



ĐẠI HỌC QUỐC GIA TP.HCM  
ĐẠI HỌC KINH TẾ - LUẬT

# ĐỒ ÁN CUỐI KỲ

MÔN HỌC: PHÂN TÍCH DỮ LIỆU WEB

## PHÂN TÍCH DỮ LIỆU YOUTUBE TRENDING



GVHD: TH.S. ĐẶNG NHÂN CÁCH  
Nhóm NTV:

- |                       |            |
|-----------------------|------------|
| 1. Nguyễn Anh Nhật    | K174111311 |
| 2. Nguyễn Quốc Triệu  | K174111323 |
| 3. Thiềm Ánh Tường Vy | K174111329 |

TP. HCM, ngày 08/06/2020

## MỤC LỤC

LỜI MỞ ĐẦU .....	1
CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI .....	2
1.1.Lý do chọn đề tài.....	2
1.1.1.Bối cảnh ngành truyền thông hiện nay .....	2
1.1.2.Tiềm năng YouTube .....	3
1.2.Mục tiêu đề tài .....	5
1.3.Phạm vi và đối tượng nghiên cứu .....	6
1.4.Phương pháp nghiên cứu .....	6
1.5.Ý nghĩa đề tài .....	6
1.6.Kết cấu đề tài .....	7
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT .....	8
2.1.Ngôn ngữ lập trình (Python) .....	8
2.2.Thư viện Python.....	10
2.2.1.Request .....	10
2.2.2.BeautifulSoup .....	10
2.2.3.Pandas .....	10
2.2.4.Numpy .....	10
2.2.5.Matplotlib .....	11
2.2.6.Seaborn.....	11
2.3.YouTube .....	12
2.3.1.Tổng quan .....	12
2.3.2.Youtube API .....	14
CHƯƠNG 3: TRIỂN KHAI PHÂN TÍCH DỮ LIỆU .....	16
3.1.Quy trình phân tích .....	16
3.1.1.Đặt vấn đề.....	16
3.1.2.Phân tích dữ liệu.....	16
3.1.3.Phân tích kết quả và trực quan hoá dữ liệu .....	28
3.2.Kết quả thu được .....	52
3.2.1.Tóm tắt kết quả phân tích .....	52
CHƯƠNG 4: KẾT LUẬN VÀ ĐÁNH GIÁ.....	54

4.1.Tóm tắt nội dung và kết quả của đề tài .....	54
4.2.Uu điểm .....	54
4.3.Nhược điểm .....	54
4.4.Hướng phát triển của đề tài.....	54
BÁO CÁO QUÁ TRÌNH LÀM VIỆC NHÓM .....	56
TÀI LIỆU THAM KHẢO.....	58
PHỤ LỤC 1: SOURCE CODE ĐỒ ÁN.....	59

## DANH MỤC HÌNH ẢNH

Hình 1.1: Thống kê số lượng người dùng Internet và mạng xã hội tháng 1 năm 2019...	2
Hình 1.2: Logo YouTube .....	3
Hình 1.3: Bảng xếp hạng các mạng xã hội hoạt động tích cực nhất .....	4
Hình 1.4: Kết quả so sánh thời lượng người dùng Việt Nam dành cho YouTube khu vực Châu Á – Thái Bình Dương .....	4
Hình 2.1: Cấu trúc Dataframe của bảng dữ liệu Khách hàng .....	10
Hình 2.2: Biểu đồ được tạo bởi thư viện Matplotlib .....	11
Hình 2.3: Biểu đồ được tạo bởi thư viện Seaborn thể hiện dữ liệu .....	12
Hình 2.4: Youtube API áp dụng cho nhiều nền tảng khác nhau .....	15
Hình 3.1: Youtube API áp dụng cho nhiều nền tảng khác nhau .....	17
Hình 3.2: Trang lấy API Key của Google .....	18
Hình 3.3: Code import thư viện dùng để Crawling dữ liệu .....	18
Hình 3.4: Các trường dữ liệu mà ta muốn lấy từ Youtube API .....	19
Hình 3.5: Khai báo các header để tương tác Youtube API .....	19
Hình 3.6: Code thể hiện cho hàm setup và hàm prepare_feature .....	20
Hình 3.7: Code thể hiện cho hàm api_request .....	20
Hình 3.8: Hàm get_pages .....	21
Hình 3.9: Hàm get_videos .....	22
Hình 3.10: Lưu dữ liệu vào file CSV .....	23
Hình 3.11: Minh họa dữ liệu thu được .....	23
Hình 3.12: Cài đặt các thư viện .....	24
Hình 3.13: Cấu hình cách thể hiện dữ liệu .....	24
Hình 3.14: Đọc dữ liệu bằng file CSV .....	24
Hình 3.15: Đưa dữ liệu vào Dataframe của Pandas .....	24
Hình 3.16: Dòng lệnh để đưa dữ liệu thành bảng .....	25
Hình 3.17: Các dòng và cột dữ liệu video xu hướng trên YouTube .....	25
Hình 3.18: Các dòng và cột dữ liệu video xu hướng trên Youtube .....	26
Hình 3.19: Các thuộc tính của các dòng dữ liệu .....	26
Hình 3.20: Dòng lệnh lấy thông tin dữ liệu .....	27

Hình 3.21: Kết quả sau khi lấy thông tin của dữ liệu .....	27
Hình 3.22: Dòng lệnh giúp làm sạch dữ liệu .....	28
Hình 3.23: Kết quả sau khi làm sạch dữ liệu .....	28
Hình 3.24: Dòng lệnh nhận biết thời gian dữ liệu thu thập .....	28
Hình 3.25: Biểu đồ cột của ngày thu thập dữ liệu .....	29
Hình 3.26: Dòng lệnh giúp biểu thị giá trị thống kê .....	29
Hình 3.27: Kết quả sau khi thu thập giá trị thống kê .....	29
Hình 3.28: Dòng lệnh chọn những video có lượt views từ cao đến thấp .....	30
Hình 3.29: Dòng lệnh hình thành bảng biểu thị video thịnh hành .....	30
Hình 3.30: Bảng biểu thị những video thịnh hành .....	31
Hình 3.31: Dòng lệnh biểu thị lượt views bằng biểu đồ cột .....	31
Hình 3.32: Biểu đồ cột hiển thị số lượng video có lượt views dưới 10 triệu .....	32
Hình 3.33: Dòng lệnh biểu thị lượt views bằng biểu đồ tròn .....	32
Hình 3.34: Biểu đồ tròn hiển thị số lượng video có lượt views dưới và trên 10 triệu ..	33
Hình 3.35: Dòng lệnh biểu thị lượt likes bằng biểu đồ cột .....	33
Hình 3.36: Biểu đồ cột hiển thị số lượng likes dưới 100.000 .....	34
Hình 3.37: Dòng lệnh biểu thị lượt likes bằng biểu đồ tròn .....	34
Hình 3.38: Biểu đồ tròn hiển thị số lượng video có lượt likes trên và dưới 60.000 ....	35
Hình 3.39: Dòng lệnh biểu thị lượt comment bằng biểu đồ cột .....	35
Hình 3.40: Biểu đồ cột hiển thị số lượng comment dưới 40.000 .....	36
Hình 3.41: Dòng lệnh biểu thị lượt likes bằng biểu đồ tròn .....	36
Hình 3.42: Biểu đồ tròn hiển thị số lượng video có lượt comment trên và dưới 20.000 .....	37
Hình 3.43: Dòng lệnh hiển thị tập dữ liệu .....	37
Hình 3.44: Kết quả của bảng dữ liệu không phải dạng số .....	37
Hình 3.45: Dòng lệnh hiển thị video title đổi tên .....	38
Hình 3.46: Kết quả sau khi nhóm các video có title bị thay đổi .....	38
Hình 3.47: Dòng lệnh biểu thị video viết hoa trong tên bằng biểu đồ tròn .....	39
Hình 3.48: Biểu đồ tròn hiển thị số lượng video viết hoa trong tên .....	39
Hình 3.49: Dòng lệnh biểu thị độ thị video title .....	40
Hình 3.50: Biểu đồ cột thể hiện độ dài video title .....	40
Hình 3.51: Dòng lệnh biểu thị mối quan hệ tiêu đề video và lượt views .....	41

Hình 3.52: Biểu đồ scatter plot thể hiện mối quan hệ tiêu đề video và lượt views .....	41
Hình 3.53: Dòng lệnh biểu thị sự tương quan giữa các biến .....	42
Hình 3.54: Bảng thể hiện sự tương quan giữa các biến .....	42
Hình 3.55: Dòng lệnh biểu thị sự tương quan giữa các biến bằng bản đồ nhiệt .....	42
Hình 3.56: Bản đồ nhiệt thể hiện sự tương quan giữa các biến .....	43
Hình 3.57: Dòng lệnh lọc stopword và các từ khóa .....	43
Hình 3.58: Kết quả sau khi lọc stopword và các từ khóa .....	44
Hình 3.59: Dòng lệnh vẽ wordcloud cho title .....	44
Hình 3.60: Wordcloud cho title .....	44
Hình 3.61: Dòng lệnh lọc các tag .....	45
Hình 3.62: Kết quả sau khi lọc các tag .....	45
Hình 3.63: Dòng lệnh vẽ wordcloud cho tag .....	45
Hình 3.64: Wordcloud cho tag .....	46
Hình 3.65: Dòng lệnh biểu thị kênh có trending video tại Việt Nam .....	46
Hình 3.66: Biểu đồ cột hiện thị các kênh có trending video tại Việt Nam .....	47
Hình 3.67: Dòng lệnh sử dụng file JSON .....	47
Hình 3.68: Dòng lệnh biểu thị số lượng trending video của từng thể loại .....	48
Hình 3.69: Biểu đồ cột thể hiện số lượng trending video của từng thể loại .....	48
Hình 3.70: Dòng lệnh biểu thị lượng video được đăng tải các ngày trong tuần .....	49
Hình 3.71: Biểu đồ cột thể hiện lượng video được đăng tải các ngày trong tuần .....	49
Hình 3.72: Dòng lệnh biểu thị lượng video được đăng tải các giờ trong ngày .....	50
Hình 3.73: Biểu đồ cột thể hiện lượng video được đăng tải các giờ trong ngày .....	50
Hình 3.74: Dòng lệnh biểu thị lượng video đóng tính năng bình luận .....	51
Hình 3.75: Biểu đồ tròn thể hiện lượng video đóng tính năng bình luận .....	51
Hình 3.76 Dòng lệnh biểu thị lượng video đóng tính năng đánh giá .....	52
Hình 3.77: Biểu đồ tròn thể hiện lượng video đóng tính năng bình luận .....	52

## **DANH MỤC BẢNG BIỂU**

Bảng 3-1. Mô tả môi trường thực nghiệm .....	16
--	----

## LỜI MỞ ĐẦU

*Thời đại Internet phát triển kéo theo sự bùng nổ của các trang mạng xã hội. Đó là nơi người dùng chia sẻ những cảm xúc, bài viết, hình ảnh, video v.v. kết nối với bạn bè khắp nơi trên thế giới. Trong đó, YouTube đang là một trong những “ông lớn” cạnh tranh với Facebook cho ngôi vị quán quân. Người dùng YouTube ngày càng tăng lên không ngừng vì những nội dung đầy sáng tạo trên đó và vì công việc YouTuber cũng đang dần trở nên “hot” hơn bao giờ hết bởi mức lương khổng lồ mà YouTube mang lại. Đó cũng chính là lí do mà nhóm lựa chọn phân tích dữ liệu YouTube trending, từ đó xây dựng bộ dữ liệu tham khảo cho Doanh nghiệp để đưa ra chiến lược quảng bá thương hiệu phù hợp. Bên cạnh đó, cũng có thể xây dựng bộ tiêu chí video lọt top trending trên YouTube cho các YouTuber.*

*Trong quá trình thực hiện và hoàn thiện đồ án, nhóm đã nhận được sự giúp đỡ, động viên và chỉ dạy hết mình của quý thầy cô, anh chị và các bạn. Bằng sự biết ơn chân thành, nhóm xin trân trọng gửi lời cảm ơn đến Ban Lãnh Đạo trường đại học Kinh tế - Luật, quý thầy cô Khoa Hệ Thống Thông Tin vì đã luôn tạo điều kiện tốt nhất cho chúng em học thêm những kiến thức hữu ích. Đặc biệt, nhóm xin gửi đến Thầy Đặng Nhân Cách - giảng viên trực tiếp giảng dạy - hướng dẫn môn học “Phân tích dữ liệu web” lời cảm ơn sâu sắc nhất vì đã luôn luôn hỗ trợ, theo sát và định hướng cách triển khai phù hợp với năng lực của chúng em. Đồng thời, thầy luôn đưa ra những ý kiến xây dựng giúp chúng em hoàn thành đồ án với kết quả tốt nhất dù cho có những khó khăn, bất cập với tình hình dịch bệnh. Chúng em luôn xem những lời động viên, quan tâm của thầy là động lực để phấn đấu. Một lần nữa, nhóm xin chân thành cảm ơn thầy. Chúc thầy thật nhiều sức khỏe để tiếp tục sứ mệnh trồng người và giúp đỡ cho nhiều thế hệ sinh viên.*

*Bên cạnh đó, với kiến thức chuyên môn hạn chế, thời gian thực hiện ngắn hạn và bản thân còn thiếu nhiều kinh nghiệm thực tiễn nên nội dung của đồ án không tránh khỏi những thiếu sót, nhóm kính mong sẽ nhận được sự góp ý từ thầy để hoàn thiện hơn. Nhóm xin chân thành cảm ơn!*

**NTV Group.**



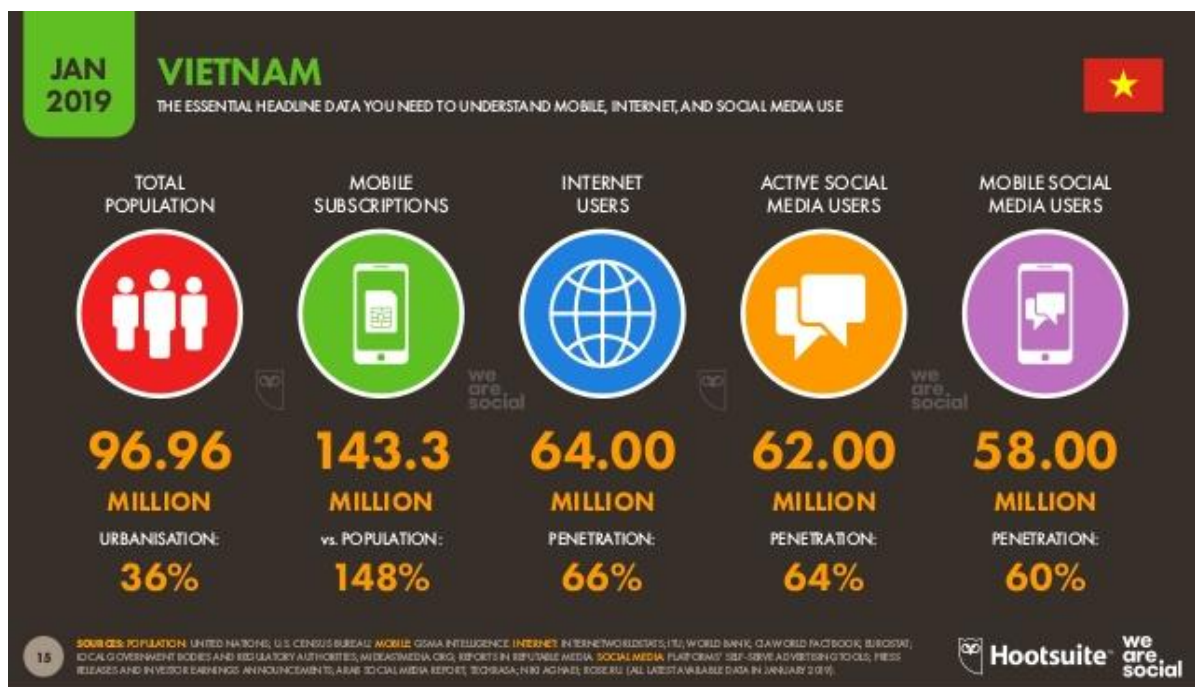
# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## 1.1. Lý do chọn đề tài

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

Theo báo cáo từ các nghiên cứu mới nhất về thị trường Việt Nam thời gian gần đây cho thấy tiềm năng phát triển ở lĩnh vực truyền thông kỹ thuật số đang ngày một rộng mở và hứa hẹn đem đến nhiều cơ hội việc làm lẫn thăng tiến cho những cá nhân có định hướng theo đuổi lĩnh vực này.

Tính đến hết tháng 1/2019, dân số Việt Nam đạt hơn 96 triệu người, hơn 64 triệu người dùng internet – 66% dân số, trong đó có hơn 62 triệu người có tài khoản mạng xã hội – chiếm 97% người dùng internet và 64% tổng dân số; trung bình một người online mạng xã hội khoảng 18 giờ 1 tuần – theo *Digital in 2019 in VietNam Report* bởi wearesocial.com. [1]



Hình 1.1: Thống kê số lượng người dùng Internet và mạng xã hội tháng 1 năm 2019

Đi cùng với sự phát triển của hạ tầng viễn thông, giúp internet phủ rộng khắp lãnh thổ Việt Nam, là tốc độ tăng trưởng lượng người tiếp cận và hoạt động tích cực

trên internet, đặc biệt là trên các mạng xã hội, đã mở ra nhiều cơ hội nghề nghiệp phát triển tăng thu nhập đầy tiềm năng. Và Youtube là một trong số đó.

