us what is the context of the video. We analyze the frequency of terms in video titles using WordCloud. First, we need to remove the stopwords such as ”you”, ”I”, ”the”, ”etc”... which do not contribute to the meaning of the title. It can be seen that most common words are ”Minecraft”, ”Game”, ”Call”, ”Funny”, ”Zombie”...

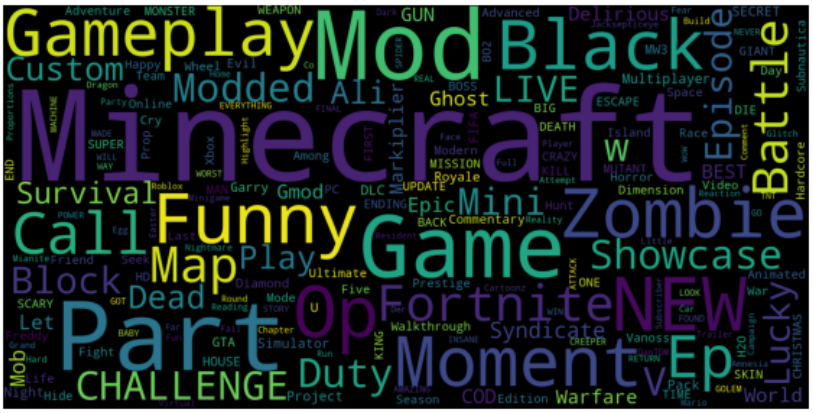

Figure 4.8 WordCloud for word in Title

4.4 NUMBER OF TAGS

Most-viewed videos seems to have 0-35 tags.

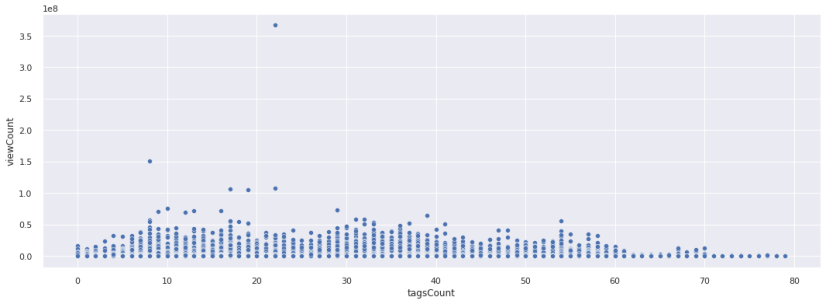

Figure 4.9 Total tags per video

4.5 WEEKDAYS

Tuesday seems to be the highest published day, but overall videos are uploaded equally for everyday.

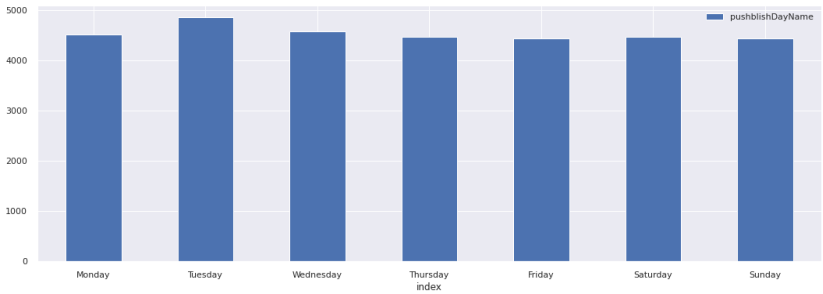

Figure 4.10 Total video uploaded by weekdays

4.6 FUTURE RESEARCH

To further develop the research project, several steps can be taken. Firstly, the dataset can be expanded to include smaller channels within the data science scope. This will provide a more comprehensive understanding of the landscape. Secondly, sentiment analysis can be performed on the comments section to determine which videos receive more positive comments and which videos receive less positive feedback. Additionally, market research can be conducted by analyzing the questions asked in the comment threads, this will help identify common questions and potential market gaps that could be filled. Lastly, the research can be conducted for other niches such as vlogs or beauty channels to compare the patterns in viewership and video characteristics among different niches. This will provide valuable insights into the nuances and differences between niches.

