



■ Final Project:

TikTok Data Analytics & TikTok Script Writing Assistant

Ứng dụng phân tích dữ liệu
thông minh - 21KHDL

Khoa học dữ liệu ứng dụng
- 21KHDL



01

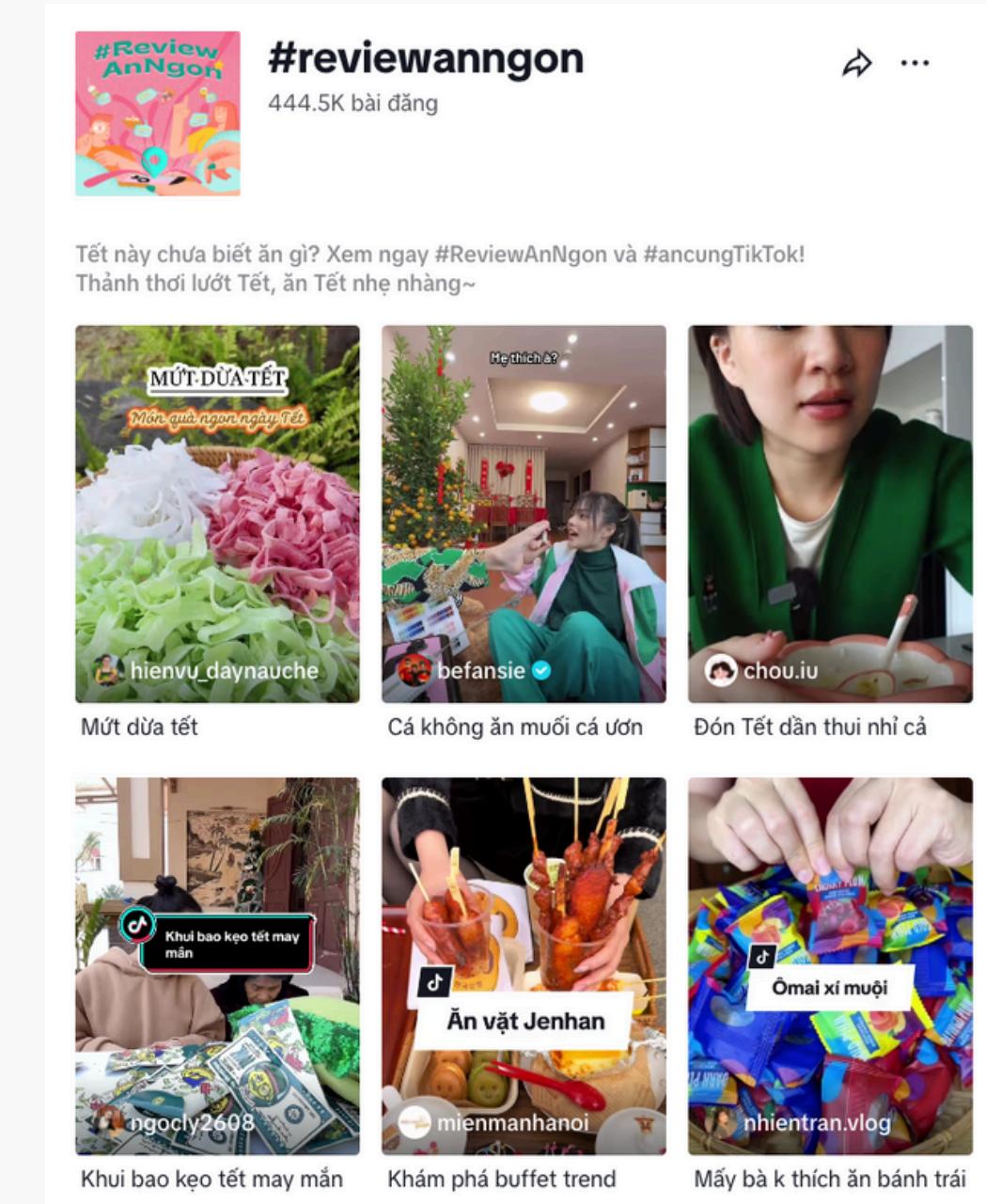
DATA COLLECTION

Dữ liệu đã thu thập

Dữ liệu: ~70,000 thông tin video trong 1,5 tháng cuối năm 2023, cả năm 2024 và 3 tháng đầu năm 2025 và thông tin tài khoản của 264 người dùng TikTok thuộc chủ đề Ăm thực.

Phương pháp thu thập:

- Vid 1
1. Thu thập thủ công một số hashtag nổi bật về ẩm thực và tiến hành crawl video theo các hashtag này.
 2. Crawl video của user có ít nhất 2 video có sử dụng các hashtag trên.
 3. Lấy ra 50 hashtag ẩm thực được nhiều người dùng nhất (không trùng lặp với hashtag cũ) và crawl theo hashtag này.
 4. Merge 2 danh sách video và lọc ra danh sách user có ít nhất 2 video và có các chỉ số thống kê về:
 - **Lượt xem tổng > 100,000 (Q1)**
 - **Lượt xem trung bình > 40,000 (Q1)**
 - **Tỉ lệ tương tác=(like+share+comment)/view > 0.03 (Q2)**
- Vid 2



