# BÁO CÁO ĐÔ ÁN

Phân tích dữ liệu Youtube Trending để một Video trở thành xu hướng

Thành viên báo cáo: Nguyễn Anh Nhật Thiềm Ánh0Tường Vy Nguyễn Quốc Triệu

# NỘI DUNG

Nội dung báo cáo gồm 4 chương:











# 1.1

#### Lý do chọn đề tài



#### 1.1.1. Bối cảnh ngành truyền thông hiện nay

#### Tính đến hết quý 3/2019

Dân số Việt Nam đạt hơn 96 triệu người

Hơn **64 triệu người** dùng internet – **67%** dân số

Hơn **62 triệu người** có tài khoản mạng xã hội

– chiếm 97% người dùng internet và 62%

tổng dân số



# 1.1 Lý do chọn đề tài



1.1.1. Bối cảnh ngành truyền thông hiện nay



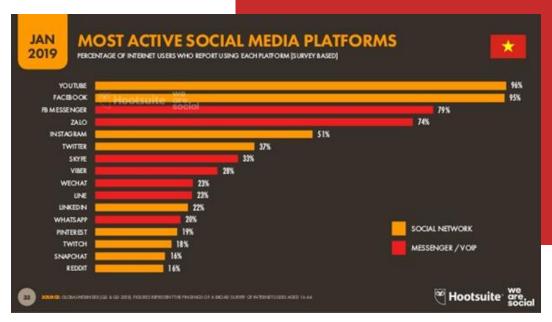
Và Youtube là một trong số đó

### Lý do chọn đề tài



#### 1.1.2. Tiềm năng của Youtube

- Là **một trang mạng xã hội chuyên chia sẻ video trực tuyến**, người xem có thể tương tác trong cuộc trò chuyện hai chiều.
- Một số những yếu tố tạo nên tính cạnh tranh của Youtube:



Lượng người tham gia lâu năm và trung thành

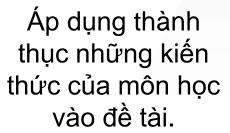
Nơi doanh nghiệp lựa chọn để đưa sản phẩm giới thiệu đến khách hàng tiềm năng

Nền tảng công nghệ thông tin hoạt động ổn định và được cải tiến liên tục

## Mục tiêu đề tài









Bộ dữ liệu đầy đủ, đáng tin cậy từ các nguồn uy tín.



Đào được dữ liệu từ YouTube về để tiến hành phân tích, báo cáo



Phát triển đề tài có tính thực tiễn

#### Phạm vi nghiên cứu





Phạm vi nghiên cứu:

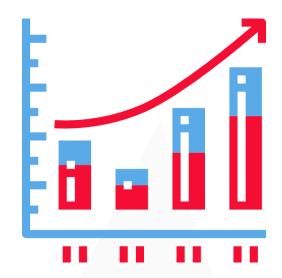
YouTube API; Bộ các thư viện; Các phần mềm hỗ trợ



Đối tượng nghiên cứu: Video trong top trending trên YouTube

## Phương pháp nghiên cứu





Phương pháp thống kê, so sánh



Phương pháp phân tích, tổng hợp

# 1.5 Ý nghĩa đề tài





Đối với người dùng



Đối với cá nhân thành viên nhóm



#### Ngôn ngữ lập trình



#### 2.1.1. Python

a) Định nghĩa



- Python là một ngôn ngữ lập trình cấp cao, hướng đối tượng,
   được giải thích với ngữ nghĩa động.
- Phù hợp cho việc phát triển ứng dụng nhanh
- Cú pháp đơn giản, dễ học nhấn mạnh khả năng đọc và do đó giảm chi phí bảo trì chương trình

## Ngôn ngữ lập trình



#### 2.1.1. Python

b) Tính năng chính của Python



Ngôn ngữ lập trình đơn giản



Tính năng mở rộng Ngôn ngữ di động



Mã nguồn mở và miễn phí





Ngôn ngữ định hướng





Hỗ trợ lập trình GUI



Ngôn ngữ cấp cao



Ngôn ngữ tích hợp Ngôn ngữ diễn giải



Thư viện tiêu chuẩn lớn

#### Thư viện Python



#### 2.2.1. Request

Thư viện xử lý để giúp người dùng không phải thêm các chuỗi truy vấn (query strings) một các thủ công vào các URLs, form-encode hoặc PUT và POST dữ liệu





```
downloads 991M license Apache 2.0 wheel yes python 2.7 | 3.5 | 3.6 | 3.7

Requests is an elegant and simple HTTP library for Python, built for human beings.

Behold, the power of Requests:

>>> r = requests.get('https://api.github.com/user', auth=('user', 'pass'))
>>> r.status_code
```

Release v2.22.0. (Installation)

>>> r.headers['content-type']

>>> r.encoding

'application/json; charset=utf8'



#### 2.2.2. BeautifulSoup

BeautifulSoup là một thư viện giúp bạn dễ dàng lấy thông tin từ các trang web



#### Example #4

```
# print soup.contents
print len(soup.contents) # 1
print soup.contents[0].name # u'html'

print len(list(soup.children)) # 1
print len(list(soup.descendants)) # 25
# -----
print(soup.html.string) # None

#for string in soup.strings:
    # print(repr(string))

for string in soup.stripped_strings:
    print(repr(string))
```

```
1
html
1
25
None

u"The Dormouse's story"
u"The Dormouse's story"
u'Once upon a time there were three little sisters; and the u'Elsie'
u','
u'Lacie'
u'and'
u'Tillie'
u';\nand they lived at the bottom of a well.'
u'...'
```

## Thư viện Python



#### 2.2.3. Pandas

Cung cấp các cấu trúc dữ liệu mạnh mẽ, linh hoạt giúp các thao tác và phân tích dữ liệu dễ dàng hơn. DataFrame là một trong những cấu trúc dữ liệu rất mạnh của Pandas.



	person_ID	name	first	last	middle	email	phone	fax	title	age	is_young	birthday
35	3903	None	Ann	Dunlap	NaN	DunlapA@univ.edu	963.555.9067	963.777.4290	Assistant Professor	25	False	1985
36	3095	None	Rich	Shields	Pena	ShieldsR@univ.edu	963.555.9197	963.777.7215	Professor	25	False	1994
37	2383	None	Winnie	Page	NaN	PageW@univ.edu	963.555.9366	963.777.3202	Curator	25	False	1991
38	2146	None	Ezra	Sparks	NaN	SparksE@univ.edu	963.555.9390	963.777.9273	Assistant Professor	25	False	1996
39	3958	None	Elba	Kaufman	NaN	KaufmanE@univ.edu	963.555.9507	963.777.3298	Professor	25	False	1994



#### 2.2.4. Numpy

Numpy là một core library phục vụ cho khoa học máy tính của Python, hỗ trợ cho việc tính toán các mảng nhiều chiều, có kích thước lớn với các hàm đã được tối ưu áp dụng lên các mảng nhiều chiều đó