### ***1.1.2. Tiềm năng YouTube***

YouTube được thành lập vào năm 2005 bởi Chad Hurley, Steve Chen và Jawed Karim. Với vai trò là một trang mạng xã hội chuyên chia sẻ video trực tuyến, người xem không chỉ xem nội dung một cách thụ động mà họ còn tương tác trong cuộc trò chuyện hai chiều với một cộng đồng cũng xem video như họ. [2]



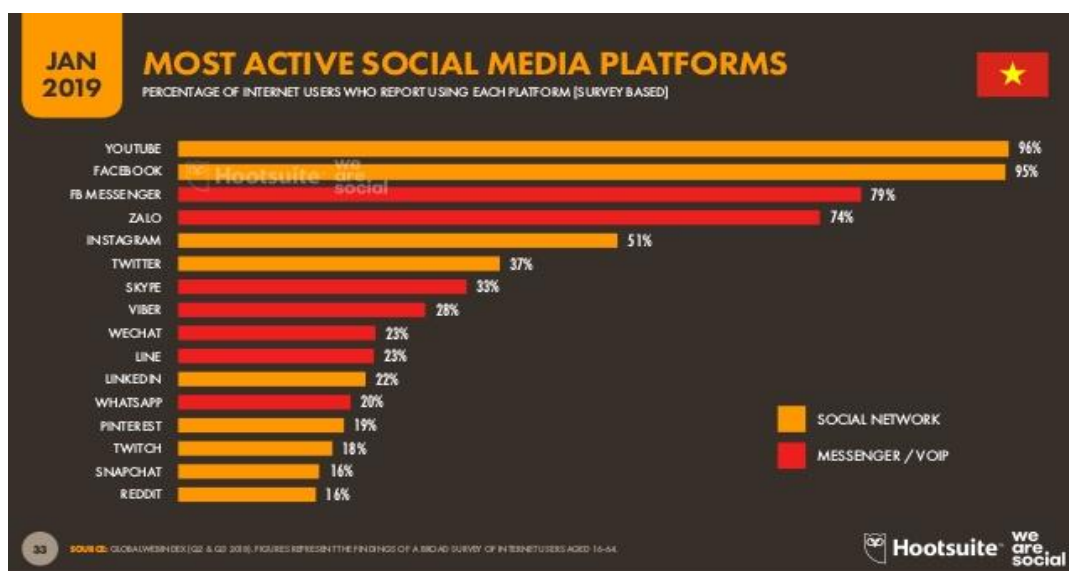
*Hình 1.2: Logo YouTube*

Khán giả gắn bó với YouTube tạo cơ hội cho người sáng tạo, thương hiệu và nhà quảng cáo xuất hiện trước khán giả vào những thời điểm quan trọng trong hành trình tiêu dùng (thời điểm này xảy ra khi mọi người truy cập vào trang để tìm hiểu thông tin, học hỏi, khám phá,... và phát sinh ra nhu cầu tiêu dùng, từ đó định hình lựa chọn và giúp họ đưa ra quyết định). Người xem, nhà quảng cáo, người sáng tạo và kể cả YouTube tương tác với nhau và cùng “nuôi dưỡng” hệ sinh thái YouTube.

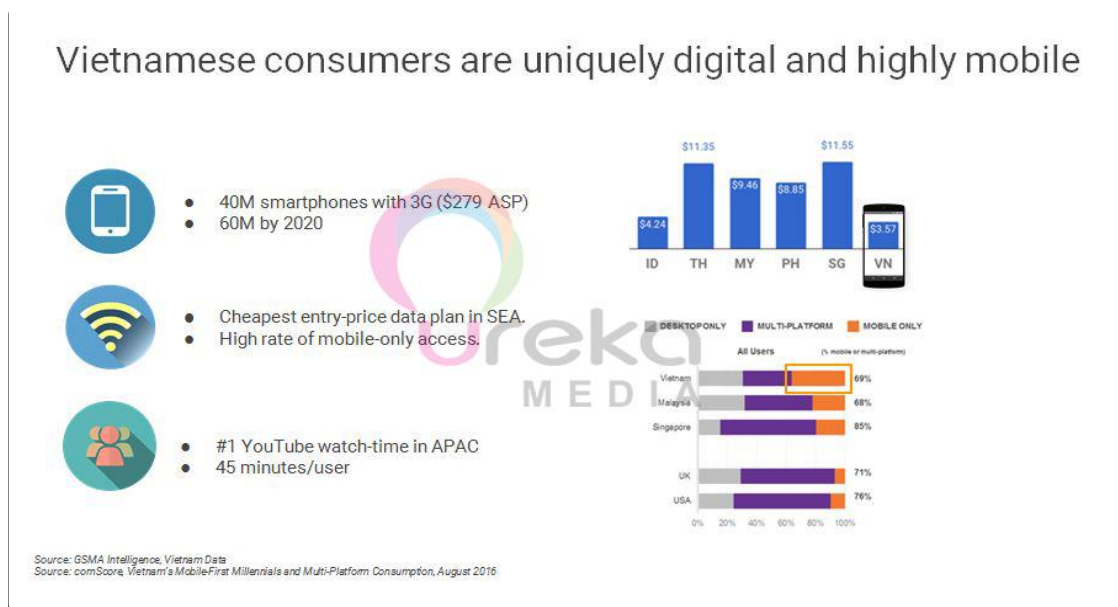
Một số những yếu tố tạo nên tính cạnh tranh của Youtube đối với các nền tảng mạng xã hội video tương tự:

- Lượng người tham gia (người sáng tạo và người xem) lâu năm và trung thành – Youtube đã chiếm được thị phần lớn trong thị trường Việt Nam với “kho tàng” video vô cùng phong phú, đáp ứng gần như đầy đủ mọi nhu cầu của người xem.

Youtube được xếp hạng là mạng xã hội hoạt động tích cực nhất ở Việt Nam (vượt qua cả Facebook) – theo *Digital in 2019 in VietNam Report* của wearesocial.com. Theo báo cáo của Google, người dùng Việt Nam là đối tượng xem Youtube nhiều nhất khu vực Châu Á – Thái Bình Dương, đứng thứ 5 trên thế giới về việc tiêu tốn thời gian cho Youtube – theo báo cáo của *GSMA Intelligence VietNam Data*. [1]



Hình 1.3: Bảng xếp hạng các mạng xã hội hoạt động tích cực nhất



Hình 1.4: Kết quả so sánh thời lượng người dùng Việt Nam dành cho YouTube khu vực Châu Á – Thái Bình Dương [3]

- Nhờ thu hút được đông đảo người xem cũng như người sáng tạo nội dung, Youtube đã trở thành nơi mà vô số thương hiệu, doanh nghiệp lựa chọn để đưa sản phẩm mình giới thiệu đến lượng lớn khách hàng tiềm năng. Đây cũng chính là đích đến quyết định cho sự “tồn tại” của trang mạng xã hội này khi thu được lợi nhuận từ các hợp đồng quảng cáo và chi trả cho những kênh Youtube làm đối tác với mình.
  - Nền tảng công nghệ thông tin hoạt động ổn định và được cải tiến liên tục (giao diện người dùng thân thiện, nhiều tính năng nâng cao trải nghiệm người dùng v.v.) do Youtube được công ty mẹ là Google hậu thuẫn.
- ⇒ Từ những yếu tố trên kết hợp với sự phát triển nền công nghiệp thiết bị kỹ thuật số đã thu hút đông đảo các cá nhân, tổ chức (công ty) tham gia sản xuất ra nhiều video chất lượng và gặt hái thành công, không chỉ dừng lại ở việc tăng thu nhập mà còn đem lại cho họ nhiều giá trị khác (danh tiếng, các mối quan hệ, phát triển thương hiệu v.v.).

Tiềm năng của Youtube có ý nghĩa đặc biệt với nhóm người đang có ý định trở thành các Youtuber, nhưng điều quan trọng hơn hết sau đó lại là cách thức hay phương hướng thực hiện. Vì không phải YouTuber nào cũng có những video lọt được vào top Trending YouTube, thậm chí nhiều YouTuber làm video lâu năm nhưng vẫn không đạt được kết quả. Do đó, nhóm hi vọng phần nào ý tưởng của nhóm sẽ trở thành một lời khuyên, gợi ý hữu ích cho các YouTuber sản xuất ra video chất lượng hơn, đạt được thành công lọt top trending. Bên cạnh đó, dựa vào những dữ liệu nhóm phân tích được, các doanh nghiệp cũng có thể xác định được xu hướng người xem tại Việt Nam, qua đó đưa ra các chiến lược Marketing phù hợp trên thị trường YouTube đầy tiềm năng này.

## **1.2. Mục tiêu đề tài**

Mục tiêu lớn nhất của đề tài là nghiên cứu, phát hiện ra những yếu tố, đặc điểm chung của các Video lọt top trending. Từ đó, trên những dữ liệu nhóm phân tích, thống kê có thể dựa vào để các doanh nghiệp tham khảo đưa ra các chiến lược marketing hoặc thiết lập bộ tiêu chí các Video lọt top trending.

Đề tài được thiết kế và phát triển đáp ứng các yêu cầu dưới đây:

- Áp dụng những kiến thức của môn học vào đề tài, vận dụng thành thực và hiểu rõ về các kĩ thuật được đưa vào đề án.
- 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 dưới dạng biểu đồ, biểu mẫu.
- Phát triển đề tài có tính thực tiễn, có khả năng triển khai và đưa vào ứng dụng thực tế, giúp người dùng dễ dàng sử dụng và đáp ứng nhu cầu.

### **1.3. Phạm vi và đối tượng nghiên cứu**

- Phạm vi nghiên cứu: YouTube API; Bộ các thư viện: Pandas, Numpy, Matplotlib, Seaborn, v.v.; Các phần mềm hỗ trợ: Google Colab, Ngôn ngữ Python, v.v.
- Đối tượng nghiên cứu: Video trong top trending trên YouTube: Thời lượng, thời gian, tiêu đề, nội dung, chủ đề, v.v.

### **1.4. Phương pháp nghiên cứu**

- Phương pháp thống kê, so sánh: Thống kê video đang trong top trending trên YouTube, so sánh sự khác biệt về nội dung, chủ đề, thời gian, v.v. của các video với nhau.
- Phương pháp phân tích, tổng hợp: Từ những dữ liệu về video trong top trending YouTube đã thống kê được, tiến hành phân tích và tổng hợp dữ liệu. Từ đó, rút ra những nhận xét, kết luận để xây dựng và hoàn thiện bộ tiêu chí. Đây là bước đi quan trọng, hình thành hướng đi chính và mục tiêu của đề tài.

### **1.5. Ý nghĩa đề tài**

- Đối với người dùng: Đề tài trên cơ sở nghiên cứu, phân tích dữ liệu từ những video trong top trending YouTube. Từ đó đưa ra những phân tích hữu ích biên chuẩn thành bộ tài liệu cho các doanh nghiệp nghiên cứu, ra quyết định cho việc tiếp cận khách hàng, xây dựng, quảng bá thương hiệu thông qua YouTube. Ngoài ra, dựa vào đó cũng có thể xây dựng thành bộ tiêu chí video

lọt top trending để các YouTuber tham khảo xây dựng, sáng tạo video của mình thêm tốt hơn, sớm đạt được thành tựu lọt top trending trên YouTube.

- Đối với cá nhân thành viên nhóm: Hiểu rõ hơn về Python, về cách lấy và phân tích dữ liệu để phục vụ cho công việc sau này.

## **1.6. Kết cấu đề tài**

Đề tài được chia làm 04 chương như sau:

**Chương 1: Tổng quan về đề tài.** Trình bày về tổng quan về đề tài bao gồm lý do hình thành đề tài, mục tiêu, phương pháp thực hiện, ý nghĩa thực tiễn và bố cục của đề tài.

**Chương 2: Cơ sở lý thuyết.** Trình bày cơ sở lý thuyết về khái niệm và tính năng ngôn ngữ lập trình Python được áp dụng để thực hiện đồ án, bộ thư viện Python: Matplotlib, Pandas, Numpy, v.v. và áp dụng YouTube API lấy dữ liệu từ YouTube.

**Chương 3: Triển khai phân tích dữ liệu.** Trình bày về quy trình thực hiện bao gồm: cách lấy dữ liệu, cách phân tích dữ liệu, cách mô hình hóa dữ liệu và kết quả thu được.

**Chương 4: Kết luận và đánh giá.** Tóm tắt lại nội dung, đánh giá kết quả của đề tài, trình bày những ưu điểm, nhược điểm và hướng phát triển của đề tài.

## CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

### 2.1. Ngôn ngữ lập trình (Python)

#### a) Định nghĩa Python

Python là một ngôn ngữ lập trình cấp cao, hướng đối tượng, được giải thích với ngữ nghĩa động. Với cấu trúc dữ liệu được xây dựng chính chu, kết hợp với kiểu gõ động và liên kết động, phù hợp cho việc phát triển ứng dụng nhanh, cũng như sử dụng như một ngôn ngữ kịch bản hoặc một ngôn ngữ liên kết để kết nối các thành phần hiện có với nhau. Cú pháp đơn giản, dễ học của Python nhấn mạnh khả năng đọc và do đó giảm chi phí bảo trì chương trình. Python hỗ trợ các module và gói, giúp cho việc module chương trình và tái sử dụng mã được dễ dàng hơn. Trình thông dịch và thư viện tiêu chuẩn mở rộng của Python có sẵn ở dạng nguồn hoặc dạng nhị phân, miễn phí cho tất cả các nền tảng phổ biến và có thể được phân phối tự do. [4]

#### b) Tính năng chính của Python [5]

🚦 Ngôn ngữ lập trình đơn giản, dễ học:

Python có cú pháp rất đơn giản, rõ ràng. Nó dễ đọc và viết hơn rất nhiều khi so sánh với những ngôn ngữ lập trình khác như C++, Java, C#. Python làm cho việc lập trình trở nên thú vị, cho phép bạn tập trung vào những giải pháp chứ không phải cú pháp.

🚦 Mã nguồn mở và miễn phí:

Bạn có thể tự do sử dụng và phân phối Python, thậm chí là dùng cho mục đích thương mại. Vì là mã nguồn mở, bạn không những có thể sử dụng các phần mềm, chương trình được viết trong Python mà còn có thể thay đổi mã nguồn của nó. Python có một cộng đồng rộng lớn, không ngừng cải thiện nó mỗi lần cập nhật.

🚦 Ngôn ngữ định hướng:

Một trong những tính năng chính của Python là lập trình hướng đối tượng. Python hỗ trợ ngôn ngữ hướng đối tượng và các khái niệm về các lớp, đóng gói đối tượng, v.v.

### 🚦 Hỗ trợ lập trình GUI:

Giao diện người dùng đồ họa có thể được thực hiện bằng cách sử dụng một module như PyQt5, PyQt4, wxPython hoặc Tk trong Python.

PyQt5 là tùy chọn phổ biến nhất để tạo các ứng dụng đồ họa với Python.

### 🚦 Ngôn ngữ cấp cao:

Python là ngôn ngữ cấp cao. Khi chúng ta viết chương trình bằng Python, chúng ta không cần phải nhớ kiến trúc hệ thống, cũng không cần phải quản lý bộ nhớ.

### 🚦 Tính năng mở rộng:

Python là một ngôn ngữ mở rộng. Chúng ta có thể viết một số mã Python sang ngôn ngữ C hoặc C++ và cũng có thể biên dịch mã đó bằng ngôn ngữ C/C++.

### 🚦 Ngôn ngữ di động:

Python cũng là ngôn ngữ di động.

Ví dụ: Nếu chúng ta có mã Python cho Windows và muốn chạy mã này trên nền tảng khác như Linux, Unix hay Mac, chúng ta không cần phải thay đổi mà vẫn có thể chạy mã này trên bất kỳ nền tảng nào.

### 🚦 Ngôn ngữ tích hợp:

Python cũng là một ngôn ngữ tích hợp vì chúng ta có thể dễ dàng tích hợp Python với các ngôn ngữ khác như C, C++, v.v.

### 🚦 Ngôn ngữ diễn giải:

Python là một ngôn ngữ diễn giải, bởi vì mã Python được thực thi từng dòng một. Giống như các ngôn ngữ khác C, C++, java, v.v., Python không cần phải biên dịch mã. Điều này giúp chúng ta dễ dàng gỡ lỗi mã của mình. Mã nguồn của Python được chuyển đổi thành một dạng mã byte ngay lập.

### 🚦 Thư viện tiêu chuẩn lớn

Python có một thư viện tiêu chuẩn lớn cung cấp bộ module và hàm phong phú để bạn không phải viết mã riêng cho từng thứ. Có rất nhiều thư viện hiện diện trong Python như các biểu thức thông thường, kiểm tra đơn vị, trình duyệt web, v.v.



🚦 Ngôn ngữ gõ động:

Python là ngôn ngữ gõ động. Điều đó có nghĩa là loại (ví dụ: int, double, long, v.v.) của một biến được quyết định vào thời gian chạy không phải trước. Bởi vì tính năng này, chúng ta không cần chỉ định loại biến.

## 2.2. Thư viện Python

### 2.2.1. Request

Đây là thư viện giúp người dùng có thể gửi các request HTTP/1.1 đơn giản nhất. Thư viện xử lý để giúp người dùng không phải thêm các chuỗi truy vấn (query strings) một cách thủ công vào các URLs, form-encode hoặc PUT và POST dữ liệu, thư viện hoạt động tốt với JSON Method.

### 2.2.2. BeautifulSoup

Beautiful Soup là một thư viện giúp bạn dễ dàng lấy thông tin từ các trang web. Nó nằm trên một trình phân tích cú pháp HTML hoặc XML, cung cấp các thành ngữ thuần Python (Pythonic) để lặp lại, tìm kiếm và sửa đổi cây phân tích cú pháp.

### 2.2.3. Pandas

Pandas là một gói thư viện viết bằng Python phổ biến cho khoa học dữ liệu và với lý do như: nó 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. Pandas kết hợp các tính năng tính toán mảng hiệu suất cao của NumPy với khả năng thao tác dữ liệu linh hoạt của bảng tính và cơ sở dữ liệu quan hệ (như SQL). Nó cung cấp chức năng lập chỉ mục chính xác để giúp dễ dàng định hình lại, cắt và trộn, thực hiện tổng hợp và chọn tập hợp dữ liệu.

Pandas là công cụ chính mà chúng ta sẽ sử dụng trong đồ án này để xử lý dữ liệu.

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

Hình 2.1: Cấu trúc DataFrame của bảng dữ liệu Khách hàng

#### 2.2.4. Numpy

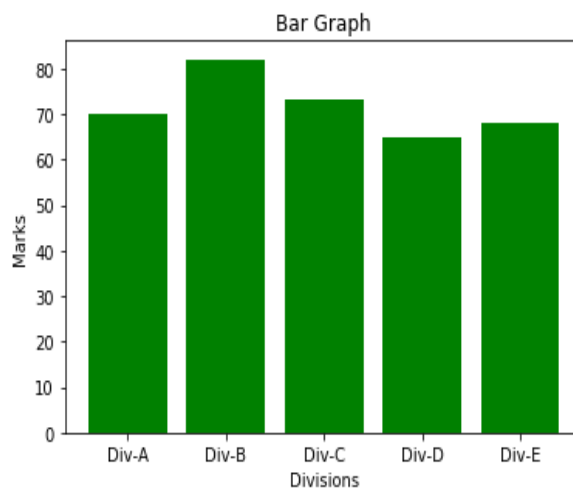
Numpy là một “core library” phục vụ cho khoa học máy tính của Python, hỗ trợ cho việc tính toán các mảng nhiều chiều, có kích thước lớn với các hàm đã được tối ưu áp dụng lên các mảng nhiều chiều đó. Numpy đặc biệt hữu ích khi thực hiện các hàm liên quan tới Đại Số Tuyến Tính.