800 users

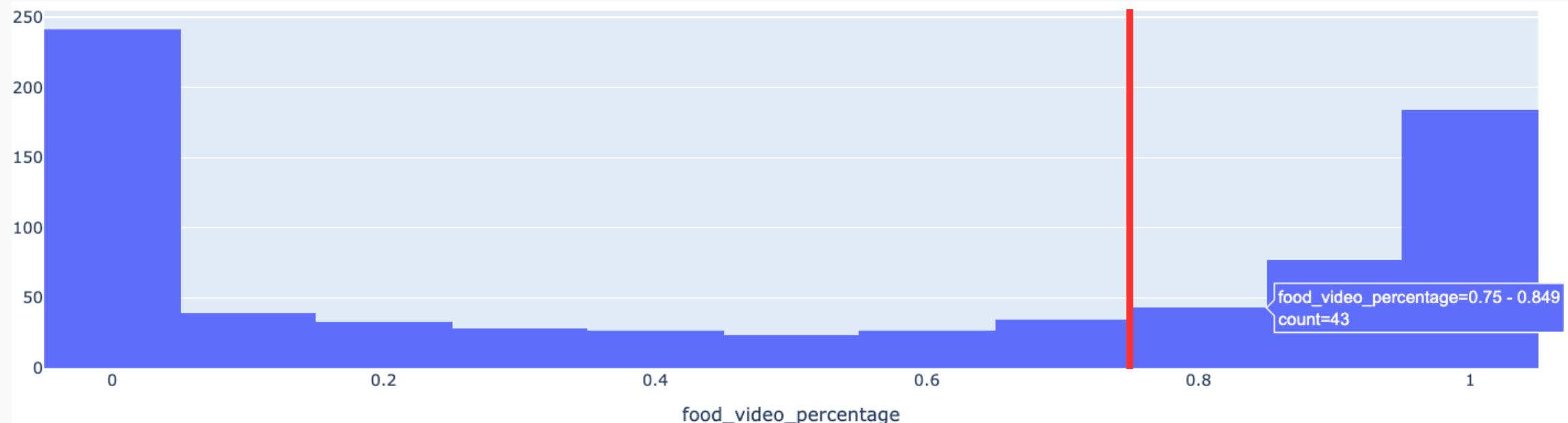
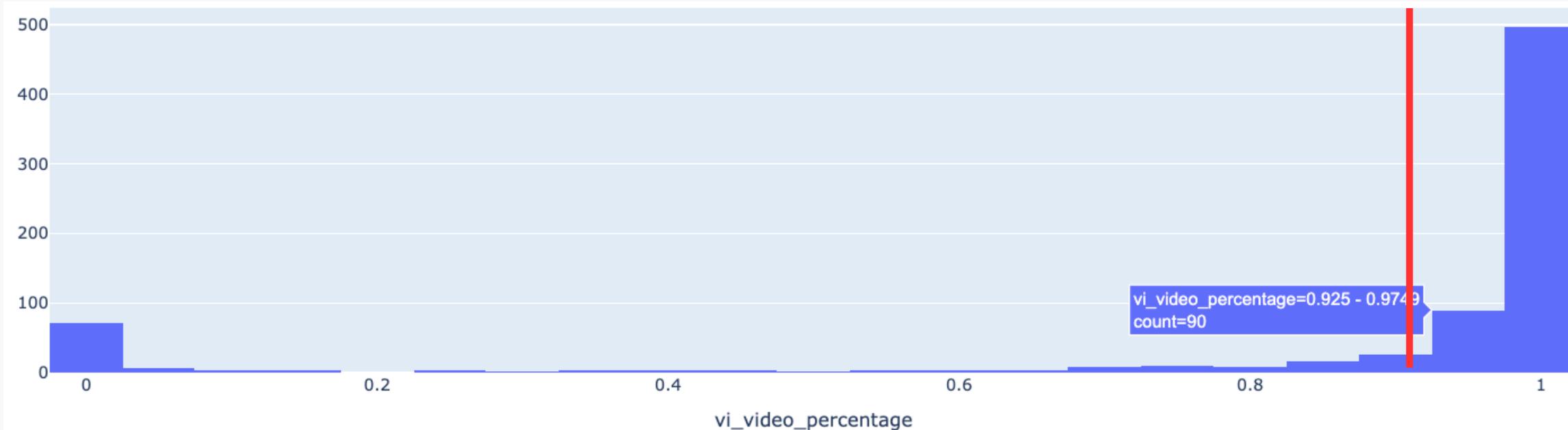
Dữ liệu đã thu thập

Dữ liệu: ~70,000 thông tin video trong 1,5 tháng cuối năm 2023, cả năm 2024 và 3 tháng đầu năm 2025 và thông tin tài khoản của 264 người dùng TikTok thuộc chủ đề Ăm thực.

Tiêu chí lọc user:

- Không có video trong năm 2025
- Có ít hơn 90% tổng lượng video bằng tiếng Việt.
- Có ít hơn 75% tổng lượng video liên quan đến ẩm thực.

264 users



Sơ lược về TikTokApi

Unofficial TikTok API in Python

This is an unofficial api wrapper for TikTok.com in python. With this api you are able to call most trending and fetch specific user information as well as much more.



This api is designed to **retrieve data** TikTok. It **can not be used post or upload content** to TikTok on the behalf of a user. It has **no support for any user-authenticated routes**, if you can't access it while being logged out on their website you can't access it here.



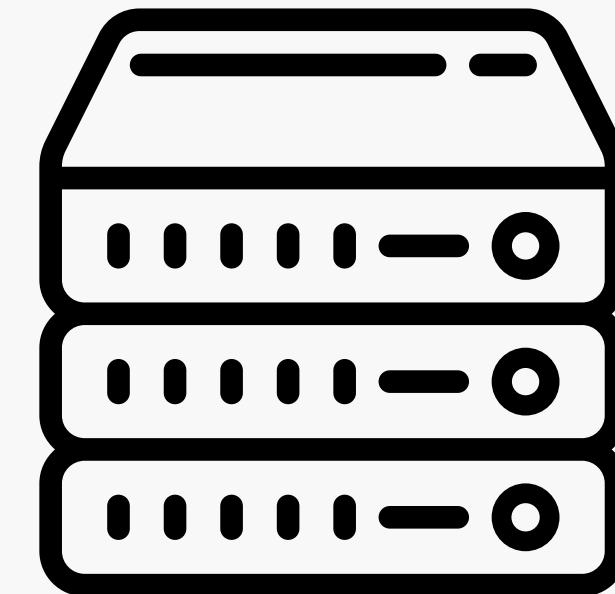
Gọi API nội bộ <https://www.tiktok.com/api/user/detail/>



Nhận về response dạng JSON

```
{  
    "stats": {  
        "diggCount": 0,  
        "followerCount": 52200,  
        "followingCount": 17,  
        "friendCount": 11,  
        "heart": 2300000,  
        "heartCount": 2300000,
```