_						_
Θ	1	2	3	4	5	
10	11	12	13	14	15	
20	21	22	23	24	25	
30	31	32	33	34	35	
40	41	42	43	44	45	
50	51	52	53	54	55	

#### Thư viện Python



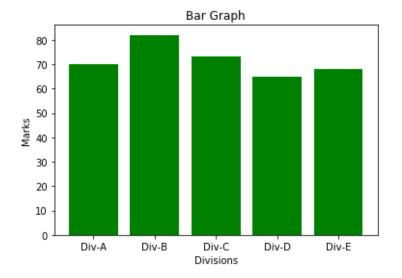
#### 2.2.5. Matplotlib

Để thực hiện các suy luận thống kê cần thiết, cần phải trực quan hóa dữ liệu của bạn và thư viện Matplotlib là một trong những giải pháp như vậy cho lập trình viên Python.



```
divisions = ["Div-A", "Div-B", "Div-C", "Div-D", "Div-E"]
division_average_marks = [70, 82, 73, 65, 68]

plt.bar(divisions, division_average_marks, color='green')
plt.title("Bar Graph")
plt.xlabel("Divisions")
plt.ylabel("Marks")
plt.show()
```

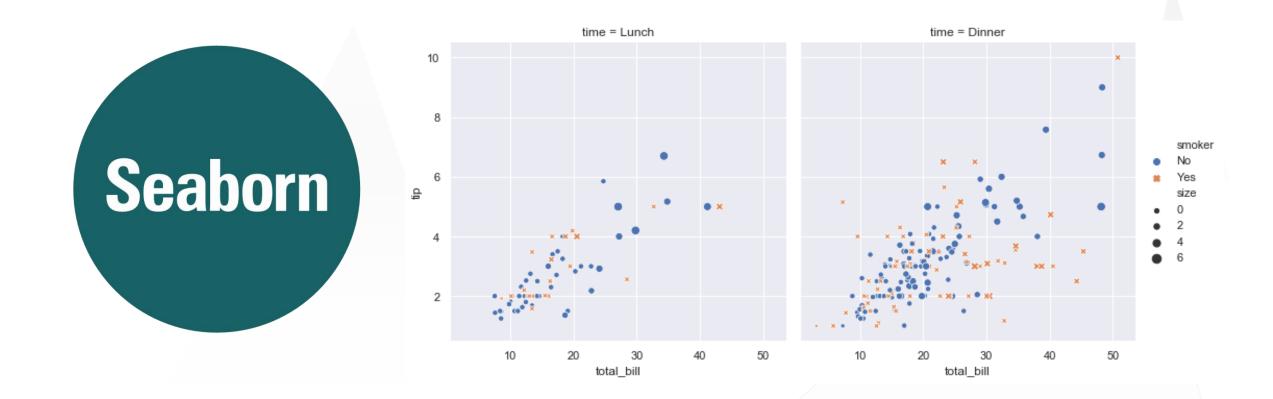


## Thư viện Python



#### 2.2.6. Seaborn

Seaborn là một thư viện để tạo đồ họa thống kê trong Python. Nó được xây dựng trên nền tảng của thư viện matplotlib và tích hợp chặt chẽ với cấu trúc dữ liệu.



## 2.3 Youtube API

#### Định nghĩa API & Youtube API

API là các phương thức, giao thức kết nối với các thư viện và ứng dụng khác. Nó là viết tắt của Application Programming Interface – giao diện lập trình ứng dụng. API cung cấp khả năng cung cấp khả năng truy xuất đến một tập các hàm hay dùng. Và từ đó có thể trao đổi dữ liệu giữa các ứng dụng.



## 2.3 Youtube API

You Tube

Định nghĩa API & Youtube API

API của Youtube cho phép nhà phát triển **tích hợp các video và chức năng** của YouTube vào các trang web hoặc ứng dụng. Youtube API cho phép các nhà phát triển **truy cập số liệu thống kê** 



#### Đọc dữ liệu



Ta sẽ đọc dữ liệu dưới dạng file CSV và đưa vào DataFrame của Pandas

```
df = pd.read_csv("C:/Users/trieu/Desktop/VN_data.csv",lineterminator='\n')
```

We set some configuration options just for improving visualization graphs; nothing crucial

```
PLOT_COLORS = ["#268bd2", "#0052CC", "#FF5722", "#b58900", "#003f5c"]
pd.options.display.float_format = '{:.2f}'.format
sns.set(style="ticks")
plt.rc('figure', figsize=(8, 5), dpi=100)
plt.rc('axes', labelpad=20, facecolor="#ffffff", linewidth=0.4, grid=True, labelsize=14)
plt.rc('patch', linewidth=0)
plt.rc('xtick.major', width=0.2)
plt.rc('ytick.major', width=0.2)
plt.rc('grid', color='#9E9E9E', linewidth=0.4)
plt.rc('font', family='Arial', weight='400', size=10)
plt.rc('text', color='#282828')
plt.rc('savefig', pad_inches=0.3, dpi=300)
```



	video_id	title	publishedAt	channelld	channelTitle	categoryld	trending_date	tags	view_count
0	WAxxfzdcNdA	em bỏ hút thuốc chưa - người yêu cũ nhắn tin h	2020-05- 17T16:15:11Z	UC90LfbAFYhRLh86Qd-Fs4zg	BICH PHUONG	10	20.26.05	em bỏ hút thuốc chưa anh bỏ hút thuốc chưa em	12541198
1	8mltWlx3cs0	#3 Đản ông đở nhất là phản bội, còn em không s	2020-05- 22T14:00:11Z	UC2Lgi2uPsOcCVF3imz7l2mg	Vie GIÅITRÍ	24	20.26.05	vie giai tri giai tri tv người ấy là ai người	5148975
2	ayJY9ieBuEU	KHÔNG THỂ CÙNG NHAU SUỐT KIÉP - HOÀ MINZY (ft	2020-05- 13T13:00:10Z	UCjm_FW7t1gam7qLldSVOclw	Hòa Minzy	10	20.26.05	hoa minzy hòa minzy hòa minzy 2020 không thế c	21413308
3	GKTpUGkhvig	FAPtv Cơm Nguội: Tập 220 - Làng Nhảm Nhí	2020-05- 23T12:58:15Z	UC0jDoh3tVXCaqJ6oTve8ebA	FAP TV	1	20.26.05	FAPTV faptv faptivi FAPtivi Faptv cơm nguộii c	3681043
4	qGjAWJ2zWWI	Agust D '대취타' MV	2020-05- 22T09:00:01Z	UC3IZKseVpdzPSBaWxBxundA	Big Hit Labels	10	20.26.05	BIGHIT 빅히트 방탄소년 단 BTS BANGTAN 방탄 Agust D  어거스트 디	44924142