CHAPTER 5

Model

5.1 TF-IDF VECTORIZATION

TF-IDF is a widely adopted method in the field of Natural Language Processing for representing the significance of words in a document. The technique calculates the Term Frequency (TF) of a word, which represents the number of occurrences of the word in a document, and the Inverse Document Frequency (IDF), which reflects the rarity of the word in the entire corpus. The product of these two values, the TF-IDF, highlights the words that are highly specific to a particular document while deemphasizing the words that are common across multiple documents.

In our analysis, we utilize the TF-IDF vectorization technique to convert all text attributes into numerical representations.

5.2 K-MEANS CLUSTERING

The K-means clustering algorithm is employed to categorize videos into eight distinct groups using the following parameters: initialization method set to 'k-means++', number of initializations warning set to 'warn', maximum number of iterations set to 300, tolerance level set to 0.0001, verbosity level set to 0, random state set to None, copying of input data set to True, and algorithm employed set to 'Lloyd'.

5.3 CLUSTER ANALYSIS

5.3.1 Cluster 1

Channels name	Videos
Markiplier	5153
jacksepticeye	4830
H2ODelirious	2566
DanTDM	1548
Total	17821

Table 5.1 Cluster 1 videos by channel

This cluster is too large to have a reasonable analysis, we need to split it into smaller clusters to analyse.

Figure 5.1 Cluster 1 WordCloud

5.3.2 Cluster 1.1

Channels name	Videos
jacksepticeye	3255
Markiplier	3238
H2ODelirious	2005
DanTDM	1142
Total	12252

Table 5.2 Cluster 1.1 videos by channel

Figure 5.2 Cluster 1.1 WordCloud

This cluster encompasses not only video games but also a variety of video genres such as 'lifestyle' and 'Q&A'. There are a total of over 12 thousands video in this cluster, more than any other cluster, with an average of 2.9 million views.

5.3.3 Cluster 1.2

Channels name	Videos
jacksepticeye	646
Markiplier	542
H2ODelirious	281
DanTDM	125
Total	1774

Table 5.3 Cluster 1.2 videos by channel

Figure 5.3 Cluster 1.2 WordCloud

The cluster features videos of popular games like 'Freddy', 'Redemption', and 'Simulator' from well-known channels such as 'Jacksepticeye' and 'Markiplier'. Most of the videos have a view count of less than 4 million, with an average of 2.8 million views.

5.3.4 Cluster 1.3

Channels name	Videos
jacksepticeye	513
Markiplier	463
H2ODelirious	151
DanTDM	80
Total	1325

Table 5.4 Cluster 1.3 videos by channel

3

Figure 5.4 Cluster 1.3 WordCloud

The cluster primarily comprises videos about horror games from various channels, with a focus on 'Jacksepticeye' and 'Markiplier' channels. It seems that scary-themed games videos are relatively popular.

There are total of 1325 videos, averaging 2,9 million views.

5.3.5 Cluster 1.4

Channels name	Videos
Markiplier	910
jacksepticeye	416
W2S	413
DanTDM	201
Total	2466