#### 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. Nó là một thư viện vẽ đồ thị rất mạnh mẽ, hữu ích cho những người làm việc với Python và NumPy. Module được sử dụng nhiều nhất của Matplotlib là Pyplot. Pyplot cung cấp giao diện như MATLAB nhưng thay vào đó, nó sử dụng Python và nó là nguồn mở.

```
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()
```



Hình 2.2: Biểu đồ được tạo bởi thư viện Matplotlib

#### 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. Nó cung cấp một giao thức cấp cao để vẽ đồ họa thống kê đẹp và đầy đủ thông tin. [6]



Hình 2.3: Biểu đồ được tạo bởi thư viện Seaborn thể hiện dữ liệu

## 2.3. YouTube

### 2.3.1. Tổng quan

#### a) Định nghĩa 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. Ngoài ra, API được sử dụng khi lập trình các thành phần giao diện người dùng đồ họa (GUI). [7]

#### b) Ví dụ về API [7]

ProgramizableWeb, một trang web theo dõi hơn 7.500 API, liệt kê Google Maps, Twitter, YouTube, Flickr và Quảng cáo sản phẩm Amazon là một số API phổ biến nhất. Danh sách sau đây chứa một số ví dụ về các API phổ biến:

- API Google Maps: API Google Maps cho phép các nhà phát triển nhúng Google Maps trên các trang web bằng giao diện JavaScript

hoặc Flash. API Google Maps được thiết kế để hoạt động trên thiết bị di động và trình duyệt máy tính để bàn.

- API Flickr: API Flickr được các nhà phát triển sử dụng để truy cập dữ liệu cộng đồng chia sẻ ảnh Flickr. API Flickr bao gồm một tập hợp các phương thức có thể gọi và một số điểm cuối API.
- API Twitter: Twitter cung cấp hai API. API REST cho phép các nhà phát triển truy cập dữ liệu Twitter cốt lõi và API tìm kiếm cung cấp các phương thức để các nhà phát triển tương tác với dữ liệu Tìm kiếm và xu hướng của Twitter.
- API quảng cáo sản phẩm của Amazon: API quảng cáo sản phẩm của Amazon cho phép các nhà phát triển truy cập vào chức năng khám phá và lựa chọn sản phẩm của Amazon để quảng cáo các sản phẩm của Amazon để kiếm tiền từ một trang web.

c) Đặc điểm của API hiện đại [8]

Ngày trước, API thường được mô tả là giao diện kết nối chung cho một ứng dụng. Nhưng hiện nay, API hiện đại có một số đặc điểm làm cho chúng trở nên hữu ích và có giá trị hơn:

- Các API hiện đại tuân thủ các tiêu chuẩn (thường là HTTP và REST), có tính dễ sử dụng và dễ hiểu và thân thiện với các nhà phát triển.
- API được xử lý giống như sản phẩm hơn là code. Chúng được thiết kế cho các đối tượng cụ thể (ví dụ: API cho thiết bị di động...).
- Vì chúng được chuẩn hóa nhiều hơn nên tính bảo mật và quản trị mạnh hơn, cũng như được theo dõi và quản lý hiệu suất, quy mô tốt hơn.
- Như bất kỳ phần mềm sản phẩm nào khác, API hiện đại có chu kỳ phát triển phần mềm riêng của nó về thiết kế, thử nghiệm, xây dựng, quản lý.

d) Ưu và nhược điểm của API

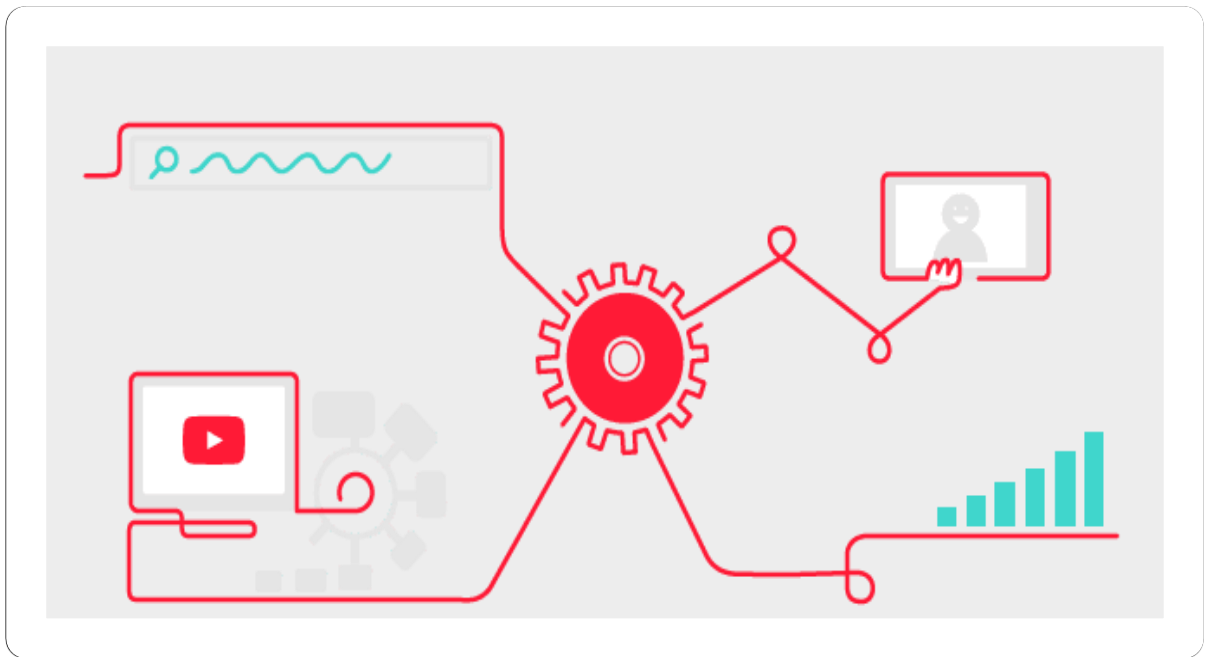
- Ưu điểm của API:

- Kết nối mọi lúc nhờ vào Internet.
- Giao tiếp hai chiều phải được xác nhận trong các giao dịch.

- Vì giao tiếp là API hai chiều nên thông tin rất đáng tin cậy.
- Cung cấp trải nghiệm thân thiện với người dùng.
- Cung cấp giải pháp phát triển khi các nhà phát triển tìm thấy cách sử dụng mới để trao đổi API.
- Cấu hình đơn giản khi được so sánh với WCF.
- Mã nguồn mở.
- Hỗ trợ chức năng RESTful một cách đầy đủ.
- Hỗ trợ đầy đủ các thành phần MVC như: routing, controller, action result, filter, model binder, IoC container, dependency injection, unit test.
- Khả năng trình diễn cao.
- Khuyết điểm của API:
  - Tốn nhiều chi phí phát triển, vận hành, chỉnh sửa.
  - Đòi hỏi kiến thức chuyên sâu.
  - Có thể gặp vấn đề bảo mật khi bị tấn công hệ thống.

### **2.3.2. 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 video và dữ liệu kênh YouTube thông qua hai loại: REST và XML-RPC. Google mô tả Tài nguyên API của YouTube là "API và Công cụ cho phép bạn mang trải nghiệm YouTube đến trang web, ứng dụng hoặc thiết bị của bạn". API YouTube bao gồm API YouTube Analytics, API dữ liệu YouTube, API phát trực tiếp YouTube, API trình phát YouTube và các API khác.



*Hình 2.4: Youtube API áp dụng cho nhiều nền tảng khác nhau*

## CHƯƠNG 3: TRIỂN KHAI PHÂN TÍCH DỮ LIỆU

### 3.1. Quy trình phân tích

Trước khi đi vào phân tích, chúng tôi sẽ mô tả môi trường thực nghiệm mà nhóm đã thực hiện:

Hệ điều hành	Windows 10 Education, Mac OS.
Vi xử lý	Intel Core i5, 2.70 GHz
Ram	8.00 GB
Tốc độ mạng trung bình	50.0 Mbps
Ngôn ngữ khai phá dữ liệu	Python
Miền dữ liệu	Dữ liệu Youtube Trending với các thông số liên quan về Video thu thập.
Nguồn dữ liệu	Youtube.com
Tổng số Video thu thập và trích lọc được	1572

Bảng 3-1. Mô tả môi trường thực nghiệm

#### 3.1.1. Đặt vấn đề

Youtube đã và dần trở thành một kênh mạng xã hội chia sẻ Video trực tuyến lớn nhất hành tinh, cùng với đó là việc xuất hiện người người muốn tiếp cận với Youtube để trở thành nhà sáng tạo nội dung, hoặc một doanh nghiệp muốn tìm hiểu thị trường này nhằm có những chiến dịch quảng cáo, tiếp thị tối ưu nhất. Nhận thấy những vấn đề cấp thiết đó, nhóm đã xây dựng mẫu một phương thức phân tích cơ bản dành cho các Video thuộc Youtube Trending Việt Nam, đây là căn cứ và cơ sở giúp cho các cá nhân, tổ chức tiếp cận thị trường này dễ dàng hơn.

#### 3.1.2. Phân tích dữ liệu

- a) Lấy dữ liệu từ Youtube thông qua Youtube API bằng Python và làm sạch dữ liệu (Data Cleaning)

Google cung cấp cho chúng ta bốn API để phân tích từ Youtube, trong phạm vi bài nghiên cứu này, nhóm chỉ sử dụng YouTube Data API v3 để thu thập dữ liệu từ các kênh:

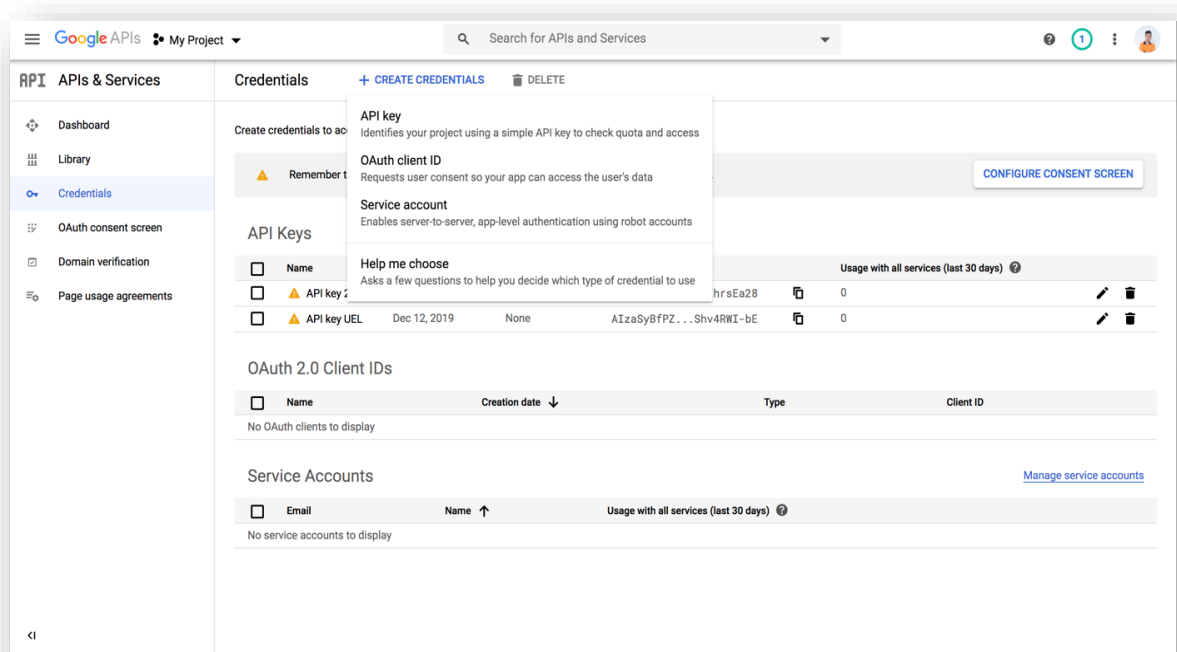
YouTube Analytics API	Retrieves your YouTube Analytics data.
YouTube Data API v3	Supports core YouTube features, such as uploading videos, creating and managing playlists, searching for content, and much more.
YouTube Live API v3	Supports core YouTube features, such as uploading videos, creating and managing playlists, searching for content, and much more.
YouTube Reporting API	Schedules reporting jobs containing your YouTube Analytics data and downloads the resulting bulk data reports in the form of CSV files.

Hình 3.1: Youtube API áp dụng cho nhiều nền tảng khác nhau

Để tiến hành thu thập, ta tiến hành qua các bước sau:

**Bước 1:** Xin Google cấp API để ta có thể tương tác với Youtube API:

- Vào trang: <https://console.developers.google.com/>
- Nhấn vào **Create Credentials -> API Key**



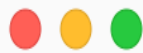
Hình 3.2: Trang lấy API Key của Google

- Sau khi tạo API Key, cửa sổ sẽ xuất hiện API mà Google cung cấp cho bạn, lưu ý là phải tuyệt đối bảo mật Key này để tránh các sự cố về bảo



mật. API Key này có thể coi như là chiếc “chìa khoá” để bạn có thể tương tác với Youtube Data API v3

**Bước 2:** Import các thư viện request, sys, time, os, argparse



```
import requests, sys, time, os, argparse
```

*Hình 3.3: Code import thư viện dùng để Crawling dữ liệu*

**Bước 3:** Khai báo một số Parameters để tương tác với các trường dữ liệu mà ta muốn Youtube API trả về:

	Các thuộc tính	Giải thích
	video_id	Mã của video đăng trên Youtube
	title	Tiêu đề của video
	publishedAt	Thời gian video được tải lên Youtube
	channelId	Mã của kênh Youtube đăng tải video
	channelTitle	Tên của kênh Youtube đăng tải video
	categoryId	Mã phân loại thể loại video được quy định bởi Youtube
	trending_date	Ngày video đạt độ phổ biến trên youtube
	tags	Từ khóa liên quan đến video
	view_count	Số lượt người xem video
	likes	Số lượt thích video
	dislikes	Số lượt không thích video
	comment_count	Tổng số bình luận video
	thumbnail_link	Liên kết hình ảnh trình bày video
	comments_disabled	Video có tắt bình luận hay không
	ratings_disabled	Video có tắt đánh giá hay không
	description	Mô tả video

Hình 3.4: Các trường dữ liệu mà ta muốn lấy từ Youtube API

```

snippet_features = ["title",
                    "publishedAt",
                    "channelId",
                    "channelTitle",
                    "categoryId"]

# Any characters to exclude, generally these are things that become problematic in CSV files
unsafe_characters = ['\n', '"]

# Used to identify columns, currently hardcoded order
header = ["video_id"] + snippet_features + ["trending_date", "tags", "view_count", "likes", "dislikes",
                                           "comment_count", "thumbnail_link", "comments_disabled",
                                           "ratings_disabled", "description"]

```

Hình 3.5: Khai báo các header để tương tác Youtube API

**Bước 4:** Ta xây dựng hai hàm **setup** và **prepare\_feature** để truyền API Key vào chuẩn bị cho việc tương tác dữ liệu

```
def setup(api_path, code_path):
    with open(api_path, 'r') as file:
        api_key = file.readline()

    with open(code_path) as file:
        country_codes = [x.rstrip() for x in file]

    return api_key, country_codes

def prepare_feature(feature):
    # Removes any character from the unsafe characters list and surrounds the whole item in quotes
    for ch in unsafe_characters:
        feature = str(feature).replace(ch, "")
    return f'"{feature}"'
```

*Hình 3.6: Code thể hiện cho hàm setup và hàm prepare\_feature*

**Bước 5:** Gửi request đến Youtube để lấy dữ liệu về thông qua hàm **api\_request**

```
def api_request(page_token, country_code):
    # Builds the URL and requests the JSON from it
    request_url = f"https://www.googleapis.com/youtube/v3/videos"
    part=id,statistics,snippet{page_token}chart=mostPopular&regionCode={country_code}&maxResults=50&key={api_key}"
    request = requests.get(request_url)
    if request.status_code == 429:
        print("Temp-Banned due to excess requests, please wait and continue later")
        sys.exit()
    return request.json()
```

*Hình 3.7: Code thể hiện cho hàm api\_request*

## Bước 6: Xây dựng hai hàm `get_videos` và `get_pages` để lấy cụ thể dữ liệu cho từng Video

```
def get_pages(country_code, next_page_token="&"):
    country_data = []

    # Because the API uses page tokens (which are literally just the same function of numbers
    # everywhere) it is much
    # more inconvenient to iterate over pages, but that is what is done here.
    while next_page_token is not None:
        # A page of data i.e. a list of videos and all needed data
        video_data_page = api_request(next_page_token, country_code)

        # Get the next page token and build a string which can be injected into the request with it,
        # unless it's None,
        # then let the whole thing be None so that the loop ends after this cycle
        next_page_token = video_data_page.get("nextPageToken", None)
        next_page_token = f"&pageToken={next_page_token}&" if next_page_token is not None else
        next_page_token

        # Get all of the items as a list and let get_videos return the needed features
        items = video_data_page.get('items', [])
        country_data += get_videos(items)

    return country_data
```

*Hình 3.8: Hàm `get_pages`*

```

def get_videos(items):
    lines = []
    for video in items:
        comments_disabled = False
        ratings_disabled = False

        # We can assume something is wrong with the video if it has no statistics, often this means it
        # has been deleted
        # so we can just skip it
        if "statistics" not in video:
            continue

        # A full explanation of all of these features can be found on the GitHub page for this project
        video_id = prepare_feature(video['id'])

        # Snippet and statistics are sub-dicts of video, containing the most useful info
        snippet = video['snippet']
        statistics = video['statistics']

        # This list contains all of the features in snippet that are 1 deep and require no special
        # processing
        features = [prepare_feature(snippet.get(feature, "")) for feature in snippet_features]

        # The following are special case features which require unique processing, or are not within
        # the snippet dict
        description = snippet.get("description", "")
        thumbnail_link = snippet.get("thumbnails", dict()).get("default", dict()).get("url", "")
        trending_date = time.strftime("%y.%d.%m")
        tags = get_tags(snippet.get("tags", ["[none]"]))
        view_count = statistics.get("viewCount", 0)

        # This may be unclear, essentially the way the API works is that if a video has comments or
        # ratings disabled
        # then it has no feature for it, thus if they don't exist in the statistics dict we know they
        # are disabled
        if 'likeCount' in statistics and 'dislikeCount' in statistics:
            likes = statistics['likeCount']
            dislikes = statistics['dislikeCount']
        else:
            ratings_disabled = True
            likes = 0
            dislikes = 0

        if 'commentCount' in statistics:
            comment_count = statistics['commentCount']
        else:
            comments_disabled = True
            comment_count = 0

        # Compiles all of the various bits of info into one consistently formatted line
        line = [video_id] + features + [prepare_feature(x) for x in [trending_date, tags, view_count,
likes, dislikes,
comment_count, thumbnail_link,
comments_disabled,
ratings_disabled, description]]

        lines.append(", ".join(line))
    return lines