count	likes	dislikes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	description	filename
541198	329323	10193	33498	https://i.ytimg.com/vi/WAxxfzdcNdA/default.jpg	False	False	Bích Phương - em bỏ hút thuốc chưa? (feat. tra	data_VN/VN01.csv
48975	47047	3533	6017	https://i.ytimg.com/vi/8mltWlx3cs0/default.jpg	False	False	#NgườiẤyLảAi #NguoiAyLaAiMua3 #VieChannelHTV2 	data_VN/VN01.csv
13308	637643	20295	52810	https://i.ytimg.com/vi/ayJY9ieBuEU/default.jpg	False	False	KHÔNG THỂ CÙNG NHAU SUỐT KIÉP - HOÀ MINZY (ft	data_VN/VN01.csv
81043	92551	3305	2970	https://i.ytimg.com/vi/GKTpUGkhvig/default.jpg	False	False	FAPtv Cơm Nguội: Tập 220 - Làng Nhảm NhíSăn X	data_VN/VN01.csv
24142	5752355	99816	743758	https://i.ytimg.com/vi/qGjAWJ2zWWI/default.jpg	False	False	Agust D '대취타' MV Agust D - 'D-2'	data_VN/VN01.cs

Mô tả dữ liệu

DownloadGoogl...



Giải thích cột dữ liệu

	Các thuộc tính	Giải thích
	video_id	Mã của video đăng trên Youtube
TE	title	Tiêu đề của video
•	publishedAt	Thời gian video được tải lên Youtube
<b>I</b>	channelId	Mã của kênh Youtube đăng tải video
NBC	channelTitle	Tên của kênh Youtube đăng tải video
	categoryld	Mã phân loại thể loại video được quy định bởi Youtube
6	trending_date	Ngày video đạt độ phổ biến trên youtube
•	tags	Từ khóa liên quan đến video
333	view_count	Số lượt người xem video
***	likes	Số lượt thích video
100	dislikes	Số lượt không thích video
<b>=</b>	comment_count	Tổng số bình luận video
₽:	thumbnail_link	Liên kết hình ảnh trình bày video
0	comments_disabled	Video có tắt bình luận hay không
0	ratings_disabled	Video có tắt đánh giá hay không
<b>4</b>	description	Mô tả video

#### Làm sạch dữ liệu



df[df["description"].apply(lambda x: pd.isna(x))].head(3)												
	video_id	title	publishedAt	channelld	channelTitle	categoryld	trending_date	tags	view_count	likes	dislikes	comment_c
117	JAgsJJnOZzQ	Yug mes nyuam them nqi part10	2020-05- 25T00:52:03Z	UCzFOZ_ezFekHPy3RL4KmbyQ	Paj ncaim toj Channel	22	20.26.05	[none]	93197	1740	170	





```
cdf = df["trending_date"].apply(lambda x: x[6:]).value_counts() \
            .to_frame().reset_index() \
            .rename(columns={"index": "date", "trending_date": "No_of_videos"})
print(cdf)
fig, ax = plt.subplots()
_ = sns.barplot(x="date", y="No_of_videos", data=cdf,
               palette=sns.color_palette(['#ff764a', '#ffa600'], n_colors=7), ax=ax)
_ = ax.set(xlabel="Date", ylabel="No. of videos")
 date No_of_videos
   05
                 772
1 06
         800
         700
         600
   No. of videos
         500
         300
         200
         100
                                 05
                                                                          06
                                                    Date
```

#### df.describe()

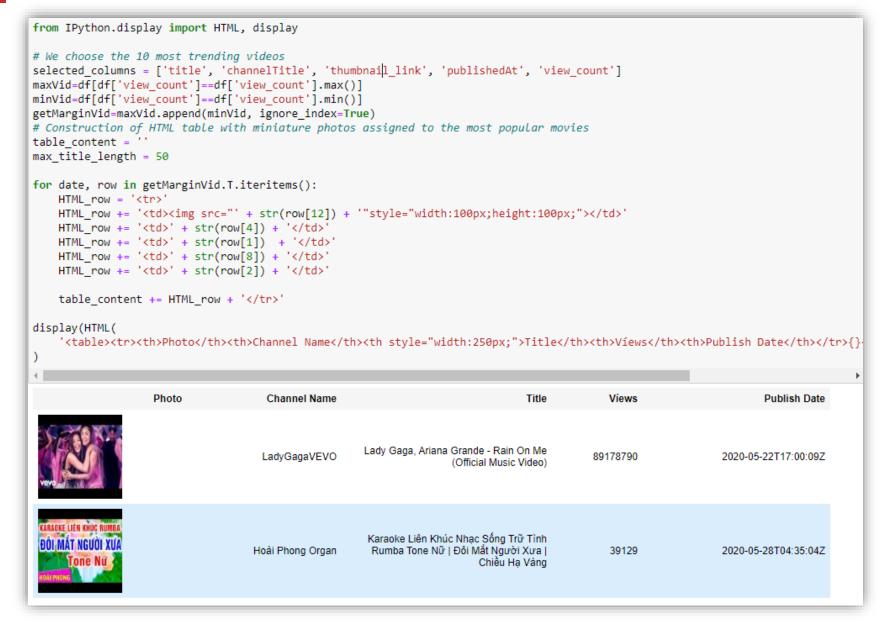
	categoryld	view_count	likes	dislikes	comment_count
count	1572.00	1572.00	1572.00	1572.00	1572.00
mean	15.41	4229646.61	164127.98	5170.64	16430.74
std	8.73	9973229.21	669695.63	19109.05	80280.03
min	1.00	39129.00	0.00	0.00	0.00
25%	10.00	586796.00	3760.50	347.75	232.25
50%	17.00	1394699.50	13474.00	926.50	798.00
75%	24.00	3205167.75	50753.50	2120.25	3445.75
max	28.00	89178790.00	6553197.00	189535.00	872476.00

Dựa vào bảng trên, chúng ta rút ra được một số kết luận như sau:

- Số lượt views trung bình của Video Treding là 4,229,646.61 . Trung vị của lượt Views là 1,394,699.50 , điều đó có nghĩa là một nửa số video trong tập dữ liệu có lượt Views lớn hơn số này, và nửa còn lại sẽ bé hơn số này
- Số lượt likes trung bình của mỗi video là 164,127.98, trong khi đó trung bình của lượt dislikes là 5170.64.
- Mỗi video ở Youtube Treding Việt Nam trung bình có khoảng 16,430 lượt comments trong khi đó trung vị là 798
- Ta cũng có thể thấy, Video có lượt views lớn nhất là 89,178,790.00, video có lượt views nhỏ nhất là 39,129.00, bảng dưới đây cho ta thấy điều đó