Table 5.5 Cluster 1.4 videos by channel

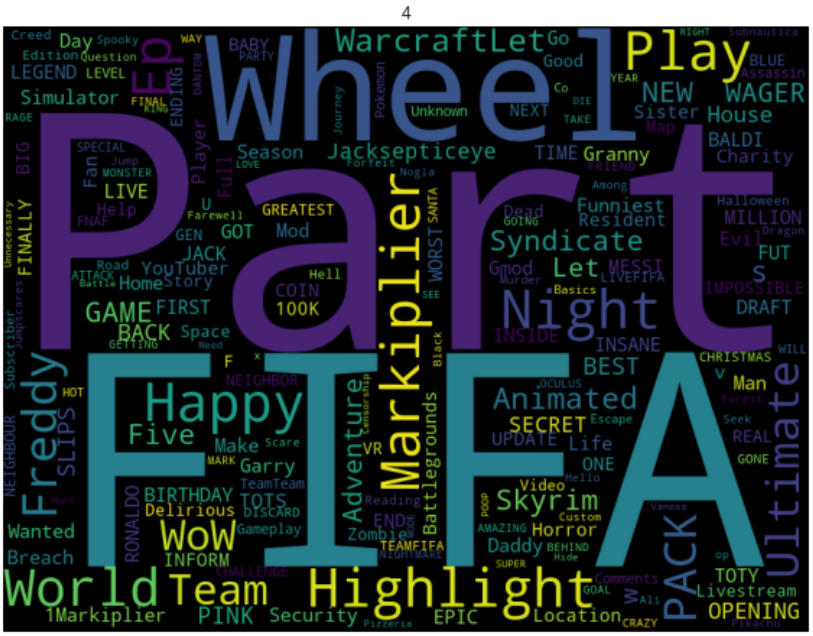

Figure 5.5 Cluster 1.4 WordCloud

This cluster primarily showcases videos of the popular soccer games ”FIFA” and ”Happy Wheels,” from the well-known channels ”Jacksepticeye,” ”Markiplier,” and ”W2S.” Most of the videos have a view count of less than 6 million, with an average of 4.1 million views.

5.3.6 Cluster 2

Table 5.6 Cluster 2 videos by channel

Channels name	Videos
Ali-A	870
Syndicate	785
H2ODelirious	49
VanossGaming	48
Total	1753

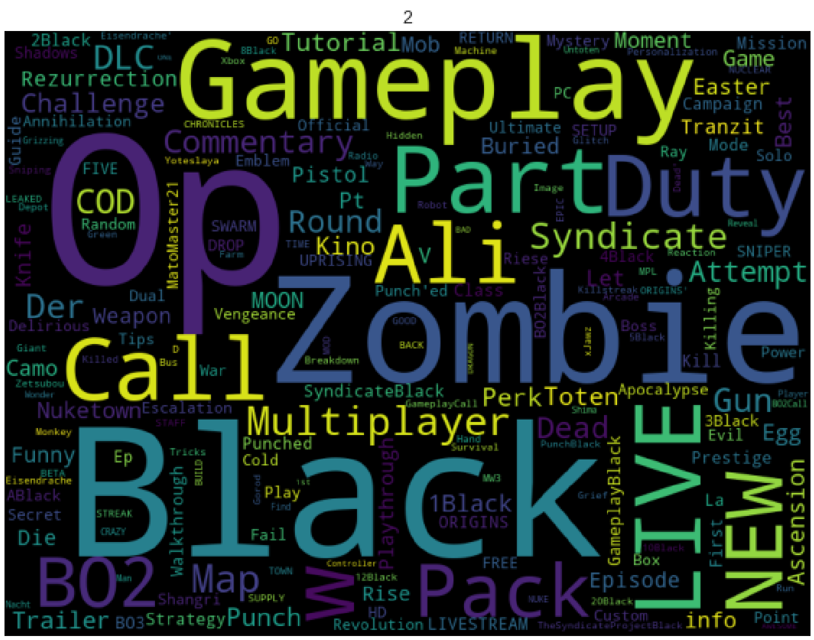

Figure 5.6 Cluster 2 WordCloud

This cluster contains videos about ’Black ops’, zombies mode in ’Black ops’, ’Call of duty’ from ’Ali-A’ and ’Syndicate’ channel. Most of videos have less than 2 million views and the mean view count is 1.2 millions.