Playwright



TikTok Server

02

DATA EXPLORING & PREPROCESSING

Data Exploring & Preprocessing

01 Xác định trước kiểu dữ liệu cho các thuộc tính phân loại

- Một số cột có **kiểu dữ liệu phân loại** nhưng có **giá trị dưới dạng chữ số**
 - Thuộc tính phân loại “author.commentSetting” (*có cho phép người xem bình luận vào video hay không*) có 2 giá trị khác nhau là: “0” và “1”

```
author.commentSetting
0.0      54226
1.0      16770
Name: count, dtype: int64
```

01 Xác định trước kiểu dữ liệu cho các thuộc tính phân loại

- Mặc định, **pandas** sẽ xem các cột này có kiểu dữ liệu dạng số
 - Điều này **không phù hợp với ý nghĩa thực sự** và có thể **gây sai sót** trong quá trình phân tích
- Do đó, ta cần **chỉ định kiểu dữ liệu của các cột** này trước khi đọc dữ liệu

```
dtypes = {  
    "id": np.object_,  
    "video.id": np.object_,  
    "video.videoID": np.object_,  
    "user.commerceUserInfo.category": np.object_,  
    "user.commerceUserInfo.categoryButton": np.object_,  
    "BAInfo": np.object_,  
    # ...  
}
```

02 Tìm hiểu về tập dữ liệu thô

- Tập dữ liệu thô có **71260 hàng** và **174 cột**
 - Mục tiêu là **giảm số chiều** của dữ liệu, **giữ lại các thông tin quan trọng nhất**
- Mỗi hàng chứa thông tin thống kê về 1 video như:
 - Số lượt xem, số lượt like, v.v.
 - Thông tin định danh về **bài hát** được sử dụng
 - Thông tin định danh về **chủ tài khoản TikTok**
 - v.v.

03 Ý nghĩa của một số cột quan trọng

Thông tin chi tiết về tài khoản:

- **author.uniqueId**: ID dùng để định danh công khai tài khoản.
- **author.nickname**: Tên hiển thị (*nickname*) của tài khoản.
- **author.duetSetting**: Cài đặt cho phép thực hiện tính năng duet (hát cùng, ghép video) với video của tác giả.
- **author.downloadSetting**: Cài đặt cho phép tải xuống video hay không.
- **author.privateAccount**: Cờ cho biết tài khoản được đặt ở chế độ riêng tư hay không.
- v.v.

03 Ý nghĩa của một số cột quan trọng

Thống kê của tác giả:

- **authorStats.followerCount:** Số lượng người theo dõi tài khoản.
- **authorStats.followingCount:** Số lượng tài khoản mà tác giả đang theo dõi.
- **authorStats.friendCount:** Số bạn bè (thường là mối quan hệ hai chiều).
- **authorStats.heartCount:** Tổng số lượt yêu thích mà tài khoản nhận được.
- **authorStats.videoCount:** Số lượng video mà tác giả đã đăng tải.

03 Ý nghĩa của một số cột quan trọng

Thông tin chung về video:

- **createTime**: Thời gian đăng tải video (thường là dạng timestamp).
- **desc**: Nội dung mô tả hoặc caption của video.
- **video.id**: ID duy nhất của video.
- v.v.

Data Exploring & Preprocessing

03 Ý nghĩa của một số cột quan trọng

Thống kê và số liệu video (stats.*):

- **collectCount:** Số lượt lưu video.
- **commentCount:** Số lượt bình luận.
- **heartCount:** Số lượt thích.
- **playCount:** Số lượt xem.
- **shareCount:** Số lượt chia sẻ.
- v.v.

03 Ý nghĩa của một số cột quan trọng

Thông tin kỹ thuật video:

- **video.VQScore**: Điểm chất lượng của video do TikTok tính toán.
- **video.bitrate**: Bitrate của video.
- **video.definition**: Độ phân giải của video.
- **video.duration**: Thời lượng video tính bằng giây.
- **video.height / video.width**: Chiều cao và chiều rộng của video.
- v.v.

03 Ý nghĩa của một số cột quan trọng

Thông tin về bài hát

- **music.album:** Tên album chứa bài nhạc được sử dụng trong video.
- **music.authorName:** Tên của nghệ sĩ hoặc nhà sản xuất âm nhạc.
- **music.duration:** Thời lượng của đoạn nhạc.
- **music.id:** ID duy nhất của bài nhạc.
- **music.isCopyrighted:** Cờ cho biết bài nhạc có được bảo hộ bản quyền hay không.
- **music.title:** Tựa đề của bài nhạc.
- v.v.

Data Exploring & Preprocessing

04 Phân tích tỷ lệ trùng lặp (duplicate)

- Để xác định các hàng bị trùng lặp, ta cần dựa vào giá trị “**video.id**”
- Ta sẽ không xét các hàng bị thiếu giá trị “**video.id**”

Rows missing "video.id" data: 264
Percentage: 0.37%



Ta cần loại bỏ 264 hàng này

- Kết quả:

Trước khi loại bỏ, tập dữ liệu có 71260 hàng.
Đã loại bỏ 264 hàng thiếu giá trị ở cột `video.id`.
Tập dữ liệu sau khi loại bỏ có 70996 hàng.

04 Phân tích tỷ lệ trùng lặp (duplicate)

- Các hàng có giá trị của cột “**video.id**” giống nhau là các hàng bị trùng lặp
- Sử dụng phương thức **duplicated()** của **DataFrame**
 - Kết quả:

Dữ liệu không có hàng nào bị trùng lặp!
Suy ra, tỉ lệ hàng bị trùng lặp là **0.00%**.