```

Hình 3.9: Hàm `get_videos`

## Bước 7: Ghi file vừa nhận được về vào file csv

```
def write_to_file(country_code, country_data):  
    print(f"Writing {country_code} data to file...")  
  
    if not os.path.exists(output_dir):  
        os.makedirs(output_dir)  
  
    with open(f"{output_dir}/{time.strftime('%y.%d.%m')}_{country_code}_videos.csv", "w+",  
            encoding='utf-8') as file:  
        for row in country_data:  
            file.write(f"{row}\n")  
  
def get_data():  
    for country_code in country_codes:  
        country_data = [".join(header)] + get_pages(country_code)  
        write_to_file(country_code, country_data)
```

Hình 3.10: Lưu dữ liệu vào file CSV

⇒ Dữ liệu thu được sẽ có dạng như sau:

	video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date	tags	view_count
0	WAxxfzdcNdA	em bỏ hút thuốc chưa - người yêu cũ nhân 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 Dân ông dờ nhất là phản bội, còn em không s...	2020-05- 22T14:00:11Z	UC2Lgi2uPsOcCVF3imz7I2mg	Vie GIẢI TRÍ	24	20.26.05	vie giai tri giai tri tv người ấy là ai người ...	5148975
2	ayJY9ieBuEU	KHÔNG THỂ CÙNG NHAU SUỐT KIẾP - HOÀ MINZY (ft....	2020-05- 13T13:00:10Z	UCjm_FW7t1gam7qLIdSVOclw	Hòa Minzy	10	20.26.05	hoa minzy hòa minzy hòa minzy 2020 không thể c...	21413308
3	GKTpUGkhvig	FAPtv Com Người: Tập 220 - Làng Nhâm Nhí	2020-05- 23T12:58:15Z	UC0jDoh3tVXCaqJ6oTve8ebA	FAP TV	1	20.26.05	FAPTV faptv faptivi FAPTivi Faptv com nguoi c...	3681043
4	qGJAWJ2zWWI	Agust D '대취타' MV	2020-05- 22T09:00:01Z	UC3lZKseVpdzPSBaWxBxundA	Big Hit Labels	10	20.26.05	BIGHIT 빅히트 방탄소년 단 BTS BANGTAN 방탄 Agust D  어거스트 디...	44924142

Hình 3.11: Minh họa dữ liệu thu được

### b) Làm sạch dữ liệu và phân tích

**Bước 1:** Ta tiến hành cài đặt các thư viện cần thiết để thực hiện đề tài như:

Pandas, Numpy, Matplotlib, Seaborn, v.v

```
import pandas as pd
import numpy as np
import matplotlib as mpl
from matplotlib import pyplot as plt
import seaborn as sns

import warnings
from collections import Counter
import datetime
import wordcloud
import json
```

Hình 3.12: Cài đặt các thư viện

```
# Hiding warnings for cleaner display
warnings.filterwarnings('ignore')

# Configuring some options
%matplotlib inline
%config InlineBackend.figure_format = 'retina'
# If you want interactive plots, uncomment the next line
# %matplotlib notebook
```

Hình 3.13: Cấu hình cách thể hiện dữ liệu

**Bước 2:** Ta sẽ đọc dữ liệu dưới file CSV và đưa vào Dataframe của Pandas

```
df = pd.read_csv("/Users/anhnhath/Documents/Uel/SepVI_1920/Phan_tich_du_lieu_Web/Final/VN_data.csv",lineterminator='\n')
```

Hình 3.14: Đọc dữ liệu bằng file CSV

```
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, labels=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)
```

Hình 3.15: Đưa dữ liệu vào Dataframe của Pandas

### Bước 3: Mô tả tập dữ liệu bằng bảng

```
df.head()
```

Hình 3.16: Dòng lệnh để đưa dữ liệu thành bảng



	video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date	tags	view_count
0	WAxxfzdcNdA	em bỏ hút thuốc chưa - người yếu cũ nhân tin h...	2020-05- 17T16:15:11Z	UC90LfAFYhRLh86Qd-Fs4zg	BICH PHUONG	10	20.26.05	em bỏ hút thuốc chưa anh bỏ hút thuốc chưa em ...	12541198
1	8mltWlx3cs0	#3 Đản ông đồ nhất là phần bội, còn em không s...	2020-05- 22T14:00:11Z	UC2Lgi2uPsOcCVF3imz7I2mg	Vie GIẢITRÍ	24	20.26.05	vie giai tri giai tri tv người ấy là ai người ...	5148975
2	ayJY9ieBuEU	KHÔNG THỂ CÙNG NHAU SUỐT KIẾP - HOÀ MINZY (ft....	2020-05- 13T13:00:10Z	UCjm_FW7t1gam7qLldSVOclw	Hòa Minzy	10	20.26.05	hoa minzy hòa minzy hòa minzy 2020 không thể c...	21413308
3	GKTpUGkhvig	FAPtv Com Người: Tập 220 - Lãng Nhảm Nhĩ	2020-05- 23T12:58:15Z	UC0jDoh3tVXCaqJ6oTve8ebA	FAP TV	1	20.26.05	FAPTV faptv faptvii FAPtivii Faptv com người c...	3681043
4	qGjAWJ2zWWI	Agust D '대쉬타' MV	2020-05- 22T09:00:01Z	UC3lZKseVpdzPSBaWxBxundA	Big Hit Labels	10	20.26.05	BIGHIT 빅히트 방탄소년 단 BTS BANGTAN 방탄 Agust D  어거스트 디...	44924142

Hình 3.17: Các dòng và cột dữ liệu video xu hướng trên YouTube



count	likes	dislikes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	description	filename
41198	329323	10193	33498	<a href="https://i.ytimg.com/vi/WAxxfzdcNdA/default.jpg">https://i.ytimg.com/vi/WAxxfzdcNdA/default.jpg</a>	False	False	Bích Phương - em bỏ hũ thuốc chưa? (feat. Trai...)	data_VN/VN01.csv
48975	47047	3533	6017	<a href="https://i.ytimg.com/vi/8mltWlx3cs0/default.jpg">https://i.ytimg.com/vi/8mltWlx3cs0/default.jpg</a>	False	False	#Người Ấy Là Ai #Người Ấy Là Ai Mua 3 #VieChannelHTV2 ...	data_VN/VN01.csv
13308	637643	20295	52810	<a href="https://i.ytimg.com/vi/ayJY9ieBuEU/default.jpg">https://i.ytimg.com/vi/ayJY9ieBuEU/default.jpg</a>	False	False	KHÔNG THỂ CÙNG NHAU SUỐT KIẾP - HOÀ MINZY (ft....)	data_VN/VN01.csv
81043	92551	3305	2970	<a href="https://i.ytimg.com/vi/GKTPUGkhvig/default.jpg">https://i.ytimg.com/vi/GKTPUGkhvig/default.jpg</a>	False	False	FAPtv Com Người Tập 220 - Lăng Nhâm Nhí Sân X...	data_VN/VN01.csv
24142	5752355	99816	743758	<a href="https://i.ytimg.com/vi/qGjAWJ2zWWI/default.jpg">https://i.ytimg.com/vi/qGjAWJ2zWWI/default.jpg</a>	False	False	Agust D '대작' MV Agust D - 'D-2' DownloadGoogl...	data_VN/VN01.csv

Hình 3.18: Các dòng và cột dữ liệu video xu hướng trên Youtube

	Các thuộc tính	Giải thích
	video_id	Mã của video đăng trên Youtube
	title	Tiêu đề của video
	publishedAt	Thời gian video được tải lên Youtube
	channelId	Mã của kênh Youtube đăng tải video
	channelTitle	Tên của kênh Youtube đăng tải video
	categoryId	Mã phân loại thể loại video được quy định bởi Youtube
	trending_date	Ngày video đạt độ phổ biến trên youtube
	tags	Từ khóa liên quan đến video
	view_count	Số lượt người xem video
	likes	Số lượt thích video
	dislikes	Số lượt không thích video
	comment_count	Tổng số bình luận video
	thumbnail_link	Liên kết hình ảnh trình bày video
	comments_disabled	Video có tắt bình luận hay không
	ratings_disabled	Video có tắt đánh giá hay không
	description	Mô tả video

Hình 3.19: Các thuộc tính của các dòng dữ liệu

#### Bước 4: Thông tin hóa dữ liệu và làm sạch dữ liệu

Sử dụng các dòng lệnh để có thể thông tin hóa dữ liệu đang sử dụng có những thuộc tính gì, chúng ta tiến hành các thao tác:

```
df.info()
```

Hình 3.20: Dòng lệnh lấy thông tin dữ liệu

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   video_id              200 non-null   object
1   title                 200 non-null   object
2   publishedAt           200 non-null   object
3   channelId             200 non-null   object
4   channelTitle          200 non-null   object
5   categoryId            200 non-null   int64
6   trending_date         200 non-null   object
7   tags                  200 non-null   object
8   view_count            200 non-null   int64
9   likes                 200 non-null   int64
10  dislikes              200 non-null   int64
11  comment_count         200 non-null   int64
12  thumbnail_link        200 non-null   object
13  comments_disabled     200 non-null   bool
14  ratings_disabled      200 non-null   bool
15  description           199 non-null   object
dtypes: bool(2), int64(5), object(9)
memory usage: 22.4+ KB
```

Hình 3.21: Kết quả sau khi lấy thông tin của dữ liệu

Từ đây chúng ta có thể nhận thấy được tất cả các thuộc tính trong tập dữ liệu đều chứa dữ liệu không rỗng ngoại trừ thuộc tính “description” có chứa 1 giá trị rỗng. Để giúp tập dữ liệu được rõ ràng hơn, chúng ta sẽ tiến hành làm sạch dữ liệu qua các bước:

```
df[df["description"].apply(lambda x: pd.isna(x))].head(3)
```

Hình 3.22: Dòng lệnh giúp làm sạch dữ liệu

trending_date	tags	view_count	likes	dislikes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	description
20.26.05	[none]	93197	1740	170	62	https://i.ytimg.com/vi/JAgsJJnOZzQ/default.jpg	False	False	NaN

Hình 3.23: Kết quả sau khi làm sạch dữ liệu

Sau khi làm sạch dữ liệu từ tập dữ liệu, chúng ta dễ dàng nhận ra được 1 dòng dữ liệu có chứa giá trị rỗng với thuộc tính “description” được biểu thị bằng chữ NaN.

### 3.1.3. Phân tích kết quả và trực quan hoá dữ liệu

Sau khi đã thu thập và làm sạch dữ liệu, chúng ta sẽ tiến hành phân tích những kết quả với những dòng lệnh thích hợp. Từ đó, ta có thể trực quan hóa kết quả thành các biểu đồ, đồ thị để có sự nhìn nhận khách quan hơn.

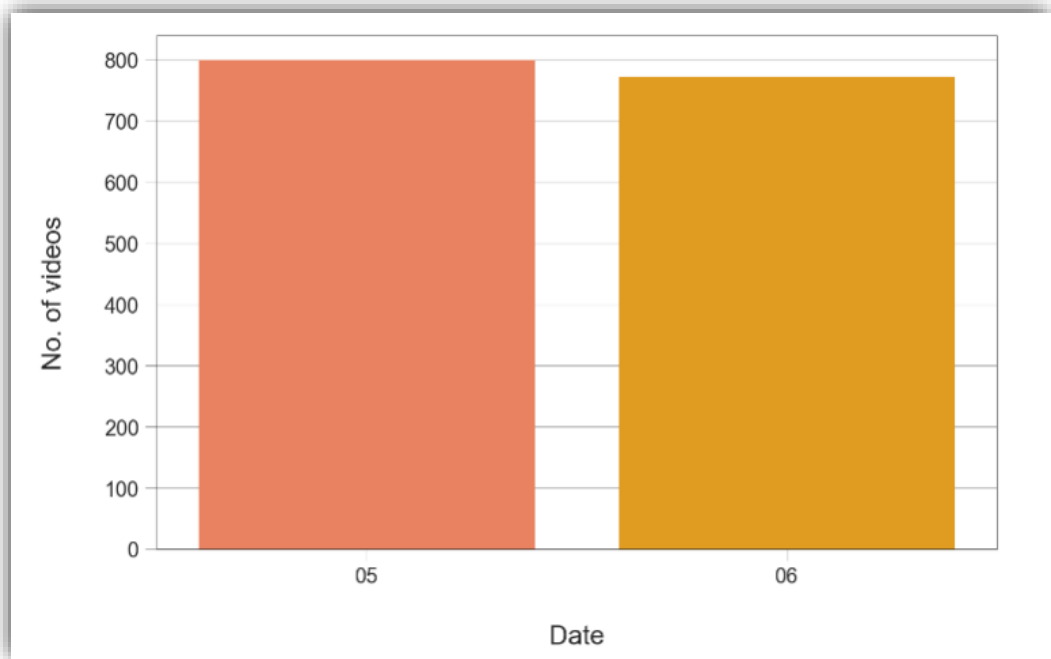
#### **Thứ nhất: Thống kê ngày lấy dữ liệu**

Qua các dòng lệnh, chúng ta có thể biết được dữ liệu được thu thập ở ngày cụ thể (dữ liệu được thu thập qua nhiều ngày). Từ đó, thể hiện mốc thời gian đã thu thập qua một biểu đồ cột.

```
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")
```

Hình 3.24: Dòng lệnh nhận biết thời gian dữ liệu thu thập



Hình 3.25: Biểu đồ cột của ngày thu thập dữ liệu

### Thứ hai: Biểu thị một số giá trị thống kê

Tiếp theo, ta tiến hành thể hiện một số giá trị của các thuộc tính có dữ liệu dạng số liệu. Từ đó, bắt đầu phân tích các số liệu nhận được và đưa ra kết luận.

```
df.describe()
```

Hình 3.26: Dòng lệnh giúp biểu thị giá trị thống kê

	categoryId	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

Hình 3.27: Kết quả sau khi thu thập giá trị thống kê

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

- Số lượt views trung bình của Video Trending 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 giá trị này, và nửa còn lại sẽ nhỏ hơn giá trị 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 Trending 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, video có lượt views nhỏ nhất là 39.129.

### ***Thứ ba: Biểu thị những video thịnh hành***

Chúng ta tiến hành lọc và hiển thị những video đang thịnh hành từ lượt views thấp nhất cho đến video thịnh hành có lượt views cao nhất qua các dòng lệnh.

```
from IPython.display import HTML, display

# We choose the 10 most trending videos
selected_columns = ['title', 'channelTitle', 'thumbnail_link', 'publishedAt', 'view_count']
maxVid=df[df['view_count']==df['view_count'].max()]
minVid=df[df['view_count']==df['view_count'].min()]
getMarginVid=maxVid.append(minVid, ignore_index=True)
# Construction of HTML table with miniature photos assigned to the most popular movies
table_content = ''
max_title_length = 50
```

*Hình 3.28: Dòng lệnh chọn những video có lượt views từ cao đến thấp*

```
for date, row in getMarginVid.T.iteritems():
    HTML_row = '<tr>'
    HTML_row += '<td></td>'
    HTML_row += '<td>' + str(row[4]) + '</td>'
    HTML_row += '<td>' + str(row[1]) + '</td>'
    HTML_row += '<td>' + str(row[8]) + '</td>'
    HTML_row += '<td>' + str(row[2]) + '</td>'

    table_content += HTML_row + '</tr>'

display(HTML(
    '<table><tr><th>Photo</th><th>Channel Name</th><th style="width:250px;">Title</th><th>Views</th><th>
```

*Hình 3.29: Dòng lệnh hình thành bảng biểu thị video thịnh hành*

Photo	Channel Name	Title	Views	Publish Date
	LadyGagaVEVO	Lady Gaga, Ariana Grande - Rain On Me (Official Music Video)	89178790	2020-05-22T17:00:09Z
	Hoài Phong Organ	Karaoke Liên Khúc Nhạc Sống Trữ Tình Rumba Tone Nữ   Đôi Mắt Người Xưa   Chiều Hạ Vàng	39129	2020-05-28T04:35:04Z

Hình 3.30: Bảng biểu thị những video thịnh hành

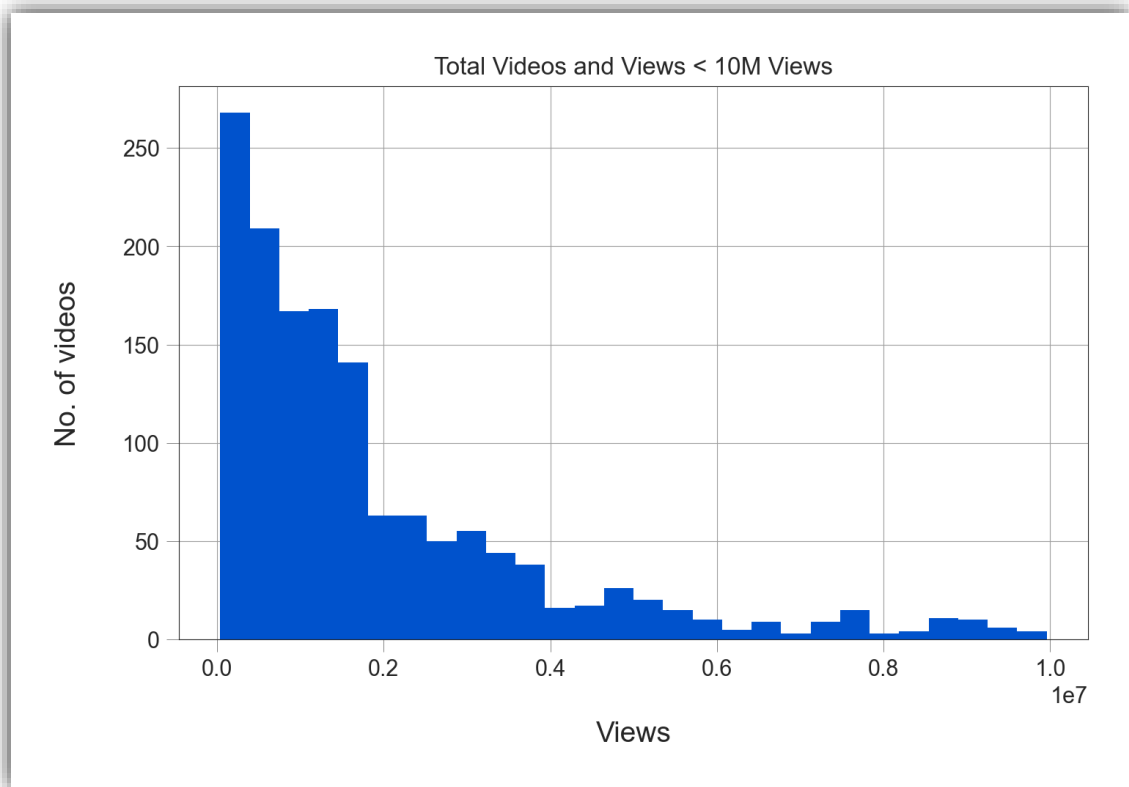
Từ dữ liệu ở bảng trên, ta nhận thấy video thịnh hành có lượt views cao nhất là 89.178.790 và lượt views thấp nhất là 39.129.

#### **Thứ tư: Biểu thị lượt xem**

Ta bắt đầu phân tích lượt xem của người dùng Youtube trong đó có bao nhiêu số lượng video có lượt views trên và dưới 10 triệu lượt. Từ đó, trực quan hóa dữ liệu thành biểu đồ cột và biểu đồ tròn.