5.3.7 Cluster 3

Channels name	Videos
PopularMMOs	698
DanTDM	5
H20Delirious	4
Ali-A	1
Total	708

Table 5.7 Cluster 3 videos by channel

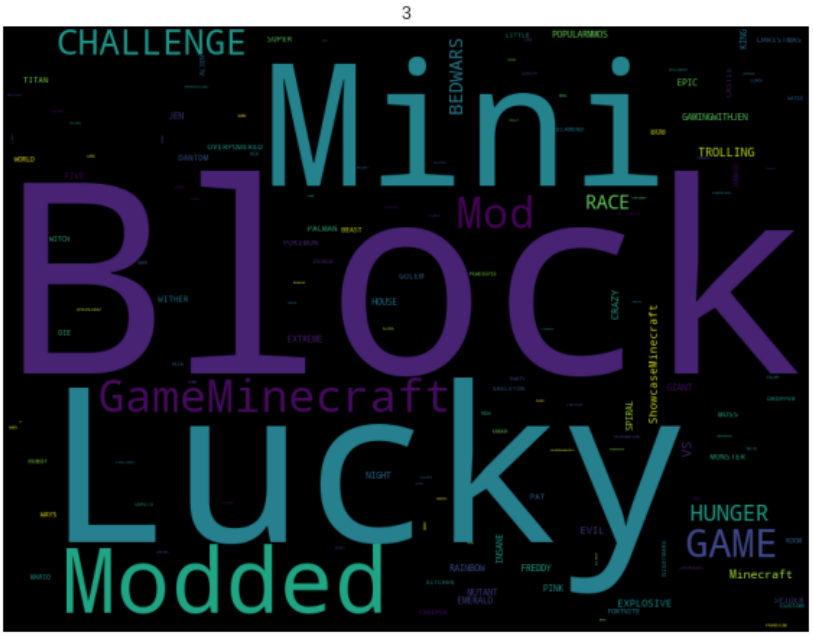

Figure 5.7 Cluster 3 WordCloud

Cluster 3 has 708 instances. The cluster is mostly from channel 'PopularMMOs', thus from the WordCloud above, this cluster is mostly about Minecraft trends which contain 'lucky block' or 'lucky' or 'modded'. The average views of videos in this cluster is slightly above 5 millions.

5.3.8 Cluster 4

Channels name	Videos
Syndicate	453
PopularMMOs	249
H20Delirious	133
jacksepticeye	94
Total	1042

Table 5.8 Cluster 4 videos by channel

4

Figure 5.8 Cluster 4 WordCloud

This cluster has videos mostly from 'VanossGaming' and 'PopularMMOs'. Videos are mostly on a wide variety of 'Minecraft'-related themes and are series-like. Cluster 2 has 1042 videos and the mean value of view is over 1.2 miliions.

5.3.9 Cluster 5

Channels name	Videos
Ali-A	990
Syndicate	124
H20Delirious	6
VanossGaming	3
Total	1124

Table 5.9 Cluster 5 videos by channel

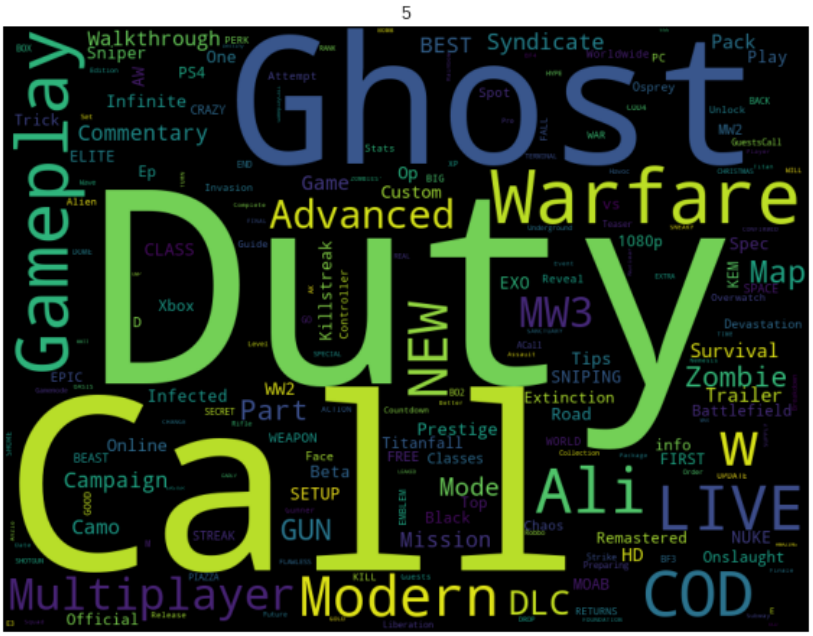

Figure 5.9 Cluster 5 WordCloud

This cluster contains videos about 'Call of Duty' series (a series of first-person shooting video games), with a focus on its zombie mode from 'Black ops 2'. The cluster has been sourced from two prominent channels, 'Ali-A' and 'Syndicate', and analyzed based on the number of views. The results indicate that the majority of the videos have a view count of less than one million, with a mean view count of 900 thousands.