➡ Ta không cần xử lý gì thêm

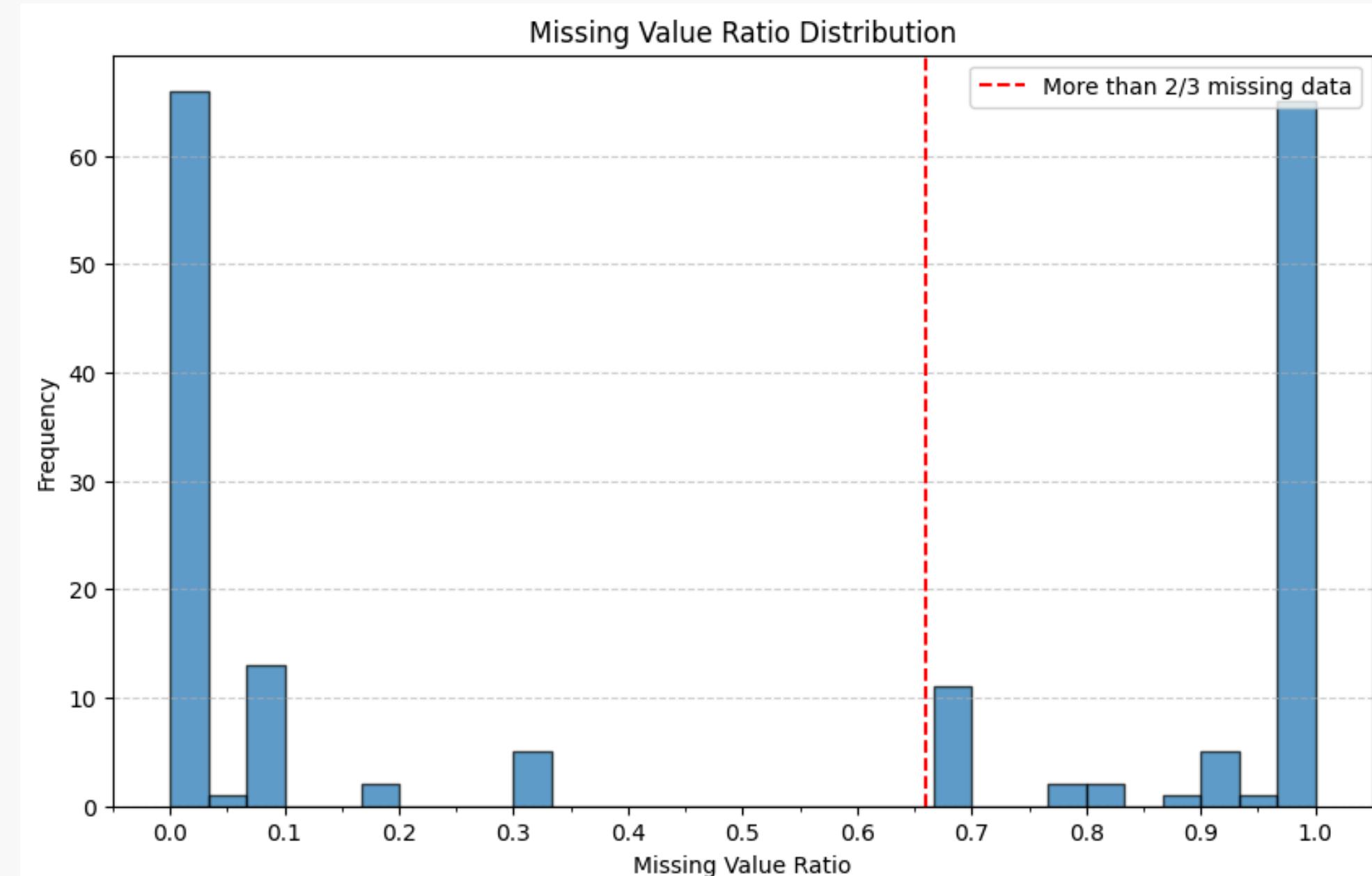
Data Exploring & Preprocessing

05 Phân tích tỷ lệ thiếu giá trị ở mỗi cột (missing rate)

- Có khoảng 50% số cột thiếu hơn $\frac{2}{3}$ giá trị
 - Đây là các cột **không đủ chất lượng** để sử dụng trong quá trình phân tích

→ Loại bỏ khỏi tập dữ liệu

Trước khi loại bỏ, tập dữ liệu có **174** cột.
Loại bỏ **87** cột thiếu hơn **66.67%** dữ liệu.
Tập dữ liệu sau khi loại bỏ có **87** cột.



87 cột có tỉ lệ thiếu giá trị lớn hơn 66.67%

06 Phân tích kiểu dữ liệu của mỗi cột

- Khi đọc dữ liệu, ta đã xác định trước kiểu dữ liệu cho một số cột đặc biệt
 - **Hầu hết các cột đều có kiểu dữ liệu phù hợp**
- Chỉ có 2 cột có kiểu dữ liệu chưa phù hợp:
 - **createTime** và **collectTime** đang có kiểu **object** thay vì **datetime**

	collectTime	createTime
0	1742378359	1738758055
1	1742378359	1723728708

→ Cần tiền xử lý 2 cột này

06 Phân tích kiểu dữ liệu của mỗi cột

- **Bước 1:** Chuyển từ dữ liệu từ dạng **timestamp** sang **datetime**

```
video_df["collectTime"] = pd.to_datetime(video_df["collectTime"], unit="s")
video_df["createTime"] = pd.to_datetime(video_df["createTime"], unit="s")
```

- **Bước 2:** Vì hầu hết nhà sáng tạo nội dung đều hoạt động tại Việt Nam, nên ta **chuyển dữ liệu sang múi giờ Việt Nam** để việc phân tích chính xác hơn

```
video_df['createTime'] = video_df['createTime'].dt.tz_localize(
    'UTC').dt.tz_convert("Asia/Ho_Chi_Minh")
video_df['collectTime'] = video_df['collectTime'].dt.tz_localize(
    'UTC').dt.tz_convert("Asia/Ho_Chi_Minh")
```

Data Exploring & Preprocessing

07

Phân tích phân bố của các giá trị trong mỗi cột có kiểu dữ liệu dạng số (numerical)

- Với mỗi cột có kiểu dữ liệu dạng số, ta sẽ tính:
 - Tỷ lệ thiểu giá trị (từ 0 đến 100).
 - Giá trị tối thiểu.
 - Giá trị tứ phân vị thứ nhất.
 - Giá trị tứ phân vị thứ hai (giá trị trung vị).
 - Giá trị tứ phân vị thứ ba.
 - Giá trị tối đa.