```
fig, ax = plt.subplots()
_ = sns.distplot(df[df["view_count"] < 10e6]["view_count"], kde=False,
                 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")
```

Hình 3.31: Dòng lệnh biểu thị lượt views bằng biểu đồ cột



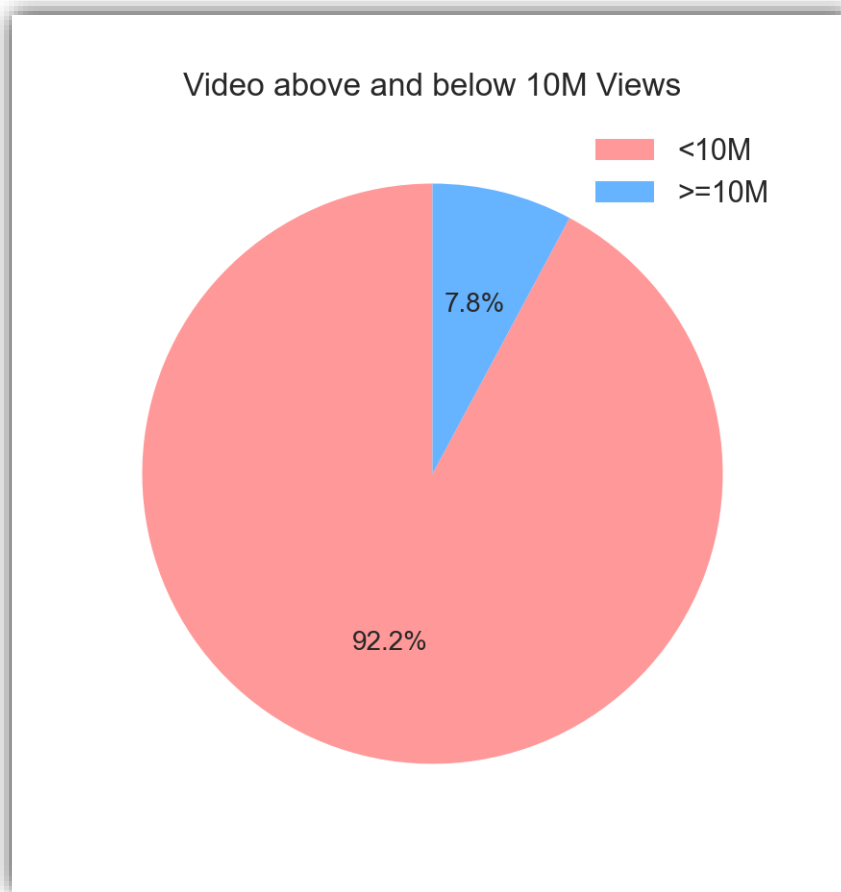
Hình 3.32: Biểu đồ cột hiển thị số lượng video có lượt views dưới 10 triệu

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 liệu

```
max_views=df[df['view_count'] < 10e6]['view_count'].count() / df['view_count'].count() * 100
min_views=df[df['view_count'] >= 10e6]['view_count'].count() / df['view_count'].count() * 100
df_views=df2 = pd.DataFrame([[max_views,min_views]], columns=['max_views','min_views'])

plt.pie(df_views,autopct='%0.1f%%',startangle=90, colors=['#ff9999','#66b3ff'])
plt.title("Video above and below 10M Views")
plt.legend(["<10M", ">=10M"])
```

Hình 3.33: Dòng lệnh biểu thị lượt views bằng biểu đồ tròn



Hình 3.34: Biểu đồ tròn hiển thị số lượng video có lượt views dưới và trên 10 triệu

Từ phân tích có thể thấy, gần 92.2% videos dưới 10 triệu Views, chỉ có khoảng 7.8% videos có trên 10 triệu Views.

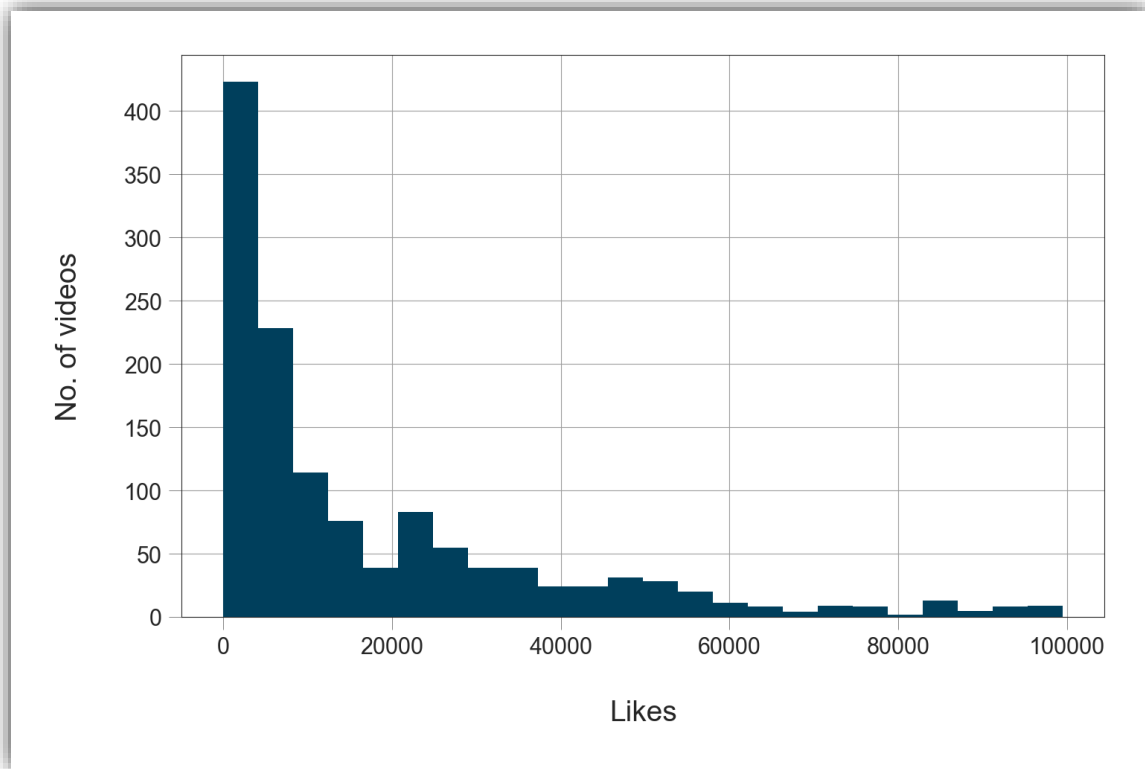
**Thứ năm: Biểu thị lượt thích**

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

```
fig, ax = plt.subplots()
_ = sns.distplot(df[df["likes"] <= 1e5]["likes"], kde=False,
                 color=PLOT_COLORS[4], hist_kws={'alpha': 1}, ax=ax)
_ = ax.set(xlabel="Likes", ylabel="No. of videos")
```

Hình 3.35: Dòng lệnh biểu thị lượt likes bằng biểu đồ cột





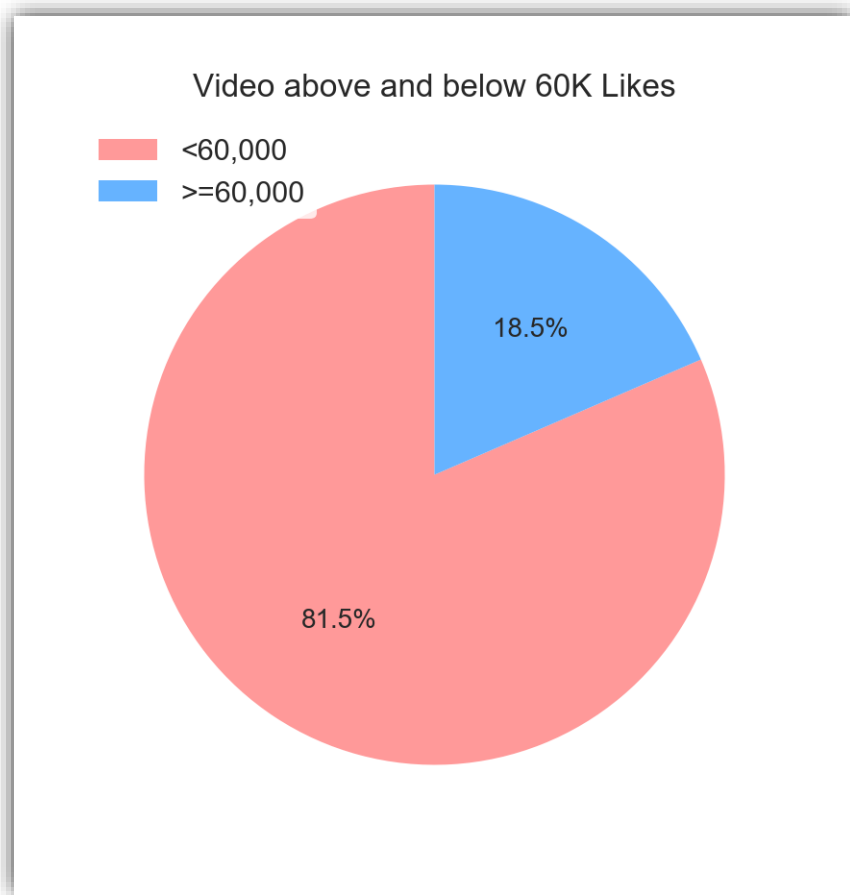
Hình 3.36: Biểu đồ cột hiển thị số lượng likes dưới 100.000

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.

```
max_likes=df[df['likes'] < 6e4]['likes'].count() / df['likes'].count() * 100
min_likes=df[df['likes'] >= 6e4]['comment_count'].count() / df['likes'].count() * 100
df_likes = pd.DataFrame([[max_likes,min_likes]], columns=['max_likes','min_likes'])

plt.pie(df_likes,autopct='%0.1f%%',startangle=90, colors=['#ff9999','#66b3ff'])
plt.title("Video above and below 10M Likes")
plt.legend(["<60,000", ">=60,000"])
```

Hình 3.37: Dòng lệnh biểu thị lượt likes bằng biểu đồ tròn



Hình 3.38: Biểu đồ tròn hiển thị số lượng video có lượt likes trên và dưới 60.000

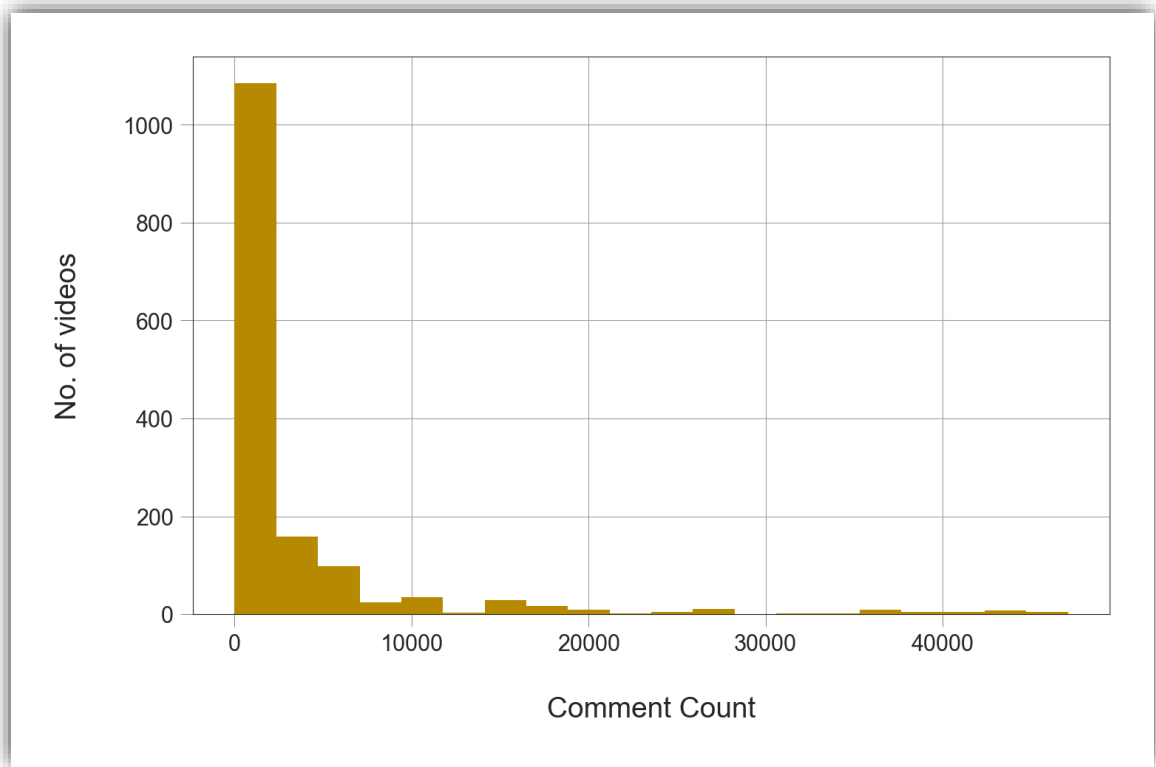
Đúng như ta dự đoán, có tới 78,2% videos có dưới 60,000 likes, và 21,8% videos có trên 60,000 likes.

#### **Thứ sáu: Biểu thị lượt bình luận**

Chúng ta cùng nhìn qua sự phân bố của lượt comment của các videos nằm trong khoảng từ 0 đến 40.000 comment bằng biểu đồ cột.

```
fig, ax = plt.subplots()
_ = sns.distplot(df[df["comment_count"] < 50000]["comment_count"], kde=False, rug=False,
                 color=PLOT_COLORS[3], hist_kws={'alpha': 1},
                 bins=20, ax=ax)
_ = ax.set(xlabel="Comment Count", ylabel="No. of videos")
```

Hình 3.39: Dòng lệnh biểu thị lượt comment bằng biểu đồ cột



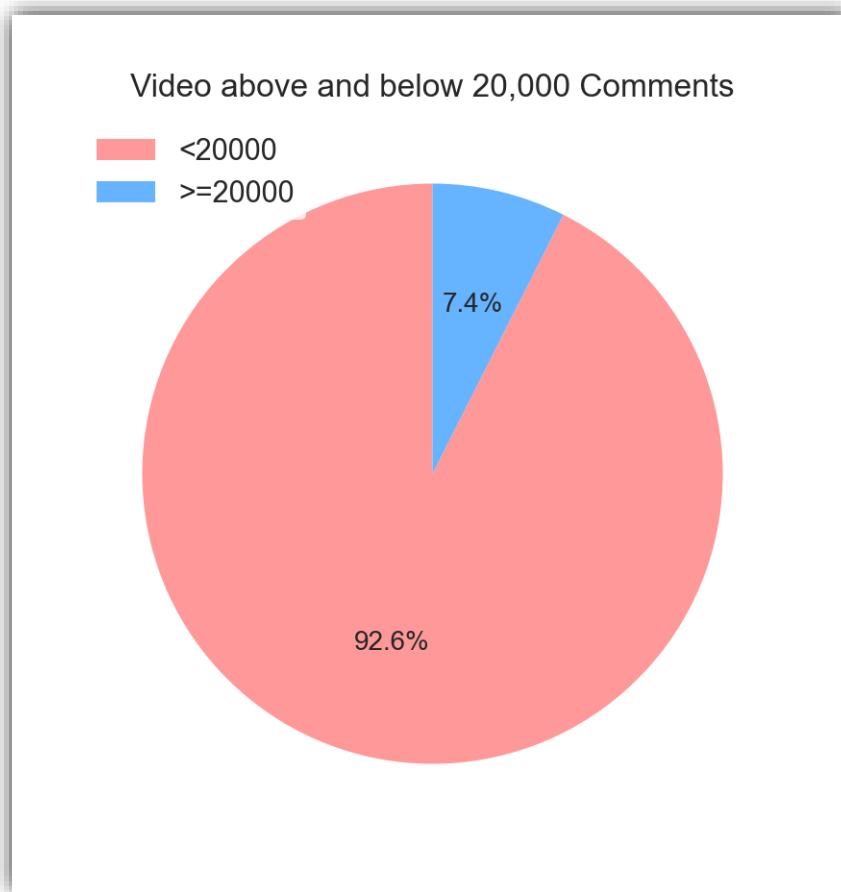
Hình 3.40: Biểu đồ cột hiển thị số lượng comment dưới 40.000

Chúng ta cùng nhìn qua tỉ lệ phân bố của lượt comment của các videos nằm trong khoảng từ 0 đến 20.000 comment bằng biểu đồ tròn.