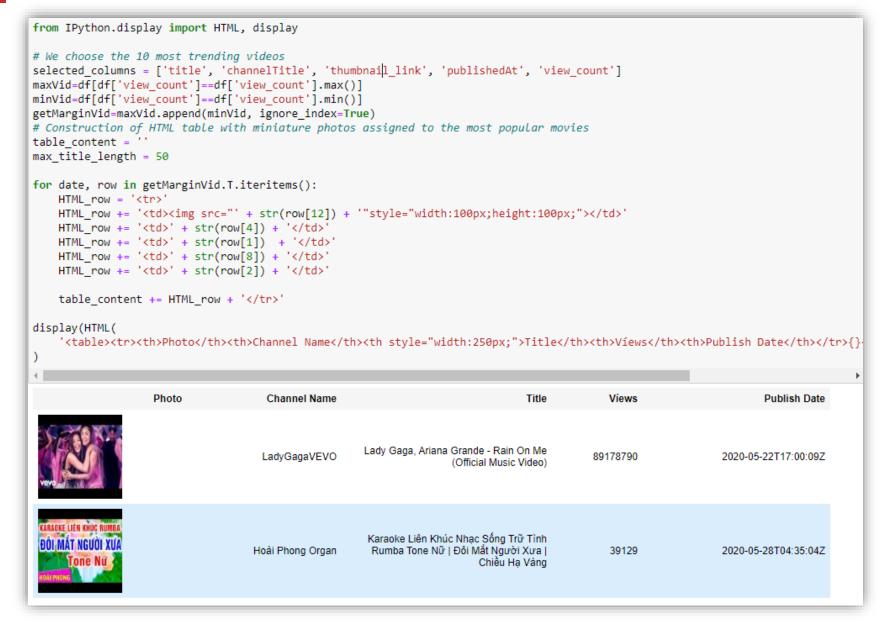


#### Biểu thị một số giá trị thống kê





#### Biểu thị một số giá trị thống kê





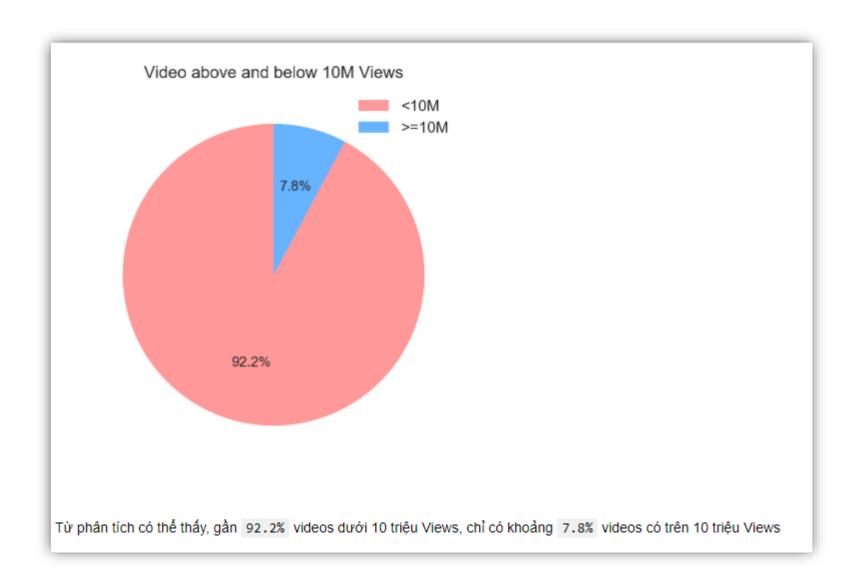
Biểu đồ lượt xem

Sau khi phân tích tập dữ liệu, tuy sự phân bố của lượt Views là rất lớn trải dài từ hơn 30 nghìn đến hơn 80 triệu Views, tuy nhiên nhóm nhận thấy lượt Views trong nhóm từ 0-10 triệu Views chiếm đa số, gần như đa số lượt video của tập dữ liêu

```
trong nhóm từ 0-10 triệu Views chiếm đa số, gần như đa số lượt video của tập dữ liệu
fig, ax = plt.subplots()
_ = sns.distplot(df[df["view_count"] < 10e6]["view_count"], kde=False,</pre>
                  color=PLOT_COLORS[1], hist_kws={'alpha': 1}, ax=ax)
_ = ax.set(xlabel="Views", ylabel="No. of videos")
plt.title("Total Videos and Views < 10M Views")
Text(0.5, 1.0, 'Total Videos and Views < 10M Views')</pre>
                                        Total Videos and Views < 10M Views
           250
           200
    No. of videos
           150
           100
            50
                                  0.2
                  0.0
                                                  0.4
                                                                  0.6
                                                                                  0.8
                                                                                                  1.0
                                                                                                   1e7
                                                        Views
```

#### Biểu đồ lượt xem



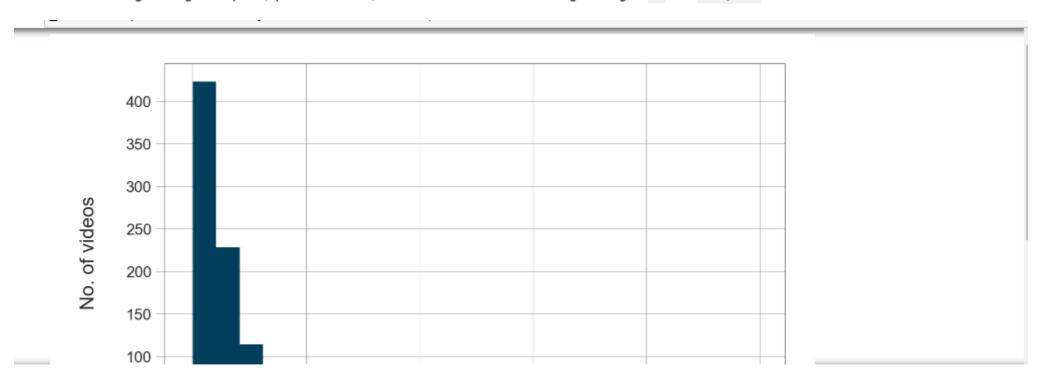


#### Biểu đồ lượt thích



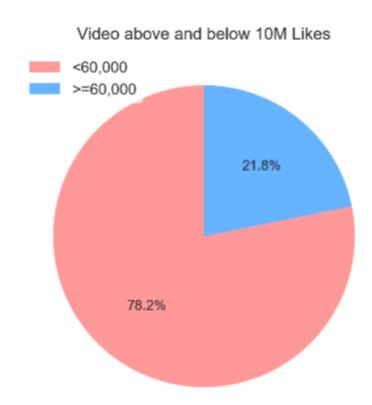
Sau views , chúng ta sẽ phân tích histogram cho cột likes

Chúng ta cùng nhìn qua sự phân bố của lượt likes của các videos nằm trong khoảng từ 0 đến 100,000 likes