Data Exploring & Preprocessing

07

Phân tích phân bố của các giá trị trong mỗi cột có kiểu dữ liệu dạng số (numerical)

Column Name	Missing Ratio (Percent)	Minimum	Lower Quartile (Q1)	Median (Q2)	Upper Quartile (Q3)	Maximum
authorStats.diggCount	0.000	16.0	1188.0	4259.0	11100.0	173000.0
authorStats.followerCount	0.000	0.0	82300.0	199900.0	381100.0	3000000.0
authorStats.followingCount	0.000	0.0	26.0	79.0	241.0	9398.0
authorStats.friendCount	0.000	0.0	0.0	0.0	0.0	0.0
authorStats.heart	0.000	71000.0	2100000.0	4800000.0	10000000.0	100200000.0
authorStats.heartCount	0.000	71000.0	2100000.0	4800000.0	10000000.0	100200000.0
authorStats.videoCount	0.000	13.0	408.0	607.0	885.0	2298.0
music.duration	0.168	1.0	57.0	81.0	130.0	1481.0
stats.collectCount	0.000	0.0	73.0	277.0	1017.0	138500.0
stats.commentCount	0.000	0.0	16.0	47.0	133.0	46900.0
stats.diggCount	0.000	0.0	635.0	2632.0	9227.0	1200000.0
stats.playCount	0.000	0.0	32800.0	108700.0	358225.0	36400000.0
stats.shareCount	0.000	0.0	27.0	128.0	572.0	190900.0
statsV2.collectCount	0.000	0.0	73.0	277.0	1017.0	138474.0
statsV2.commentCount	0.000	0.0	16.0	47.0	133.0	46900.0
statsV2.diggCount	0.000	0.0	635.0	2632.0	9227.0	1200000.0
statsV2.playCount	0.000	0.0	32800.0	108700.0	358225.0	36400000.0

Data Exploring & Preprocessing

07

Phân tích phân bố của các giá trị trong mỗi cột có kiểu dữ liệu dạng số (numerical)

- Nhận xét:

- Một vài cột có phân phối khá giống nhau:
 - Hai cột `authorStats.heart` và `authorStats.heartCount`
 - Các cột bắt đầu bằng `stats.*` và `statsV2.*`



Ta sẽ thực hiện các phân tích để kiểm tra xem các cột này có chứa thông tin trùng lặp hay không:

- Với các cột bị trùng lặp thì ta chỉ cần giữ lại 1 cột

08

Phân tích phân bố của các giá trị trong mỗi cột có kiểu dữ liệu không phải dạng số (non-numerical)

- Với mỗi cột có kiểu dữ liệu không phải dạng số, ta sẽ tính:
 - Tỷ lệ thiếu giá trị (từ 0 đến 100).
 - Số lượng các giá trị khác nhau.
 - Tỷ lệ xuất hiện (từ 0 đến 100) của mỗi giá trị.

Data Exploring & Preprocessing

08

Phân tích phân bố của các giá trị trong mỗi cột có kiểu dữ liệu không phải dạng số (non-numerical)

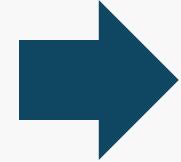
Column Name	Missing Ratio (Percent)	Number of Unique Values	Value Ratios
CategoryType	0.0	21	{'111.0': 87.1, '0.0': 5.6, '120.0': 2.6, '105.0': 2.4, '102.0': 0.4, '104.0': 0.3}
author.commentSetting	0.0	1	{'0.0': 100.0}
author.downloadSetting	0.0	2	{'0.0': 76.4, '3.0': 23.6}
author.duetSetting	0.0	3	{'0.0': 91.7, '3.0': 6.8, '1.0': 1.5}
author.ftc	0.0	1	{'False': 100.0}
author.id	0.0	264	{'7163686821477729306': 1.3, '6922569104567010305': 1.3, '7345796533974270994': 1.2, '6741580738694775809': 1.1, '7297243221116109830': 1.0, '7039230412157699074': 1.0}
author.isADVirtual	0.0	1	{'False': 100.0}
author.isEmbedBanned	0.0	1	{'False': 100.0}
author.nickname	0.0	263	{'Muoidian': 1.3, 'TÍNH TUNG TĂNG': 1.3, 'Thích Gì Ăn Đó': 1.2, 'Đi Cùng Phúc': 1.1, 'tunauan68': 1.0, 'Hôm nay có chút rảnh': 1.0}
author.openFavorite	0.0	2	{'False': 98.4, 'True': 1.6}
author.privateAccount	0.0	1	{'False': 100.0}
author.relation	0.0	1	{'0.0': 100.0}

Data Exploring & Preprocessing

08

Phân tích phân bố của các giá trị trong mỗi cột có kiểu dữ liệu không phải dạng số (non-numerical)

- Nhận xét:
 - Có nhiều cột **chỉ chứa 1 giá trị duy nhất**. Các cột này không mang lại nhiều thông tin hữu ích trong quá trình phân tích



Ta sẽ loại bỏ các cột này khỏi tập dữ liệu

09 Loại bỏ các cột không có nhiều ý nghĩa để phân tích

1. Các cột chứa giá trị **ID** (*nhưng giữ lại ID của tác giả và video*)
2. Các cột chứa thông tin **bị trùng lặp**
3. Các cột chỉ **có duy nhất 1 giá trị**
4. Loại bỏ các cột chứa thông tin về **ngôn ngữ gốc của video**

Kết quả:

Đã loại bỏ tổng cộng **37** cột.
Tập dữ liệu còn lại **50** cột.