```
max_likes=df[df['comment_count'] < 20000]['comment_count'].count() / df['comment_count'].count() * 100
min_likes=df[df['comment_count'] >= 20000]['comment_count'].count() / df['comment_count'].count() * 100
df_likes = pd.DataFrame([[max_likes,min_likes]], columns=['max_likes','min_likes'])

plt.pie(df_likes,autopct='%0.1f%',startangle=90, colors=['#ff9999','#66b3ff'])
plt.title("Video above and below 20,000 Comments")
plt.legend(["<20000", ">=20000"])
```

Hình 3.41: Dòng lệnh biểu thị lượt likes bằng biểu đồ tròn



Hình 3.42: Biểu đồ tròn hiển thị số lượng video có lượt comment trên và dưới 20.000

### Thứ bảy: Tìm hiểu về các cột không phải dạng số

Sau khi tìm hiểu các cột dữ liệu dạng số, ta sẽ tìm hiểu các phần dữ liệu không phải là số. Đầu tiên, ta thực hiện dòng lệnh:

```
df.describe(include = ['O'])
```

Hình 3.43: Dòng lệnh hiển thị tập dữ liệu

	video_id	title	publishedAt	channelId	channelTitle	trending_date	tags	thumbnail_link	description	filename	category
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)	2020-05-06T12:00:13Z	UCH-NjZ3ojREOWBZL3pYLA	WeTV Vietnam	20.01.06	[none]	https://i.ytimg.com/vi/Zsx-4qsOEks/default.jpg	Ban 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	

Hình 3.44: Kết quả của bảng dữ liệu không phải dạng số

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 ngày.

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 duy nhất. Tuy nhiên ta có thể thấy có tận 436 title, chứng tỏ 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.

```
grouped = df.groupby("video_id")
groups = []
wanted_groups = []
for key, item in grouped:
    groups.append(grouped.get_group(key))

for g in groups:
    if len(g['title'].unique()) != 1:
        wanted_groups.append(g)

wanted_groups[0]
example_title_df= pd.DataFrame(wanted_groups[0])
example_title_df.to_csv('example_title_df.csv', index=False)
```

Hình 3.45: Dòng lệnh hiển thị video title đổi tên

	A	B	C	D	E	F
1	video_id	title	publishedAt	channelId	channelTitle	categoryId
2	0ZcXDfQYPDI	[Nhạc chế] CÂY CÂY ĐẠI CA   Xuân Dích & Thế Một   Trai Ngoan Parody	2020-05-30T04:00:11Z	UCy5kU2hAo7ZJZOzSmzRAnLQ	Trai Ngoan	10
3	0ZcXDfQYPDI	[Nhạc chế] CÂY CÂY ĐẠI CA   Xuân Dích & Thế Một   Trai Ngoan Parody	2020-05-30T04:00:11Z	UCy5kU2hAo7ZJZOzSmzRAnLQ	Trai Ngoan	10
4	0ZcXDfQYPDI	[Nhạc chế] CÂY CÂY 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						

Hình 3.46: Kết quả sau khi nhóm các video có title bị thay đổi

Từ bảng trên, ta có thể nhận thấy rằng, video clip nằm ở top trending có tên “[Nhạc chế] CÂY CÂY ĐẠI CA” đã có 3 lần đổi tên title của video nhưng vẫn có chung một video\_id.

#### ***Thứ tám: Bao nhiêu chữ không viết hoa trong tên?***

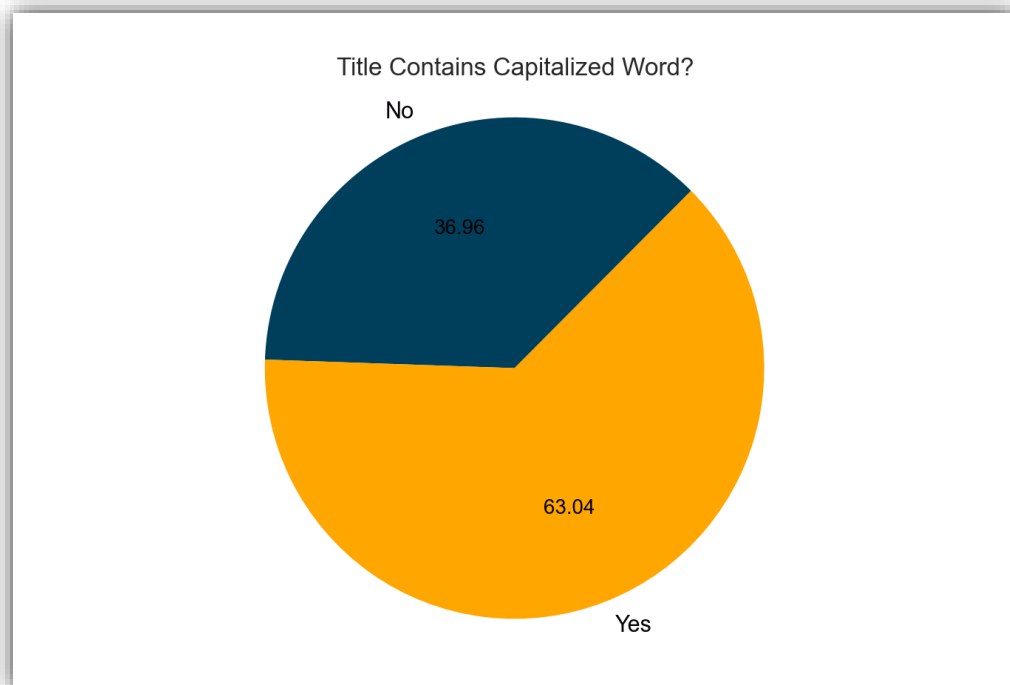
Nhóm đặt ra giả thiết rằng liệu yếu tố viết hoa hoặc không viết hoa ở title một video có tác động đến xu hướng các video nằm trong trending YouTube? Vì thế, nhóm đã tiến hành lọc dữ liệu và nhận được kết quả như sau:

```
def contains_capitalized_word(s):
    for w in s.split():
        if w.isupper():
            return True
    return False

df["contains_capitalized"] = df["title"].apply(contains_capitalized_word)

value_counts = df["contains_capitalized"].value_counts().to_dict()
fig, ax = plt.subplots()
_ = ax.pie([value_counts[False], value_counts[True]], labels=['No', 'Yes'],
           colors=['#003f5c', '#ffa600'], textprops={'color': '#040204'}, startangle=45, autopct="%0.2f")
_ = ax.axis('equal')
_ = ax.set_title('Title Contains Capitalized Word?')
```

*Hình 3.47: Dòng lệnh biểu thị video viết hoa trong tên bằng biểu đồ tròn*



*Hình 3.48: Biểu đồ tròn hiển thị số lượng video viết hoa trong tên*

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

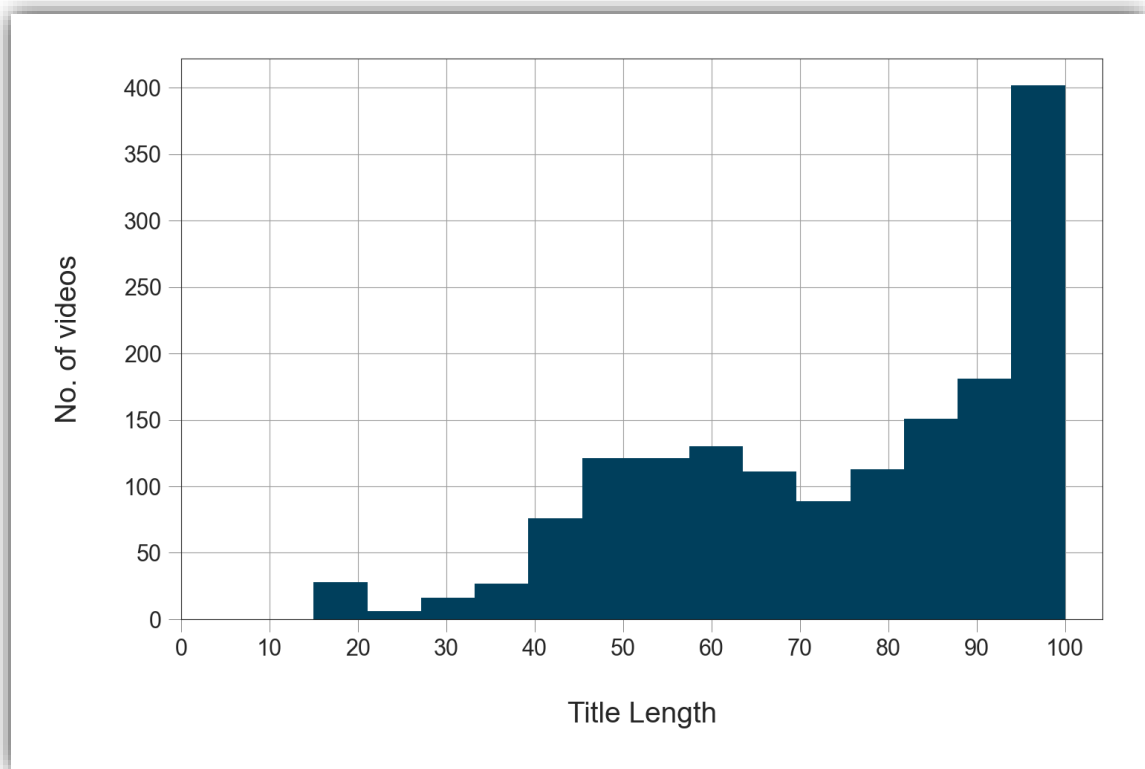
#### ***Thứ chín: Độ dài video title***

Tiếp theo, nhóm sẽ phân tích độ dài của video title của một trending video thường có độ dài bao nhiêu và mối quan hệ giữa độ dài video title và lượt views của video đó có ảnh hưởng đến nhau như thế nào.

```
df["title_length"] = df["title"].apply(lambda x: len(x))

fig, ax = plt.subplots()
_ = sns.distplot(df["title_length"], kde=False, rug=False,
                 color=PLOT_COLORS[4], hist_kws={'alpha': 1}, ax=ax)
_ = ax.set(xlabel="Title Length", ylabel="No. of videos", xticks=range(0, 110, 10))
```

*Hình 3.49: Dòng lệnh biểu thị đồ thị video title*



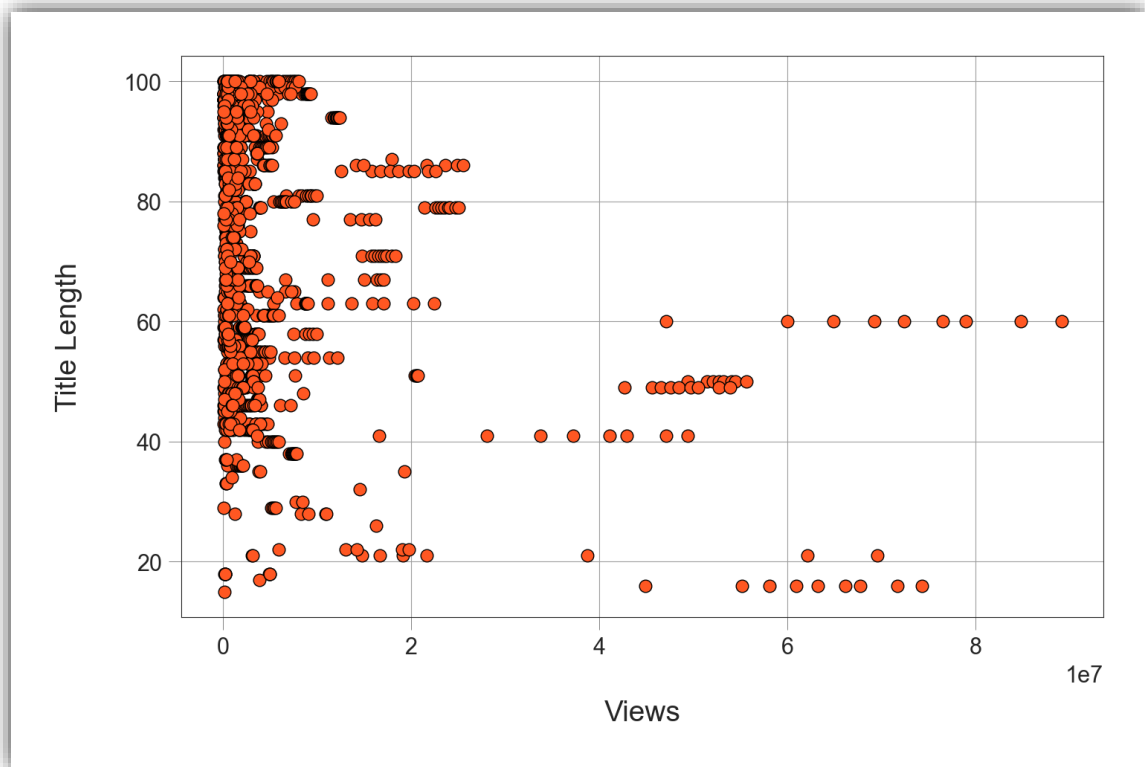
*Hình 3.50: Biểu đồ cột thể hiện độ dài video title*

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 Trending 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.

```
fig, ax = plt.subplots()
_ = ax.scatter(x=df['view_count'], y=df['title_length'], color=PLOT_COLORS[2], edgecolors="#000000", linewidths=0.5)
_ = ax.set(xlabel="Views", ylabel="Title Length")
```

Hình 3.51: Dòng lệnh biểu thị mối quan hệ tiêu đề video và lượt views



Hình 3.52: Biểu đồ scatter plot thể hiện mối quan hệ tiêu đề video và lượt views

Ta có thể thấy là không có mối quan hệ mật thiết nào giữa độ dài của tiêu đề video và lượt views của chúng, tuy nhiên những video có lượt views trên 5 triệu thường có độ dài nằm trong khoảng 15-20 hoặc 60 chữ.

#### ***Thứ mười: Sự tương quan giữa các biến của dataset***

Bây giờ, hãy 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.

```
df.corr()
```

Hình 3.53: Dòng lệnh biểu thị sự tương quan giữa các biến

	categoryId	view_count	likes	dislikes	comment_count	comments_disabled	ratings_disabled	contains_capitalized	title_length
categoryId	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

Hình 3.54: Bảng thể hiện sự tương quan giữa các biến

Ví dụ, ta 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, giữa số lượt thích và không thích (0,88), và giữa số lượt không thích và số bình luận (0,8).

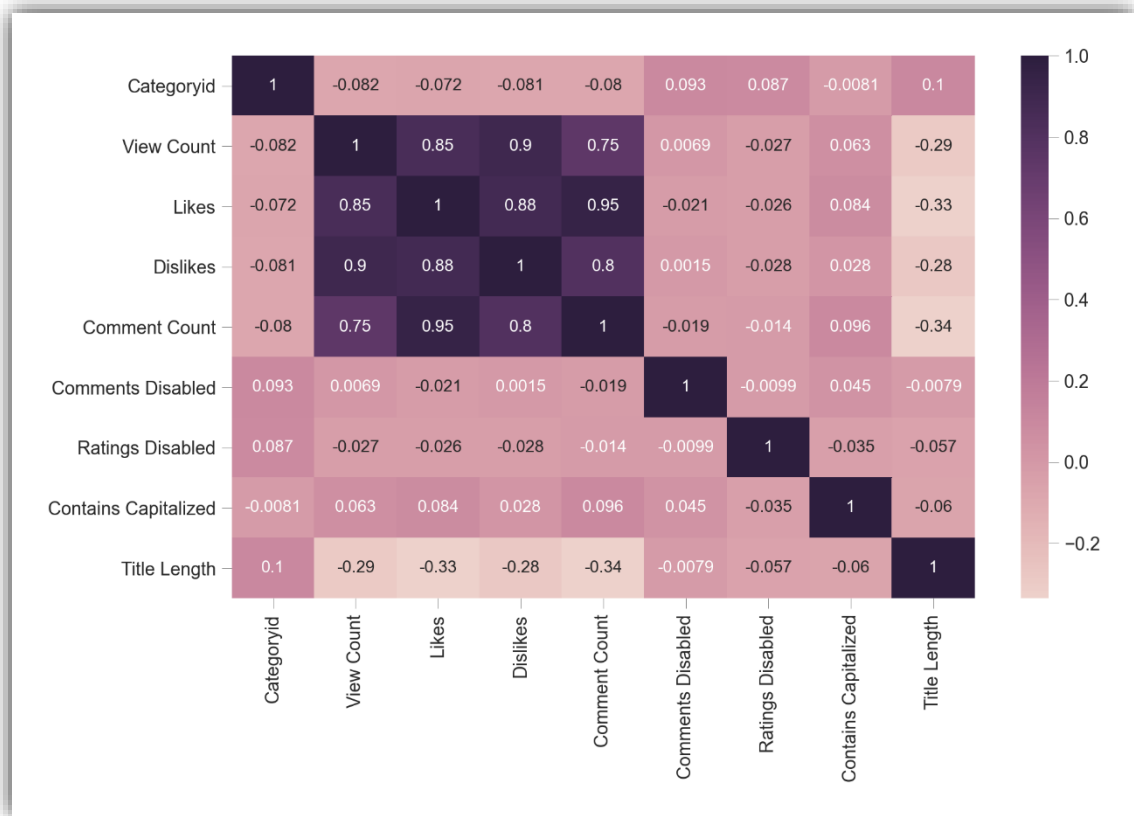
Videos có nhiều likes thì cũng sẽ có mối tương quan thuận với dislikes.

Tiếp theo, chúng ta sẽ trực quan hóa sự tương quan giữa các biến trong tập dữ liệu qua một bản đồ nhiệt để hiểu rõ hơn các mối quan hệ tương quan giữa các biến nào có ảnh hưởng sâu sắc đến nhau.

```
h_labels = [x.replace('_', ' ').title() for x in
             list(df.select_dtypes(include=['number', 'bool']).columns.values)]

fig, ax = plt.subplots(figsize=(10,6))
_ = sns.heatmap(df.corr(), annot=True, xticklabels=h_labels, yticklabels=h_labels, cmap=sns.cubehelix_palette(as_cmap=True), ax=
```

Hình 3.55: Dòng lệnh biểu thị sự tương quan giữa các biến bằng bản đồ nhiệt



Hình 3.56: Bản đồ nhiệt thể hiện sự tương quan giữa các biến

#### Thứ mười một: Wordcloud cho title và tag

Nhóm tiến hành vẽ Wordcloud cho title để thể hiện mức độ phổ biến của các từ được sử dụng thường xuyên trong các video trending của Youtube. Trước tiên, ta cần phải lọc các stopwords và các từ khóa quan trọng để phục vụ cho việc vẽ wordcloud.

```
with open('vietnamese-stopwords.txt') as f:
    stopwords = f.readlines()
stopwords = [x.strip() for x in stopwords]

title_words = list(df["title"].apply(lambda x: x.split()))
title_words = [x for y in title_words for x in y]
Counter(title_words).most_common(25)
```

Hình 3.57: Dòng lệnh lọc stopwords và các từ khóa

```
[('l', 1247),
 ('-', 857),
 ('Phim', 196),
 ('2020', 172),
 ('Tập', 162),
 ('Anh', 134),
 ('Nhất', 120),
 ('Hay', 118),
 ('Nhạc', 112),
 ('Không', 110),
 ('x', 105),
 ('Giang', 94),
 ('Em', 94),
 ('Tinh', 92),
 ('3', 85),
 ('MV', 75),
 ('2', 75),
 ('OFFICIAL', 74),
 ('Thiên', 74),
 ('Quốc', 73),
 ('không', 72),
 ('KHÔNG', 72),
 ('Official', 69),
 ('Ca', 65),
 ('ERIK', 64)]
```

Hình 3.58: Kết quả sau khi lọc stopwords và các từ khóa

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

Bước tiếp theo, ta sẽ tiến hành vẽ wordcloud cho title.