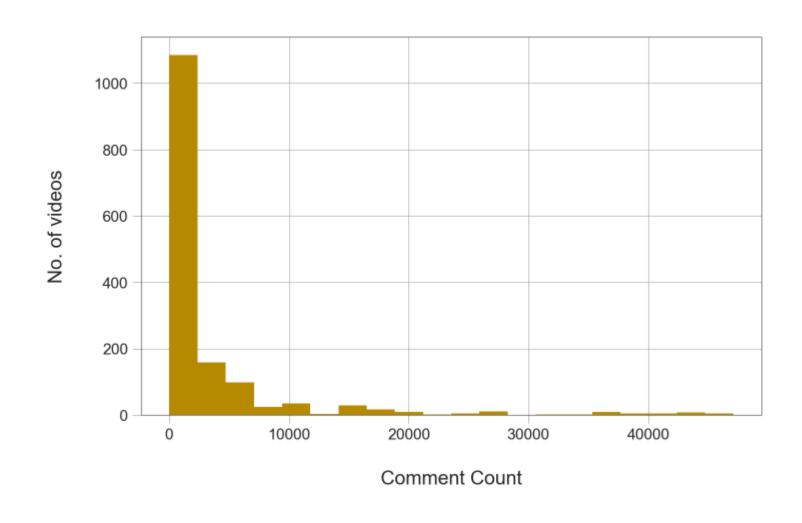


Từ biểu đồ trên, ta có thể thấy được từ 60,000 likes thì tập dữ liệu bắt đầu có sự phân hóa thành hai miền, nửa bên phải chiếm rất nhiều videos có lượt likes nhỏ hơn 60,000, bên phía còn lại thì chiếm ít hơn, ta có thể dự đoán tỉ lệ của lượt likes cũng tuân theo quy tắc 80/20, để chứng minh cho dự đoán đó ta sẽ cùng nhìn biểu đồ dưới đây.



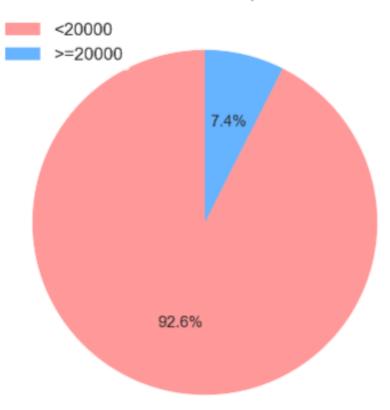








Video above and below 20,000 Comments







# Quy trình phân tích Tìm hiểu về các cột không phải dạng số

	video_id	title	publishedAt	channelld	channelTitle	trending_date	tags	thumbnail_link	description	filename	categ
count	1572	1572	1572	1572	1572	1572	1572	1572	1572	1572	
unique	430	436	424	285	285	9	419	430	443	9	
top	c896HneZZF0	ERIK - 'Em Không Sai, Chúng Ta Sai' (Official MV)	'Em nông Sai, 2020-05- núng 06T12:00:13Z Sai' ficial	UCH- _NjZ3ojREOwBZt- 3pYLA	WeTV Vietnam	20.01.06	[none]	https://i.ytimg.com/vi/Zsx- 4qsOEks/default.jpg	►Bạn muốn xem trọn bộ, hãy tải ngay ứng dụng W	data_VN/VN05.csv	
freq	9	9	17	27	27	200	30	9	27	200	

Từ bảng trên, ta có thể thấy có 9 giá trị trending\_date duy nhất, điều đó có nghĩa là tập dữ liệu này được thu thập từ 9 days.

Từ cột video\_id, ta có thể thấy tập dữ liệu có 1572 videos, nhưng ở đây chỉ có 430 videos duy nhất - điều đó có nghĩa là có rất nhiều videos xuất hiện nhiều lần trên bảng xếp hạng trending trong suốt nhiều ngày.

Ta có thể thấy ca khúc Em Không Sai, Chúng Ta Sai của ca sĩ Erik xuất hiện liên tục trong suốt 9 ngày, đây là video có tần xuất xuất hiện lớn nhất.

Bên cạnh đó có một điều rất lạ trong tập dữ liệu: Chúng ta có 430 video IDs duy nhất, theo logic thì sẽ có 430 video 430 duy nhất. Tuy nhiên ta có thể thấy có tận 436 title, chứng tổ có 6 videos đã được đổi tên trong suốt quá trình chúng trở thành trending videos.

Ta sẽ cùng nhìn vào một video title để hiểu được ví dụ về một video đã đổi tên trong quá trình đăng trên Youtube.





# Quy trình phân tích <u>Tìm hiểu về các cột không phải dạng số</u>

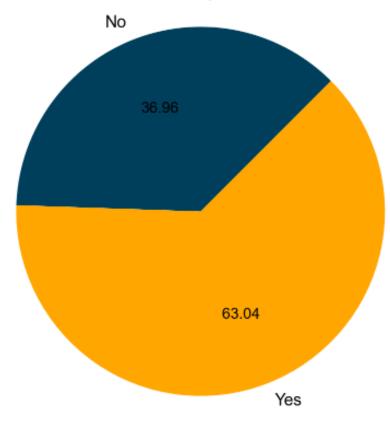
	А	В	С	D	Е	F
1	video_id	title	publishedAt	channelld	channelTitle	categoryld
2	0ZcXDfqYPDI	[Nhạc chế] CẬY CẬY ĐẠI CA   Xuân Dích & Thế Một   Trai Ngoan Parody	2020-05-30T04:00:11Z	UCy5kU2hAo7ZJZOzSmzRAnLQ	Trai Ngoan	10
3	0ZcXDfqYPDI	[Nhạc chế] CẬY CẬY ĐẠI CA   Xuân Dích & Thế Một   Trai Ngoan Parody	2020-05-30T04:00:11Z	UCy5kU2hAo7ZJZOzSmzRAnLQ	Trai Ngoan	10
4	0ZcXDfqYPDI	[Nhạc chế] CẬY CẬY ĐẠI CA   Xuân Dích & Thế Một   Trai Ngoan Parody (Quảng cáo Nhất Kiếm Giang Hồ)	2020-05-30T04:00:11Z	UCy5kU2hAo7ZJZOzSmzRAnLQ	Trai Ngoan	10
5						