5.3.10 Cluster 6

Channels name	Videos
PopularMMOs	3153
DanTDM	1898
Syndicate	808
Markiplier	95
Total	6124

Table 5.10 Cluster 6 videos by channel

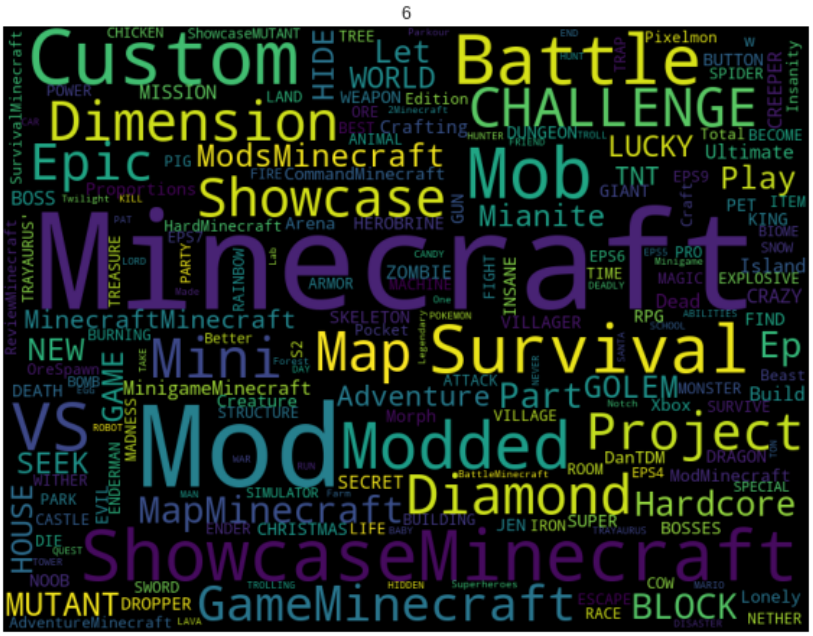

Figure 5.10 Cluster 6 WordCloud

This cluster features videos about 'Minecraft', primarily from 'PopularMMOs', 'Syndicate', and 'DanTDM' channels. The majority of these videos boast less than 4 million views, with an average view count of 3.2 million, a higher figure compared to the similar Cluster 4.

5.3.11 Cluster 7

Channels name	Videos
VanossGaming	1272
H2ODelirious	377
jacksepticeye	48
Markiplier	37
Total	1782