10 Điền giá trị thích hợp vào các cột bị thiếu giá trị

- Đối với các cột có kiểu dữ liệu số:
 - Điền giá trị **trung vị** (của các giá trị không bị thiếu)
- Đối với các cột có kiểu dữ liệu không phải số:
 - Điền 1 giá trị đặc biệt là "**others**"

➡ Cách này giúp giữ lại đặc điểm phân phối của dữ liệu gốc
nhiều nhất có thể, hạn chế ảnh hưởng đến phân tích sau này

11 Tổng kết quá trình tiền xử lý dữ liệu

- Sau quá trình tiền xử lý:
 - Tập dữ liệu còn lại **70996 hàng và 50 cột**
 - Các cột đều **có kiểu dữ liệu thích hợp và không bị thiếu giá trị**
- Tập dữ liệu sau khi tiền xử lý được lưu thành định dạng **Parquet**



Giúp giữ nguyên kiểu dữ liệu phù hợp cho mỗi cột
trong các lần sử dụng tiếp theo



03

**FEATURE
ENGINEERING**

Feature Engineering

01 Trích xuất các hashtag từ mô tả của video

- Các hashtag được đính kèm trong mô tả của video, bắt đầu với ký tự “#”

Description: Review chi nhánh Haidilao Vincom Đồng Khởi #Haidilao #tiktokfood #ăncùngtiktok

- Trích xuất danh sách hashtag, chuẩn hóa bằng cách: **viết thường và bỏ dấu câu**

Hashtags: ['haidilao', 'tiktokfood', 'ancungtiktok']

Feature Engineering

01 Trích xuất các hashtag từ mô tả của video

- Phân tích số lượng hashtag trong mỗi video:

Total number of videos: **70996**

Total videos **with hashtags**: **70650**

Percentage of videos **with hashtags**: **99.51%**

Average number of hashtags per video: **6.85**



Mỗi video trên TikTok có trung bình từ 6-7 hashtag

Feature Engineering

01 Trích xuất các hashtag từ mô tả của video

- Các hashtag phổ biến nhất thường liên quan đến ẩm thực

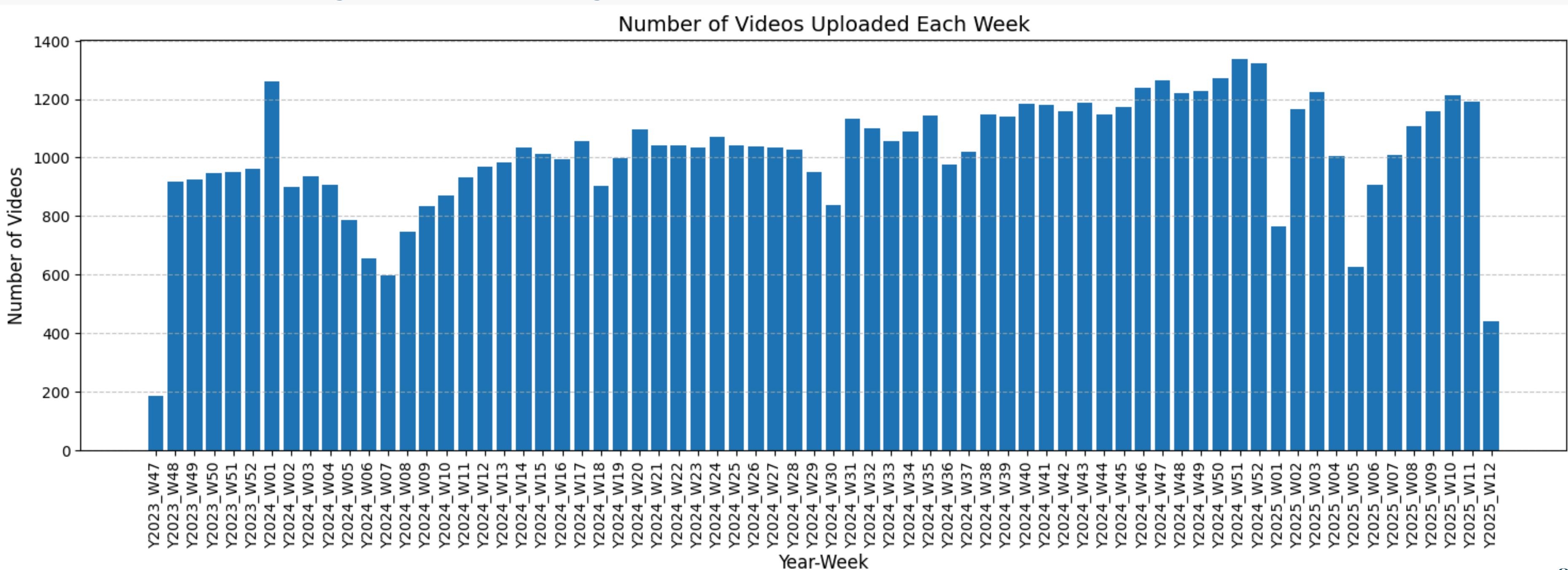
➡ Điều này phù hợp với cách mà ta thu thập dữ liệu

===== Top 20 hashtags =====	
ancungtiktok	: 50231
learnontiktok	: 30866
reviewanngon	: 22232
xuhuong	: 20930
mukbang	: 9876
fyp	: 9668
vtmgr	: 8377
food	: 7590
viral	: 7285
monngonmoingay	: 6154
foodreview	: 4863
xuhuongtiktok	: 4156
trending	: 4035
foodtiktok	: 3834
nauancungtiktok	: 3591
review	: 3430
tiktokfood	: 2955
anvat	: 2902
homnayangi	: 2801
thanhthoiluottet	: 2626

Feature Engineering

02 Trích xuất nội dung từ top video trong mỗi tuần

- Tập dữ liệu gồm video trong 70 tuần



Feature Engineering

02 Trích xuất nội dung từ top video trong mỗi tuần

- Lọc dữ liệu theo từng tuần

- Tập dữ liệu gồm video trong **70 tuần**
- Sử dụng **20% video có xếp hạng cao nhất trong mỗi tuần** để trích xuất các đặc trưng về nội dung