## Bao nhiêu video không có viết hoa trong tên?



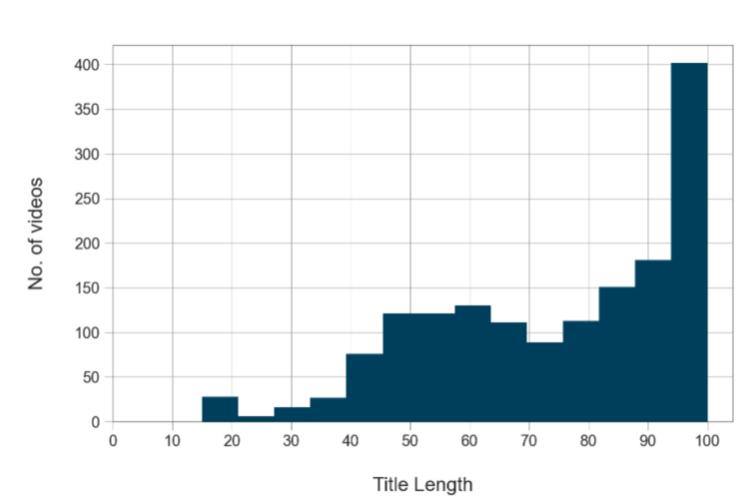


True 0.63 False 0.37

Name: contains\_capitalized, dtype: float64

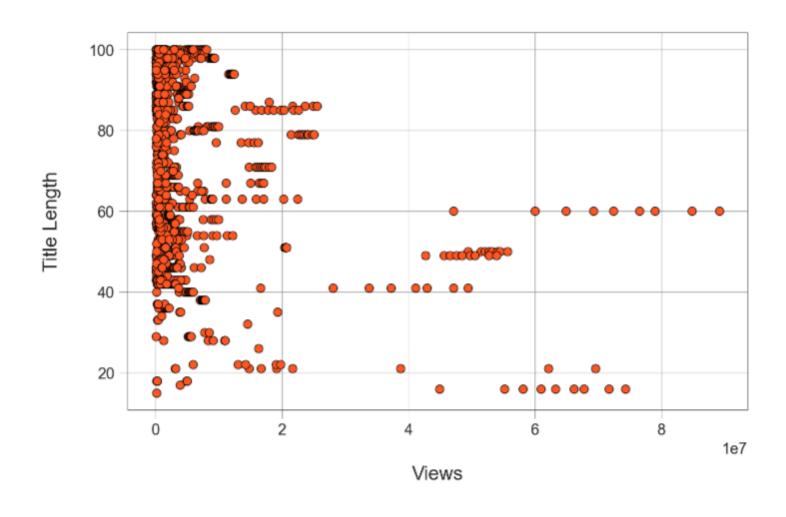
Một điều bất ngờ là có tới 36,96% videos không có bất cứ một chữ viết hoa nào trong tên





Ta có thể thấy mối quan hệ giữa độ dài tiêu đề video và số lượng videos, có thể thấy đa số các videos ở Youtube Treding Việt Nam có tiêu đề khá dài, dao động mạnh từ 40-100 từ Tuy nhiên, để kiểm chứng xem có thực sự là videos càng nhiều chữ sẽ có lượt views càng lớn hay không thì ta sẽ vẽ scatter plot







#### Sự tương quan giữa các biến của dataset

Bây giờ, hấy xem các biến số liệu tương quan với nhau như thế nào: ví dụ như nhóm muốn xem lượt xem và lượt thích tương quan như thế nào, có nghĩa là lượt xem và thích tăng giảm cùng nhau (tương quan tích cực)? Có một trong giá trị tăng khi giá trị kia giảm và ngược lại (tương quan tiêu cực)? Hay chúng không tương quan?

Tương quan được biểu diễn dưới dạng một giá trị giữa -1 và + 1 trong đó + 1 biểu thị mối tương quan dương cao nhất, -1 biểu thị mối tương quan âm cao nhất và 0 biểu thị rằng không có mối tương quan.

Chúng ta hãy xem bảng tương quan giữa các biến số liệu của nhóm (chỉ các biến số và boolean)

df.corr()

	categoryld	view_count	likes	dislikes	comment_count	$comments\_disabled$	ratings_disabled	$contains\_capitalized$	title_length
categoryld	1.00	-0.08	-0.07	-0.08	-0.08	0.09	0.09	-0.01	0.10
view_count	-0.08	1.00	0.85	0.90	0.75	0.01	-0.03	0.06	-0.29
likes	-0.07	0.85	1.00	0.88	0.95	-0.02	-0.03	0.08	-0.33
dislikes	-0.08	0.90	0.88	1.00	0.80	0.00	-0.03	0.03	-0.28
comment_count	-0.08	0.75	0.95	0.80	1.00	-0.02	-0.01	0.10	-0.34
comments_disabled	0.09	0.01	-0.02	0.00	-0.02	1.00	-0.01	0.04	-0.01
ratings_disabled	0.09	-0.03	-0.03	-0.03	-0.01	-0.01	1.00	-0.03	-0.06
contains_capitalized	-0.01	0.06	0.08	0.03	0.10	0.04	-0.03	1.00	-0.06
title_length	0.10	-0.29	-0.33	-0.28	-0.34	-0.01	-0.06	-0.06	1.00

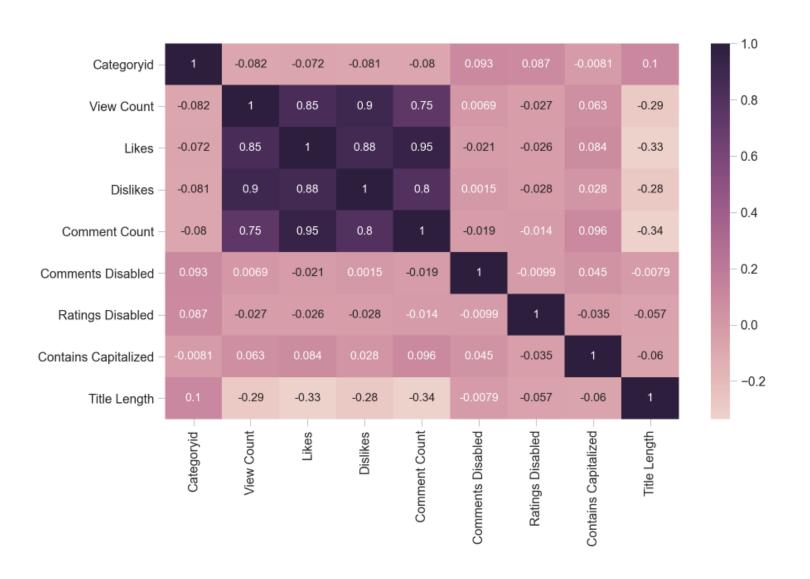
Ví dụ, chúng tôi thấy rằng lượt xem và lượt thích có mối tương quan tích cực với giá trị tương quan là 0,85; nhóm cũng thấy một mối tương quan tích cực cao (0,95) giữa số lượt thích và số bình luận, và giữa số lượt không thích và số bình luận (0,8).

Videos có nhiều likes thì cũng sẽ có mối tương quan thuận với dislikes





#### Sự tương quan giữa các biến của dataset







#### Wordcloud cho title và tag

Từ title của tập dữ liệu, ta có thể thấy một số từ xuất hiện rất nhiều lần nhứ Tập, Nhạc, Music, Official, Phim, Video đây đều là những từ thuộc các videos thuộc nhóm lĩnh vực Giải trí và âm nhạc.