```
wc = wordcloud.WordCloud(stopwords=stopwords,width=1200, height=500,
                        collocations=False, background_color="white",
                        colormap="tab20b").generate(" ".join(title_words))
plt.figure(figsize=(15,10))
plt.imshow(wc, interpolation='bilinear')
_ = plt.axis("off")
```

Hình 3.59: Dòng lệnh vẽ wordcloud cho title



Hình 3.60: Wordcloud cho title

Sau đó ta tiến hành vẽ Wordcloud cho tag để thể hiện mức độ phổ biến của các từ được sử dụng thường xuyên trong các video trending của Youtube. Trước tiên, ta cần phải lọc các tag quan trọng để phục vụ cho việc vẽ wordcloud.

```
tags_words = list(df["tags"].apply(lambda x: x.split()))
tags_words = [x for y in tags_words for x in y]
Counter(tags_words).most_common(25)
```

Hình 3.61: Dòng lệnh lọc các tag

```
[('sai', 981),
 ('ta', 633),
 ('trung', 618),
 ('hay', 563),
 ('không', 525),
 ('mới', 473),
 ('anh', 416),
 ('tập', 382),
 ('khong', 374),
 ('2020|phim', 358),
 ('tình', 357),
 ('giang', 342),
 ('chung', 332),
 ('chế', 319),
 ('quốc', 313),
 ('nhạc', 293),
 ('hay|phim', 278),
 ('chúng', 275),
 ('gia', 260),
 ('là', 243),
 ('thanh', 243),
 ('hành', 238),
 ('nhau', 235),
 ('hanh', 232),
 ('thiên', 231)]
```

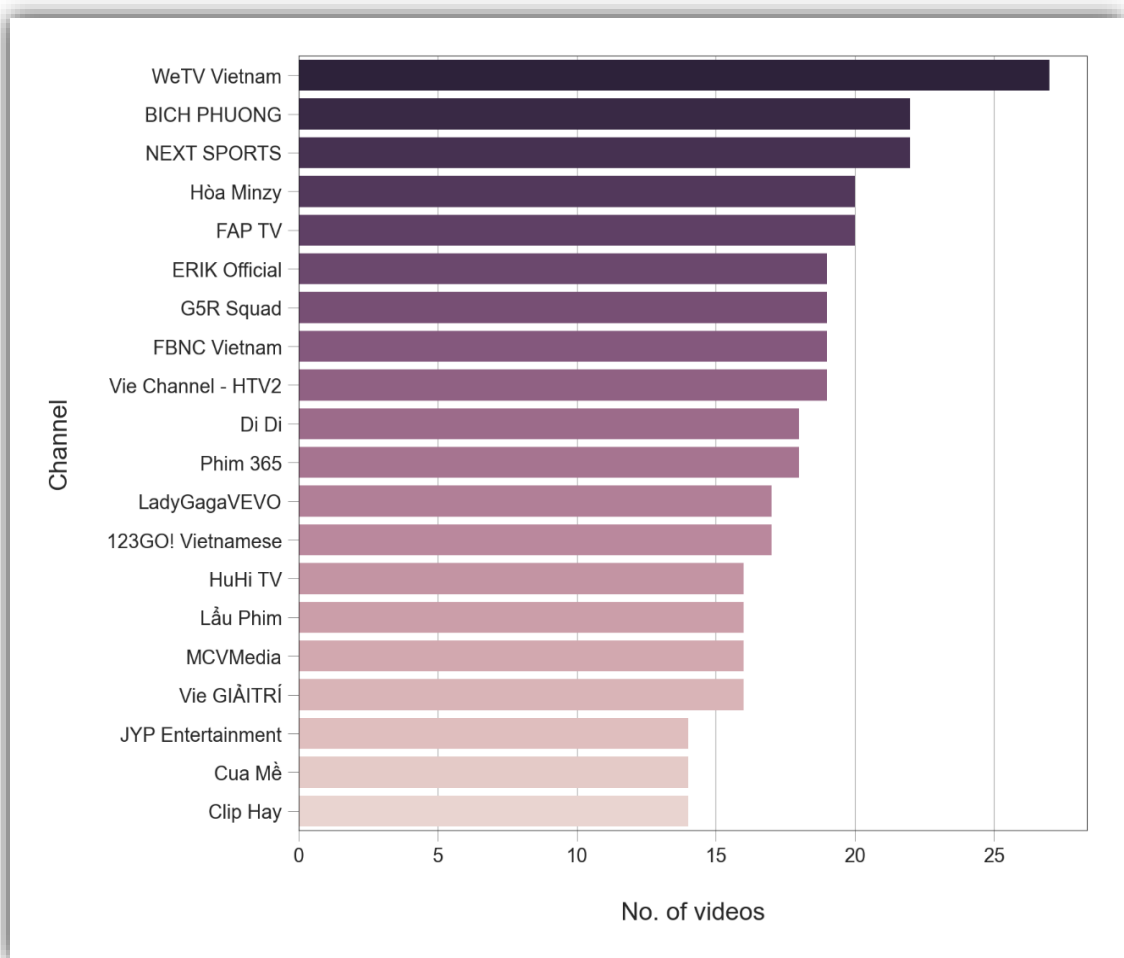
Hình 3.62: Kết quả sau khi lọc các tag

Bước tiếp theo, ta sẽ tiến hành vẽ wordcloud cho tag.

```
wc = wordcloud.WordCloud(stopwords=stopwords,width=1200, height=500,
                          collocations=False, background_color="white",
                          colormap="tab20b").generate(" ".join(tags_words))
plt.figure(figsize=(15,10))
plt.imshow(wc, interpolation='bilinear')
_ = plt.axis("off")
```

Hình 3.63: Dòng lệnh vẽ wordcloud cho tag





Hình 3.66: Biểu đồ cột hiển thị các kênh có trending video tại Việt Nam

Từ đồ thị, ta có thể thấy 3 kênh có nhiều video đạt trending nhất là kênh “WeTV Vietnam”, “BICH PHUONG”, “NEXT SPORTS” đều thuộc danh mục Âm nhạc và Giải trí.

### **Thứ mười ba: Thể loại video có trending video nhiều nhất VN**

Nhóm sẽ tiến hành lọc và sắp xếp các thể loại có nhiều trending video nhất Việt Nam, sau đó trực quan hóa kết quả bằng biểu đồ cột. Đầu tiên, chúng ta sẽ sử dụng một file JSON cung cấp tập dữ liệu chứa các thông tin về từng thể loại video tại Việt Nam.

```
with open("/Users/anhnhath/Documents/UET/SepVI_1920/Phan_tich_du_lieu_Web/Final/VN_category_id.json") as f:
    categories = json.load(f)["items"]
    cat_dict = {}
    for cat in categories:
        cat_dict[int(cat["id"])] = cat["snippet"]["title"]
    df['category_name'] = df['categoryId'].map(cat_dict)
```

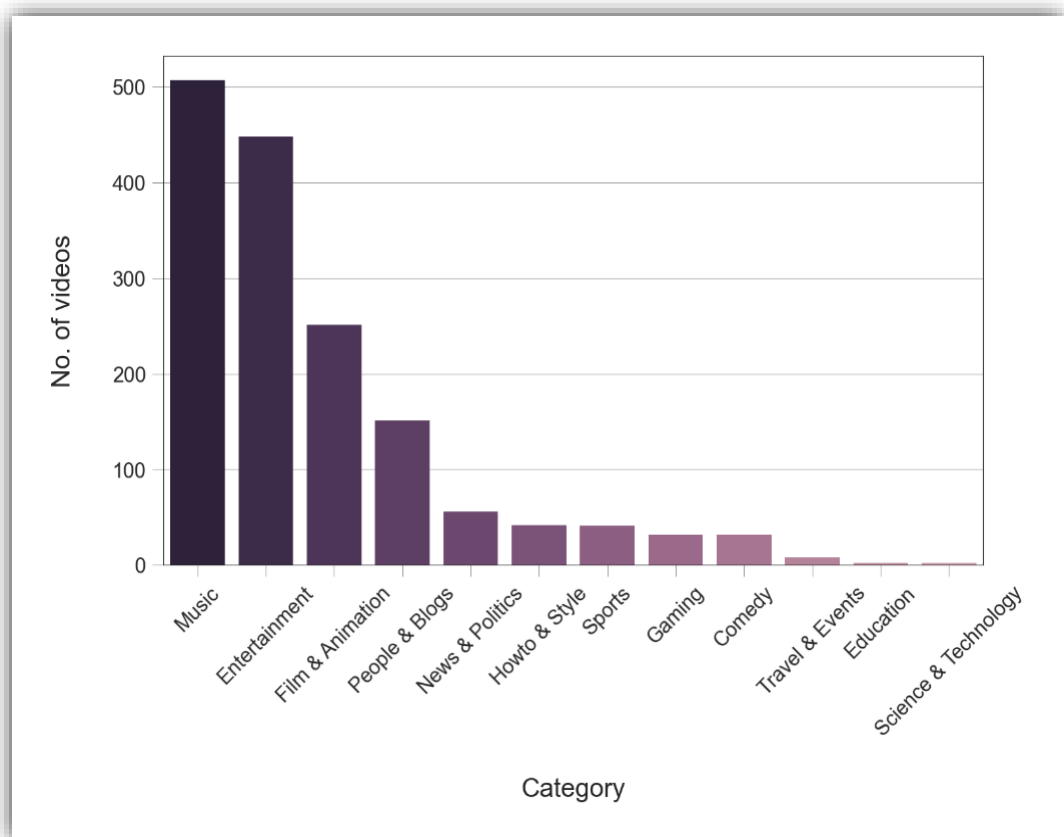
Hình 3.67: Dòng lệnh sử dụng file JSON

```

cdf = df["category_name"].value_counts().to_frame().reset_index()
cdf.rename(columns={"index": "category_name", "category_name": "No_of_videos"}, inplace=True)
fig, ax = plt.subplots()
_ = sns.barplot(x="category_name", y="No_of_videos", data=cdf,
                palette=sns.cubehelix_palette(n_colors=16, reverse=True), ax=ax)
_ = ax.set_xticklabels(ax.get_xticklabels(), rotation=45)
_ = ax.set(xlabel="Category", ylabel="No. of videos")

```

Hình 3.68: Dòng lệnh biểu thị số lượng trending video của từng thể loại



Hình 3.69: Biểu đồ cột thể hiện số lượng trending video của từng thể loại

### **Thứ mười bốn: Thể loại video có trending video nhiều nhất VN**

Nhóm tiến hành phân tích các dữ liệu về thời gian trong tuần để xét xem ngày nào có số lượng trending video được đăng tải nhiều nhất. Đầu tiên, từ tập dữ liệu có sẵn, ta thực hiện các dòng lệnh:

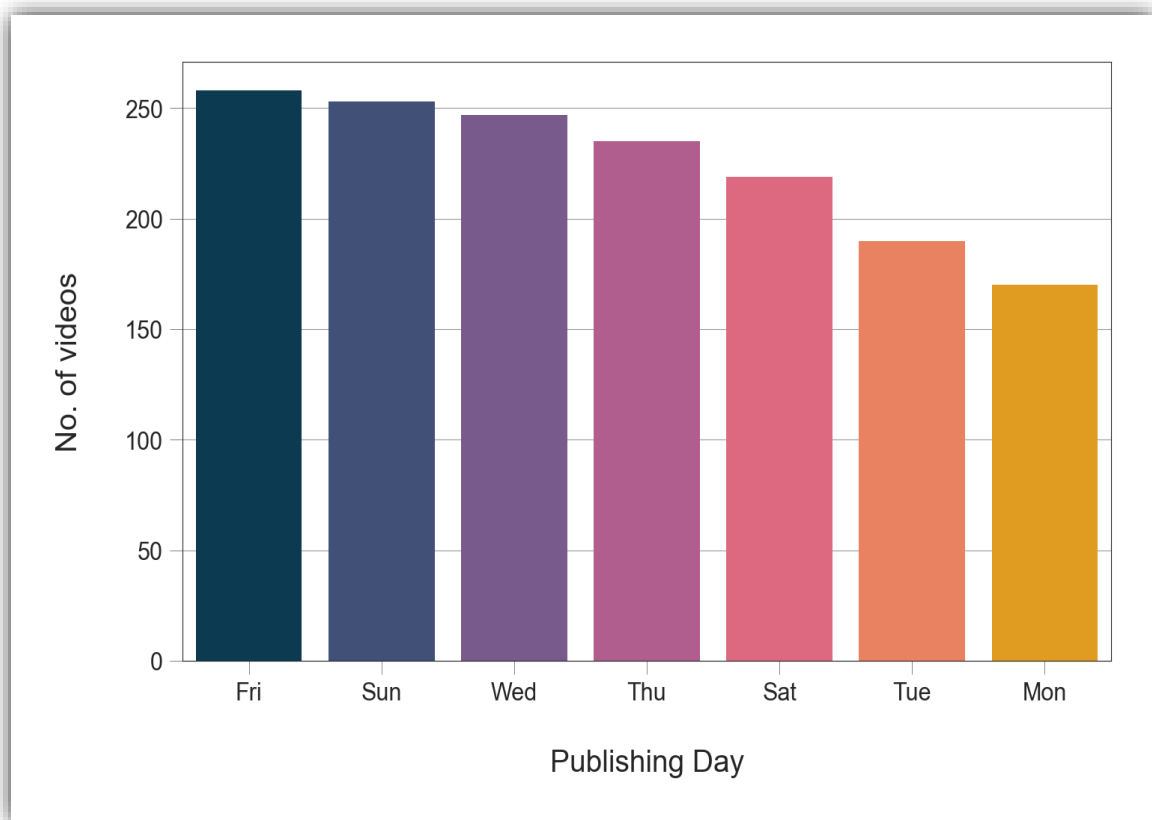


```

cdf = df["publishing_day"].value_counts()\
      .to_frame().reset_index().rename(columns={"index": "publishing_day", "publishing_day": "No_of_videos"})
fig, ax = plt.subplots()
_ = sns.barplot(x="publishing_day", y="No_of_videos", data=cdf,
                palette=sns.color_palette(['#003f5c', '#374c80', '#7a5195',
                                           '#bc5090', '#ef5675', '#ff764a', '#ffa600'], n_colors=7), ax=ax)
_ = ax.set(xlabel="Publishing Day", ylabel="No. of videos")

```

Hình 3.70: Dòng lệnh biểu thị lượng video được đăng tải các ngày trong tuần



Hình 3.71: Biểu đồ cột thể hiện lượng video được đăng tải các ngày trong tuần

Chúng ta có thể thấy rằng số lượng Video Trending được đăng vào Thứ hai và Thứ ba ít hơn đáng kể so với số lượng Video Trending được công bố vào các ngày khác trong tuần.

Bây giờ, ta sử dụng cột “Publishing\_hour” để xem giờ xuất bản nào có số lượng video xu hướng lớn nhất.

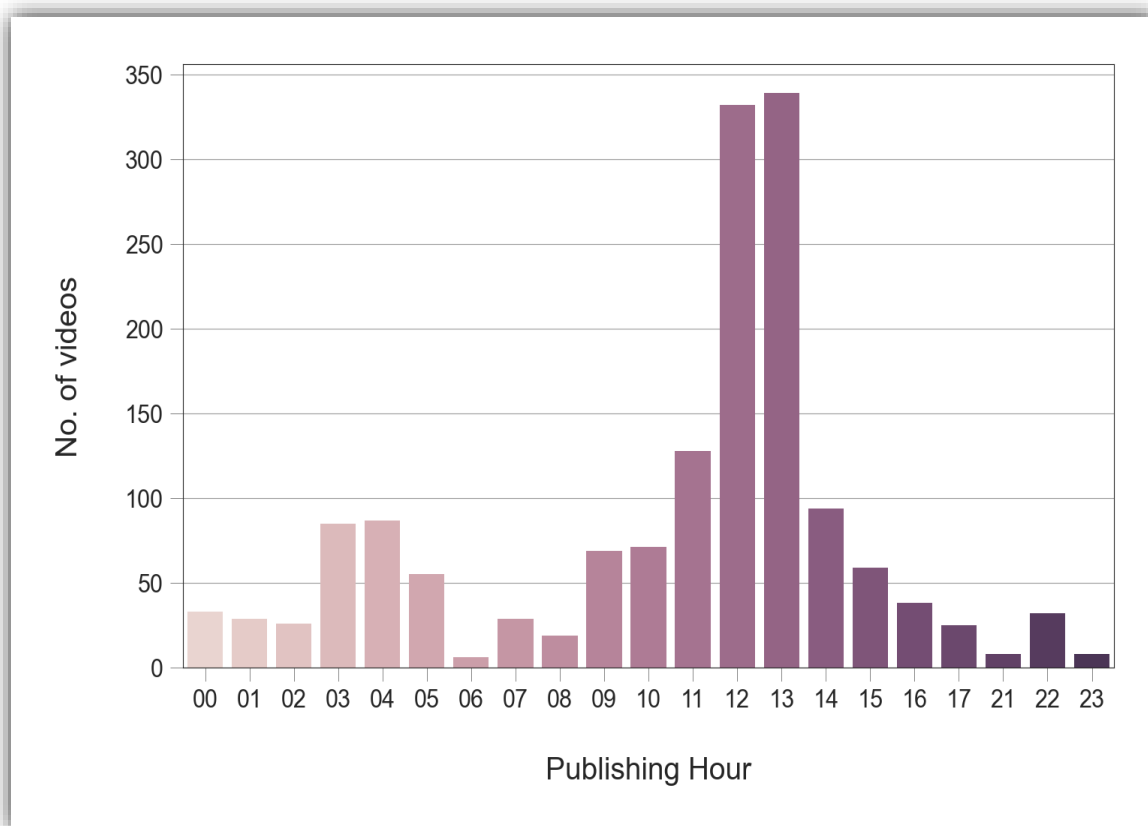