Table 5.11 Cluster 7 videos by channel

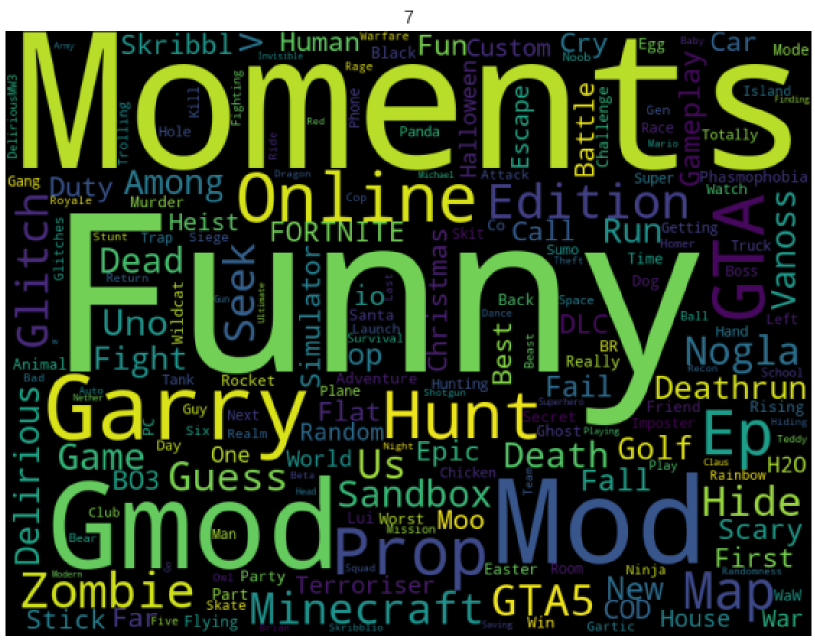

Figure 5.11 Cluster 7 WordCloud

This cluster contains videos about 'Fortnite', 'Gmod', 'Gta',... mostly from 'VanossGaming' and 'H2ODelirious' channels. Moreover, it seems like there is an abundance of compilations of 'moments', especially funny moments. There are a total of 1782 videos with an average view of 7.1 millions.

5.3.12 Cluster 8

Channels name	Videos
Ali-A	1117
H20Delirious	113
DanTDM	92
PopularMMOs	31
Total	1396

Table 5.12 Cluster 8 videos by channel

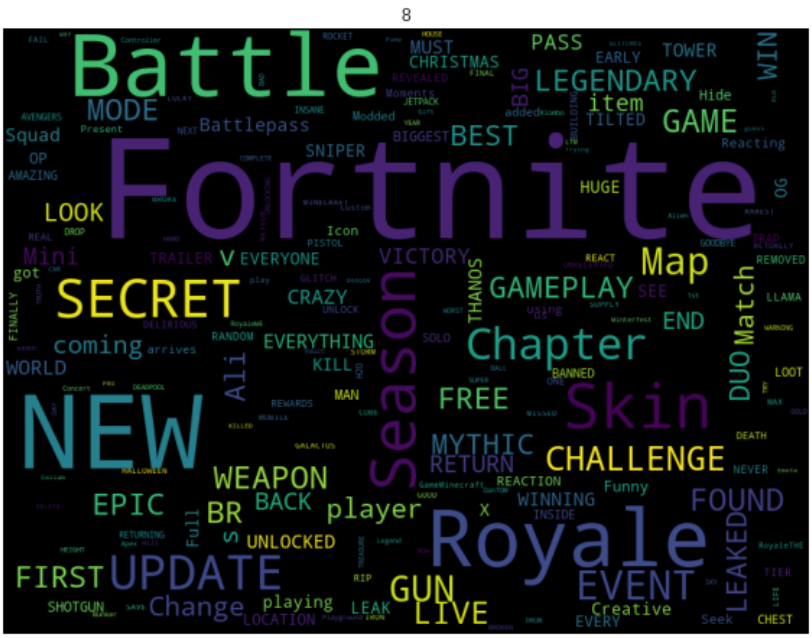

Figure 5.12 Cluster 8 WordCloud

This cluster comprises video content related to a widely popular first-person shooter game, 'Fortnite', specifically from the 'Ali-A 'channels. Most of videos have less than 5 million views and the mean view count is 2.4 millions

CHAPTER 6

Conclusion

In this project, we have explored the video data of 9 of the most popular gaming channels and obtained some insights. Higher views often means higher comments and likes though it is not guarantee. Likes seem to be a better indicator for interaction than comments and the number of likes seem to follow the ”social proof”. Most-viewed videos tend to have average title length of 30-70 characters. Too short or too long titles seem to harm viewership. Videos are usually uploaded equally everyday.

Our clustering analysis reveals several trending topics in gaming videos. ”Minecraft” and ”Fortnite” seem to be the most popular games, with numerous sub-topics such as ”mod,” ”gmod,” ”showcase,” and ”battle.” Shooting game channels like ”Syndicate” and ”Ali-A” consistently receive stable views. Additionally, viewers seem to enjoy watching role-playing and horror game videos, like ”Freddy”. YouTubers are also branching out and creating content outside of gaming, such as ”Q&A” and ”lifestyle” videos. Some videos are part of a series, with labels like ”part 1” and ”part 2.

The findings should also be taken with a grain of salt for a number of reasons. Even with over 30000 videos, the clustering part seems to be unstable and vague. There are many other factors that haven’t been taken into the analysis, including the marketing strategy of the creators and many random effects that would affect how successful a video is.

LIST OF FIGURES

LIST OF TABLES