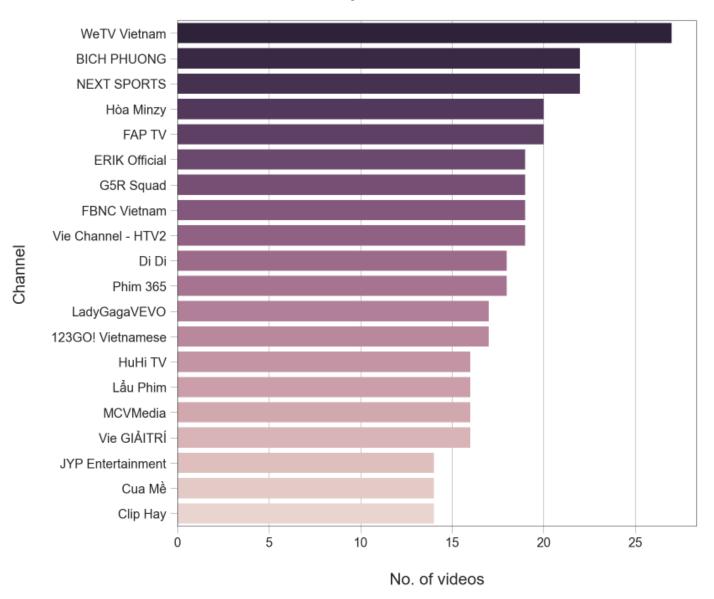
#### Wordcloud cho title và tag







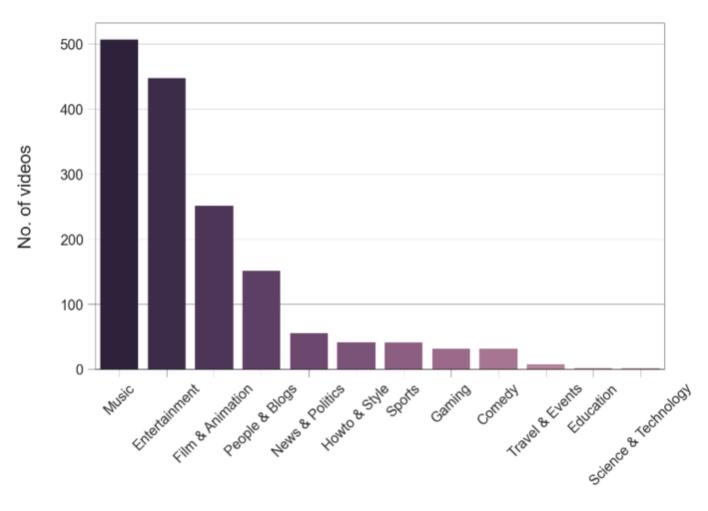
## Top các kênh có video trending nhiều nhất VN







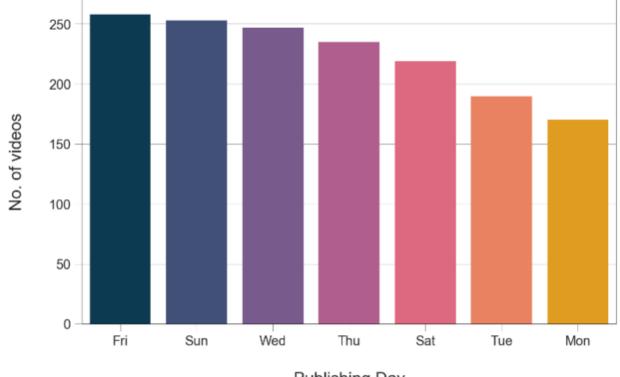
#### Thể loại video có trending video nhiều nhất VN







#### Thời gian đăng của các trending video

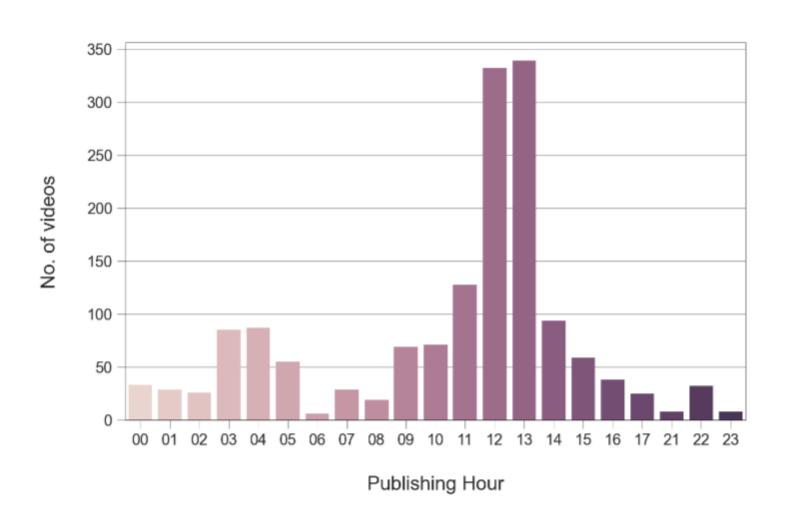


**Publishing Day** 



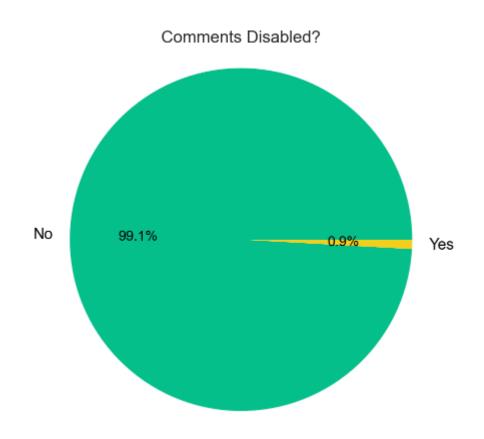


#### Thời gian đăng của các trending video



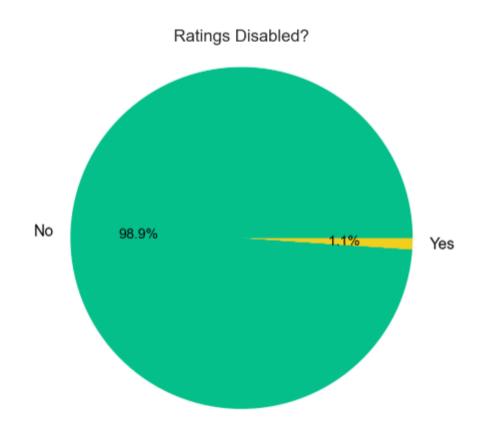


#### Có bao nhiều Videos được đóng tính năng bình luận





#### Bao nhiêu videos được khóa đánh giá (like/dislike)?







Nhóm đã thu thập và phân tích Youtube Trending Việt Nam trong 9 ngày. Tập dữ liệu được thu thập vào cuối tháng 5 và đầu tháng 6 năm 2020. Nó bao gồm 1572 videos đầu vào.