```

cdf = df["publishing_hour"].value_counts().to_frame().reset_index()\
      .rename(columns={"index": "publishing_hour", "publishing_hour": "No_of_videos"})
fig, ax = plt.subplots()
_ = sns.barplot(x="publishing_hour", y="No_of_videos", data=cdf,
                palette=sns.cubehelix_palette(n_colors=24), ax=ax)
_ = ax.set(xlabel="Publishing Hour", ylabel="No. of videos")

```

Hình 3.72: Dòng lệnh biểu thị lượng video được đăng tải các giờ trong ngày



Hình 3.73: Biểu đồ cột thể hiện lượng video được đăng tải các giờ trong ngày

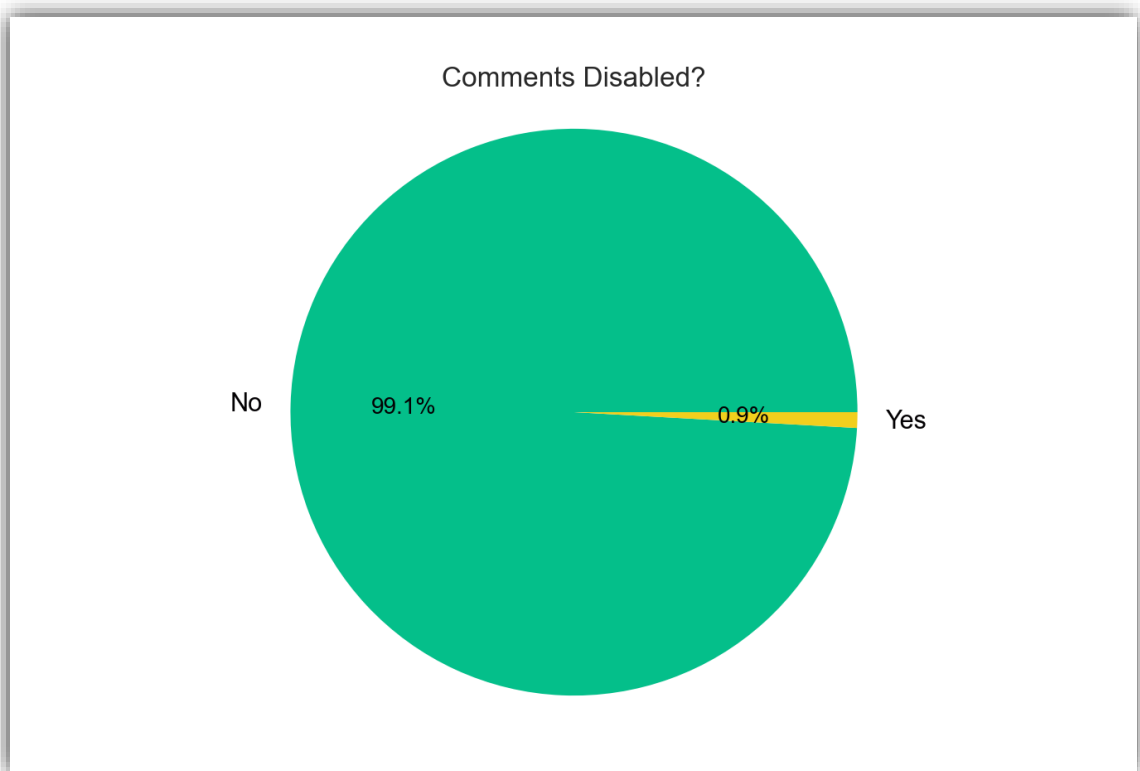
Ta có thể thấy được rằng khoảng thời gian đăng video từ 12PM-13PM là cao điểm nhất, đây là khoảng thời gian nếu bạn đăng sẽ có nhiều khả năng xuất hiện ở trending nhất. Điều này khá lạ khi so sánh với các quốc gia khác, đơn cử là nếu ở Mỹ thì videos thường được đăng vào các khung giờ buổi tối.

**Thứ mười năm: Có bao nhiêu video đóng tính năng bình luận? Bao nhiêu video đóng đánh giá (like/dislike)?**

Cuối cùng, nhóm sẽ dùng tập dữ liệu sẵn có để phân tích xem bao nhiêu trending video đã đóng tính năng bình luận và đóng tính năng đánh giá. Chúng ta có thể thực hiện các thao tác sau:

```
value_counts = df["comments_disabled"].value_counts().to_dict()
fig, ax = plt.subplots()
_ = ax.pie(x=[value_counts[False], value_counts[True]], autopct='%0.1f%%', labels=['No', 'Yes'],
          colors=['#04BF8A', '#F2CF1D'], textprops={'color': '#040204'})
_ = ax.axis('equal')
_ = ax.set_title('Comments Disabled?')
```

Hình 3.74: Dòng lệnh biểu thị lượng video đóng tính năng bình luận



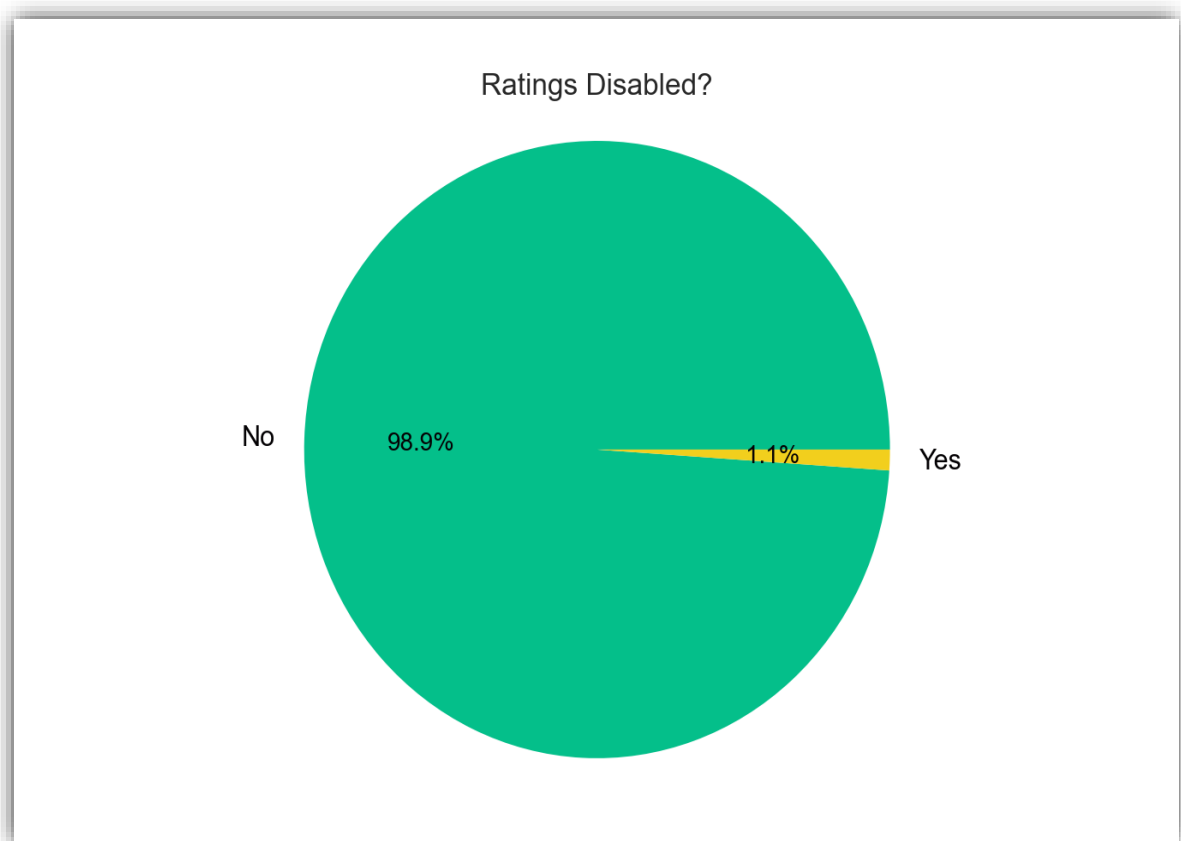
Hình 3.75: Biểu đồ tròn thể hiện lượng video đóng tính năng bình luận

Có thể thấy con số vô cùng ít (chỉ 0,9%), bởi lượt tương tác của người xem tác động rất lớn đến video.

Tiếp theo, ta sẽ bắt đầu phân tích số lượng video đã đóng tính năng đánh giá qua các dòng lệnh như sau:

```
value_counts = df["ratings_disabled"].value_counts().to_dict()
fig, ax = plt.subplots()
_ = ax.pie([value_counts[False], value_counts[True]], autopct='%0.1f%%', labels=['No', 'Yes'],
           colors=['#04BF8A', '#F2CF1D'], textprops={'color': '#040204'})
_ = ax.axis('equal')
_ = ax.set_title('Ratings Disabled?')
```

Hình 3.76 Dòng lệnh biểu thị lượng video đóng tính năng đánh giá



Hình 3.77: Biểu đồ tròn thể hiện lượng video đóng tính năng bình luận

Tương tự như việc đóng tính năng bình luận, việc đóng tính năng đánh giá cũng rất ít (chỉ 1,1%), vì các tính năng trên là nhân tố ảnh hưởng đến việc để video đạt được trending. Vì vậy, người đăng tải video sẽ hạn chế tối đa việc đóng những tính năng trên.

### 3.2. Kết quả thu được

#### 3.2.1. Tóm tắt kết quả phân tích

Dưới đây là một số kết luận mà nhóm đã thu thập được:

- 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 bao gồm 1572 videos đầu vào.
- 92,2% videos trending có dưới 10 triệu views, and 81,3% videos có dưới 5 triệu views.
- 78,2% của videos trending có dưới 60,000 likes, và 82,7% videos có dưới 100,000 likes.
- 92,6% videos trending có ít hơn 20,000 comments.
- Rất nhiều videos liên tục đứng top bảng xếp hạng trending trong nhiều ngày. Dữ liệu của nhóm có 1572 videos đầu vào nhưng chỉ có tổng cộng 430 videos là duy nhất.
- Những videos có hơn 5,000,000 views và lớn hơn thường có độ dài tiêu đề video nằm trong khoảng 15 đến 20 từ, hoặc là 60 từ.
- Các stopword - và | rất phổ biến trong tiêu đề của các videos.
- Các cụm từ như 'Official', 'Video', 'Phim', 'Nhạc', 'Tập' rất phổ biến trong tiêu đề các video.
- Có một mối tương quan tích cực rất mạnh giữa số lượt xem và số lượt thích của các Videos trending: Khi một trong số chúng tăng, số khác cũng đồng thời tăng và ngược lại.
- Có một mối tương quan tích cực mạnh mẽ giữa số lượt thích và số lượng bình luận, giữa số lượng không thích và số lượng bình luận cũng có mối tương quan tương tự.
- Danh mục có số lượng video xu hướng lớn nhất là 'Âm nhạc' với hơn 500 videos, tiếp theo là danh mục Giải trí, Phim và hoạt hình.
- Điều đáng buồn là các videos nằm trong danh mục Khoa học và công nghệ, Giáo dục là các chủ đề có ít videos trending nhất.

## CHƯƠNG 4: KẾT LUẬN VÀ ĐÁNH GIÁ

### 4.1. Tóm tắt nội dung và kết quả của đề tài

Đề tài “Phân tích dữ liệu trending YouTube” đã thu thập dữ liệu bằng ngôn ngữ Python, YouTube API và tiến hành làm sạch dữ liệu. Sau đó, nhóm tiến hành trực quan hóa dữ liệu bằng các biểu đồ, từ đó đưa ra các phân tích về dữ liệu.

Đề tài đã đạt được những kết quả như mong đợi, hoàn thành các mục tiêu đề ra và các thành viên trong nhóm học hỏi được nhiều kiến thức mới.

### 4.2. Ưu điểm

Về đề tài:

- Nhóm đã áp dụng linh hoạt những kiến thức được học và tìm tòi, khám phá thêm những kiến thức mới.
- Thu thập được những con số dữ liệu ấn tượng, giá trị phù hợp với nội dung đề tài.
- Dữ liệu của đề tài mang tính ứng dụng cao, vừa có thể xây dựng thành bộ tiêu chí video top trending tham khảo cho các YouTuber, vừa là nền tảng dữ liệu cho các doanh nghiệp nghiên cứu và triển khai chiến lược quảng bá của mình.

Về cá nhân các thành viên:

- Có sự đóng góp tích cực xây dựng đề tài
- Chịu khó tìm tòi, học hỏi
- Thực hiện đúng nội quy nhóm, có sự phối hợp ăn ý giữa các thành viên để xây dựng thành công đề tài

### 4.3. Nhược điểm

Bên cạnh những kết quả đã đạt được, đề tài vẫn tồn tại nhiều điểm hạn chế cần được khắc phục và hoàn thiện:

- Nguồn dữ liệu vẫn chưa nhiều vì thời gian hạn chế dẫn đến kết quả phân tích chưa chuyên sâu.
- Các kiến thức của nhóm vận dụng vào đề tài còn nhiều hạn chế, đôi khi vẫn còn phát sinh lỗi.
- Kết quả đồ án chỉ tạm dừng lại ở mức phân tích cơ bản chứ chưa đi sâu vào việc đưa ra những tri thức.

### 4.4. Hướng phát triển của đề tài

- Tiến hành thu thập thêm dữ liệu trong khoảng thời gian là một năm. Sau đó, tiến hành phân tích chuyên sâu thêm về các yếu tố liên quan đến những video trong top trending YouTube.  
Từ đó, 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.

## BÁO CÁO QUÁ TRÌNH LÀM VIỆC NHÓM

### I. Danh sách thành viên

STT	HỌ VÀ TÊN	MSSV
1	Nguyễn Anh Nhật (Nhóm trưởng)	K174111311
2	Nguyễn Quốc Triệu	K174111323
3	Thiêm Ánh Tường Vy	K174111329

### II. Quy định làm việc của nhóm

Cách thức làm việc:

- Trực tiếp: Họp thường xuyên tại trường hoặc quán cà phê mỗi tuần 1 – 2 buổi.
- Gián tiếp: Họp nhóm và liên hệ nhau thông qua các hệ thống trực tuyến (Facebook, Email, Drive và Zoom).

Quy định làm việc:

- Mỗi thành viên khi tham gia họp phải chuẩn bị sổ ghi chép, bút viết và máy tính cá nhân. Tham gia đầy đủ các buổi họp thường kỳ của nhóm, khi vắng phải thông báo và xin phép trước ít nhất một ngày, nếu không sẽ phải đóng phạt.
- Mỗi thành viên phải gửi nội dung được giao đến những thành viên còn lại vào Drive chung của nhóm để nắm được nội dung. Đồng thời, nhóm trưởng phải tổng hợp, thông báo nội dung buổi họp tiếp theo đến các thành viên của nhóm.
- Các thành viên phải đến đúng giờ, chuẩn bị trước nội dung mà nhóm sẽ trình bày, bàn luận. Nếu thành viên không sắp xếp được thời gian tham gia họp phải thông báo cho nhóm trưởng để có sự điều chỉnh thời gian, địa điểm phù hợp.
- Các thành viên đều phải tiến hành bàn luận, góp ý trên tinh thần nghiêm túc, nhiệt tình cho từng nội dung của thành viên còn lại để hoàn thành công đúng thời gian, đạt kết quả tốt.
- Thời lượng họp trung bình mỗi buổi từ 1 – 2 giờ để đảm bảo chất lượng cuộc họp

- Các thành viên phải tôn trọng deadline đã đặt ra và hoàn thành công việc của mình đúng thời hạn được giao (trễ deadline phạt 1 phút/ngày mỗi cá nhân)
- Sau khi nộp bài, các thành viên tiến hành kiểm tra chéo nội dung của nhau để đảm bảo sự chính xác, tính logic của đề tài.
- Trước khi báo cáo sẽ có một buổi họp mặt tổng hợp, duyệt thử.

### III. Bảng phân công nhiệm vụ

HỌ VÀ TÊN	NHIỆM VỤ	MỨC ĐỘ HOÀN THÀNH
Nguyễn Anh Nhật	<ul style="list-style-type: none"> <li>- Lên ý tưởng cho đề án</li> <li>- Đóng góp xây dựng đề án</li> <li>- Sửa code cho đề án</li> <li>- Phân công và đốc thúc tiến độ các thành viên trong nhóm, lên kế hoạch họp và mục tiêu đề tài.</li> </ul>	100%
Nguyễn Quốc Triệu	<ul style="list-style-type: none"> <li>- Đóng góp bổ sung ý tưởng</li> <li>- Đóng góp xây dựng đề án</li> <li>- Hỗ trợ đóng góp nội dung</li> <li>- Làm ppt</li> </ul>	100%
Thiền Ánh Tường Vy	<ul style="list-style-type: none"> <li>- Đóng góp bổ sung ý tưởng</li> <li>- Đóng góp xây dựng đề án</li> <li>- Tổng hợp, định dạng file word.</li> <li>- Báo cáo quá trình làm việc nhóm.</li> </ul>	100%



## TÀI LIỆU THAM KHẢO

- [1] DataReportal, “Digital 2019 Vietnam (January 2019) v01,” 06:36:24 UTC, Accessed: Jul. 01, 2020. [Online]. Available: <https://www.slideshare.net/DataReportal/digital-2019-vietnam-january-2019-v01>.
- [2] “YouTube,” *Wikipedia tiếng Việt*. Jun. 28, 2020, Accessed: Jul. 01, 2020. [Online]. Available: <https://vi.wikipedia.org/w/index.php?title=YouTube&oldid=62751646>.
- [3] urekamedia.com, “VIETNAM DIGITAL MARKETING TRENDS IN 2018.” <https://urekamedia.com/en/news/vietnam-digital-marketing-trends-in-2018-96+92> (accessed Jul. 01, 2020).
- [4] “What is Python? Executive Summary,” *Python.org*. <https://www.python.org/doc/essays/blurb/> (accessed Jul. 01, 2020).
- [5] “Python Features,” *GeeksforGeeks*, Apr. 03, 2019. <https://www.geeksforgeeks.org/python-features/> (accessed Jul. 01, 2020).
- [6] “An introduction to seaborn — seaborn 0.10.1 documentation.” <https://seaborn.pydata.org/introduction.html> (accessed Jul. 01, 2020).
- [7] V. Beal, “What is API - Application Program Interface? Webopedia Definition.” <https://www.webopedia.com/TERM/A/API.html> (accessed Jul. 01, 2020).
- [8] “API là gì? 3 đặc điểm cơ bản của API.” <https://jobs.hybrid-technologies.vn/blog/api-la-gi/> (accessed Jul. 01, 2020).

## **PHỤ LỤC 1: SOURCE CODE ĐỒ ÁN**

**Github:** [https://github.com/anhnhatucl/UEL\\_ClassProject\\_Mining-Social-Web](https://github.com/anhnhatucl/UEL_ClassProject_Mining-Social-Web)