92.2% videos trending có dưới 10 triệu views



81,3% videos có dưới 5 triệu views



Nhóm đã thu thập và phân tích Youtube Trending Việt Nam trong 9 ngày. Tập dữ liệu được thu thập vào cuối tháng 5 và đầu tháng 6 năm 2020. Nó bao gồm 1572 videos đầu vào.

78,2% của videos trending có dưới 60,000 likes



**82,7%** videos có dưới 100,000 likes



Nhóm đã thu thập và phân tích Youtube Trending Việt Nam trong 9 ngày. Tập dữ liệu được thu thập vào cuối tháng 5 và đầu tháng 6 năm 2020. Nó bao gồm 1572 videos đầu vào.



92,6% videos trending có ít hơn 20,000 comments



Nhóm đã thu thập và phân tích Youtube Trending Việt Nam trong 9 ngày. Tập dữ liệu được thu thập vào cuối tháng 5 và đầu tháng 6 năm 2020. Nó bao gồm 1572 videos đầu vào.

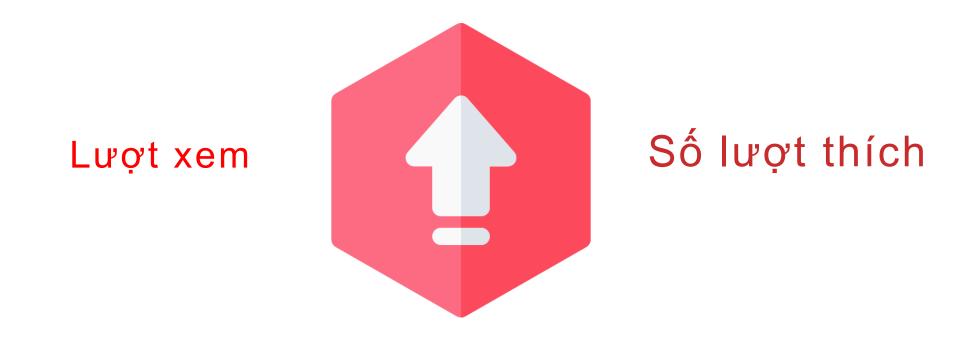
Những videos có hơn 5,000,000 views và lớn hơn thường có độ dài tiêu đề video 15 và 20 từ, hoặc là 60 từ.

Các stopword - và | rất phổ biến trong title của các videos

Các cụm từ như
'Official', 'Video',
'Phim', 'Nhạc', 'Tập'
rất phổ biến trong tiêu
đề các video



Nhóm đã thu thập và phân tích Youtube Trending Việt Nam trong 9 ngày. Tập dữ liệu được thu thập vào cuối tháng 5 và đầu tháng 6 năm 2020. Nó bao gồm 1572 videos đầu vào.





Nhóm đã thu thập và phân tích Youtube Trending Việt Nam trong 9 ngày. Tập dữ liệu được thu thập vào cuối tháng 5 và đầu tháng 6 năm 2020. Nó bao gồm 1572 videos đầu vào.

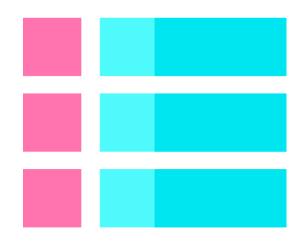
Số lượt thích



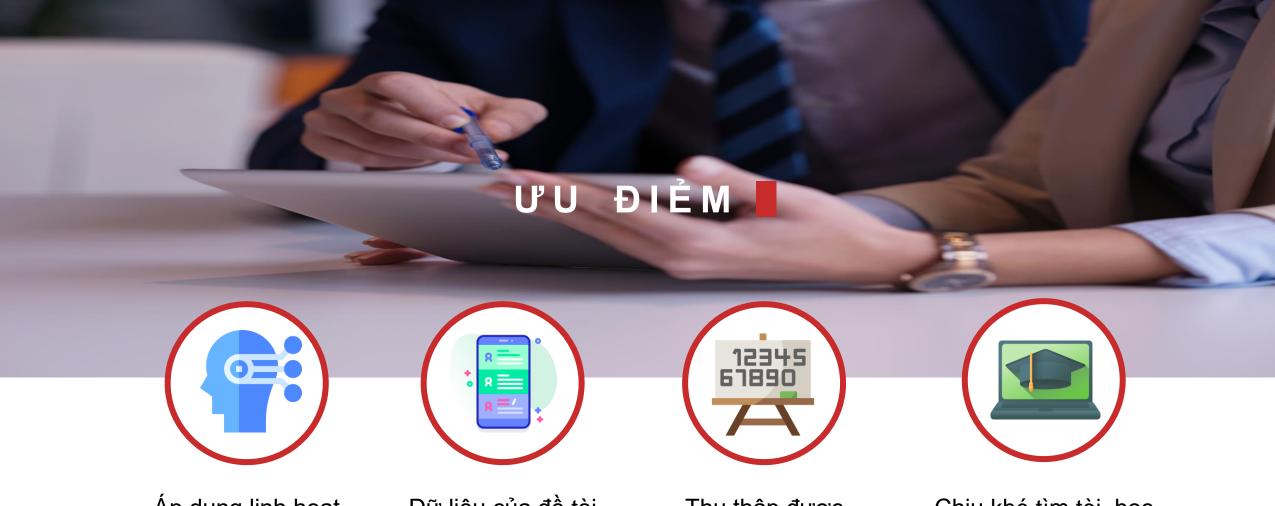
Lượt bình luận



Nhóm đã thu thập và phân tích Youtube Trending Việt Nam trong 9 ngày. Tập dữ liệu được thu thập vào cuối tháng 5 và đầu tháng 6 năm 2020. Nó bao gồm 1572 videos đầu vào.



Danh mục có số lượng video xu hướng lớn nhất là \* 'Âm nhạc' \*



Áp dụng linh hoạt những kiến thức được học Dữ liệu của đề tài mang tính ứng dụng cao

Thu thập được những con số dữ liệu ấn tượng Chịu khó tìm tòi, học hỏi



Nguồn dữ liệu vẫn chưa nhiều

Kiến thức của nhóm vận dụng vào đề tài còn nhiều hạn chế

Kết quả đồ án chỉ tạm dừng lại ở mức phân tích cơ bản



- -Hoàn thiện xây dựng bộ dữ liệu tham khảo, nghiên cứu cho các Doanh nghiệp. Đồng thời, đưa ra bộ tiêu chí video top trending YouTube cho các YouTuber.
- Thực hiện lọc các video xuất hiện nhiều lần trong top Trending, các video giữ hạng lâu trên top trending để tiến hành phân tích sâu hơn.