Tập dữ liệu dự kiến sẽ có hơn **14000 hàng**

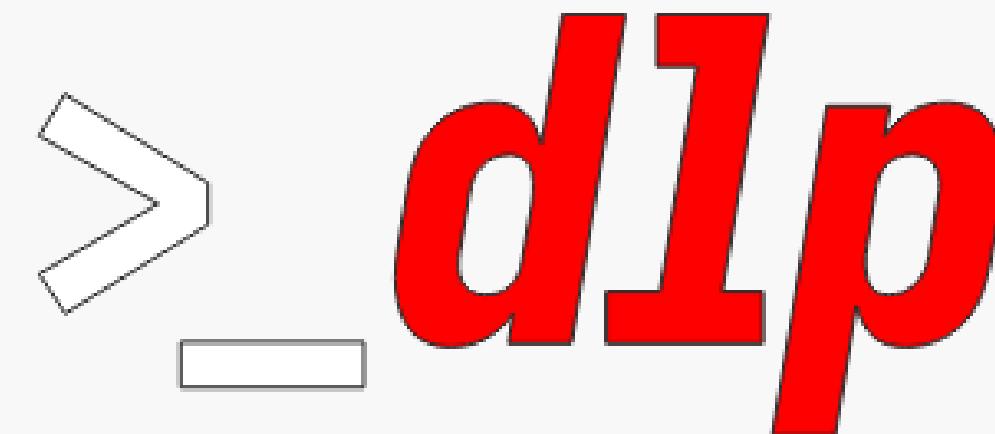
Feature Engineering

02 Trích xuất nội dung từ top video trong mỗi tuần

- **Độ đo chọn video tốt nhất theo tuần:** tổng có trọng số của lượt xem và các chỉ số tương tác

```
columns_weights = {  
    'statsV2.playCount': 0.40,  
    "engagement_rate": 0.1,  
    "statsV2.shareCount": 0.15,  
    "statsV2.commentCount": 0.10,  
    "statsV2.diggCount": 0.25  
}
```

Feature Engineering Pipeline



YT-DLP A youtube-dl fork with additional features and fixes

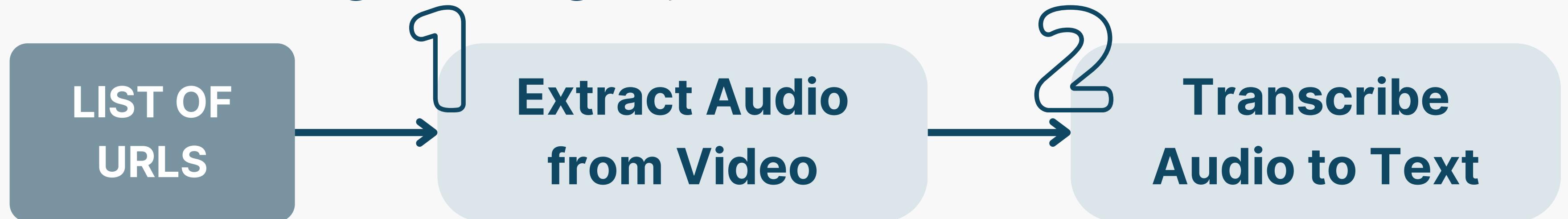
1

Extract Audio from Video

- **INPUT:** url (str)
- **OUTPUT:** file *.wav
 - Tải file audio về máy thay vì file video (sẽ nhẹ hơn)

```
# Download the audio from the YouTube video
print(f"Downloading audio from YouTube: {url}")
ydl_opts = {
    'format': 'bestaudio/best',
    'postprocessors': [{
        'key': 'FFmpegExtractAudio',
        'preferredcodec': 'wav',
    }],
    'outtmpl': output_path,
    'keepvideo': True,
}
with yt_dlp.YoutubeDL(ydl_opts) as ydl:
    try:
        ydl.download([url])
    except Exception as e:
        print(f"Error downloading audio: {e}")
return None
```

Feature Engineering Pipeline



Gemini

2

Transcribe Audio to Text

- **INPUT:** file *.wav
 - Đọc file âm thanh dưới dạng byte
- **MODEL:** Gemini-2.0-flash
- **OUTPUT:** json file

```
# Open the audio file and read the content
with open(wav_file, 'rb') as f:
    image_bytes = f.read()

try:
    # Call the API to generate content
    response = client.models.generate_content(
        model='gemini-2.0-flash',
        contents=[
            prompt,
            types.Part.from_bytes(
                data=image_bytes,
                mime_type='audio/wav',)])
```

2

Transcribe Audio to Text

- **PROMPT:**

- Đoạn mô tả nhiệm vụ ngắn gọn

```
prompt = ""  
Generate a transcript of the speech. The speech is in Vietnamese.  
If there is no speech in the file, return None.  
  
Then generate 3 takeaways from the speech.  
The takeaways should be concise and informative, written in Vietnamese.  
  
Check if the speech contains calls to action (CTA) sentences.  
Check if the speech contains elements of curiosity gap.  
  
Return the results in JSON format with fields:  
{  
    "transcript": "The transcript of the speech",  
    "takeaways": ["Takeaway 1", "Takeaway 2", "Takeaway 3"],  
    "has_call_to_action": true/false,  
    "has_curiosity_gap": true/false  
}  
"""
```

Feature Engineering Pipeline



Gemini

3

Extract Food, Location

- **INPUT:**
 - description
 - transcript
- **MODEL:** Gemini-2.0-flash
- **OUTPUT:** json format
- **PROMPT:**

```
prompt = """
```

Bạn là một chuyên gia trong lĩnh vực phân tích dữ liệu thông minh. Bạn có thể phân tích nội dung video và trích xuất thông tin từ đó. Người dùng sẽ cung cấp thông tin về mô tả và transcript của video TikTok về chủ đề ẩm thực. Hãy trích xuất thông tin về các món ăn và địa điểm được đề cập trong video đó. Bạn sẽ trả về một JSON chứa các thông tin sau:

```
{
```

"foods": Danh sách các món ăn được đề cập trong video (danh sách string),
 "city": Tên thành phố (string),
 "district": Tên quận/huyện (string)

```
}
```

Đây là một số lưu ý quan trọng:

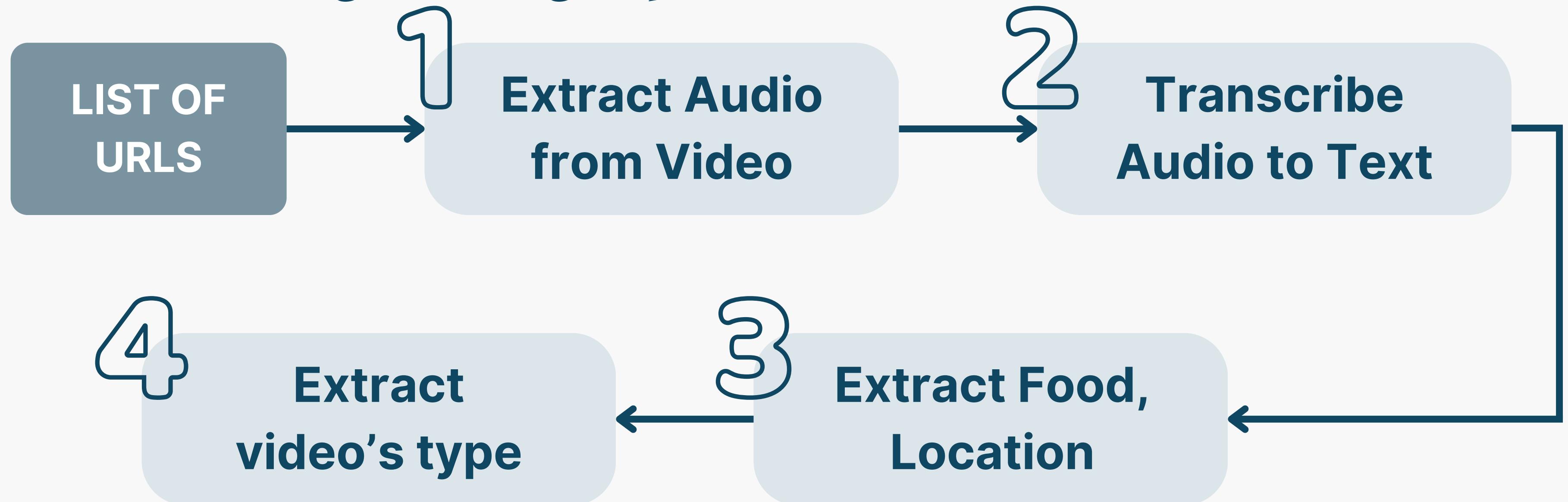
- Đối với các video không thuộc chủ đề ẩm thực thì hãy trả về None cho tất cả các trường
- Các thông tin trả về đều phải có đúng định dạng
- Trả về kết quả theo đúng định dạng JSON
- Ưu tiên thông tin chính xác
- Câu trả lời phải bám sát theo nội dung được cung cấp

Dưới đây là mô tả và transcript của video TikTok:

- Mô tả: %s
- Transcript: %s

```
"""
```

Feature Engineering Pipeline



Gemini

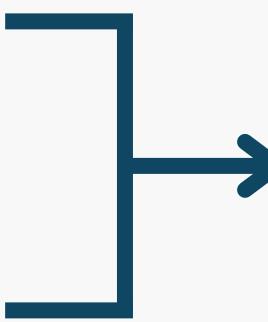
04

DASHBOARDS & WEBAPP

Dashboards & Webapp

- **4 nhóm dashboard chính:**

- Dashboard phân tích **thống kê về “user”**
- Dashboard phân tích **thống kê về “video”**
- Dashboard phân tích **xu hướng ẩm thực**
- Dashboard phân tích **nội dung video**



Yếu tố góp phần tạo nên sự thành công của video TikTok

Thông tin hữu ích cho công cụ hỗ trợ viết kịch bản

Dashboards & Webapp

- Các công nghệ chính:
 - Trực quan hóa dữ liệu: **Plotly**
 - Nhận xét biểu đồ: **Gemini API**
 - Xây dựng webapp: **Streamlit**
 - Deploy: **Streamlit Community Cloud**
- Link: <https://21khdl-tiktok-analytics.streamlit.app/>



Dashboards & Webapp

DEMO

Dùng AI nhận xét biểu đồ

01 Yêu cầu / Nhiệm vụ chính

Xác định rõ **nhiệm vụ trọng tâm** mà AI cần thực hiện

- Ví dụ: Phân tích mối quan hệ giữa các biến số đã cho

```
correlation_analysis_prompt = f"""
```

```
Hãy phân tích mối tương quan giữa 'Số người theo dõi', 'Số lượt thích'  
và 'Số lượng video' của các TikToker dựa trên dữ liệu được cung cấp."""
```

Dùng AI nhận xét biểu đồ

02

Hướng dẫn chi tiết & Ràng buộc kết quả

- Chỉ rõ nội dung và các khía cạnh cụ thể cần AI tập trung phân tích
- Xác định cấu trúc và định dạng mong muốn cho bài phân tích
→ Giúp AI tạo ra câu trả lời súc tích, đi đúng trọng tâm và dễ hiểu

```
correlation_analysis_prompt = f"""
```

Viết một đoạn phân tích súc tích (khoảng 250-350 từ) tập trung vào:

1. Mức độ tương quan (mạnh, trung bình, yếu) giữa các cặp biến
2. Hướng tương quan (dương/âm) và ý nghĩa thực tế của nó
3. Các điểm bất thường hoặc xu hướng đáng chú ý từ biểu đồ phân tán
4. Các hàm ý cho người sáng tạo nội dung TikTok

Cấu trúc phân tích nên bao gồm:

- Tổng quan về mức độ tương quan chung giữa các biến
- Phân tích chi tiết từng cặp tương quan quan trọng
- Kết luận và gợi ý thực tiễn cho người sáng tạo nội dung"""

Dùng AI nhận xét biểu đồ

03 Dữ liệu đầu vào & Thông tin tham khảo

- Bao gồm **toàn bộ dữ liệu cần thiết** cho phân tích (biểu đồ, bảng số liệu)
- Cung cấp **các định nghĩa hoặc quy tắc tham khảo** (ví dụ: thang đo tương quan)
→ Giúp AI có **cơ sở đầy đủ và chính xác** để thực hiện phân tích một cách khách quan

Định dạng dữ liệu phổ biến:

- **Ảnh**
→ Chuỗi byte
- **Bảng thống kê (DataFrame/Series)**
→ Chuỗi LaTeX/Markdown
- **Scalar value**
→ Truyền trực tiếp

```
correlation_analysis_prompt = f"""
```

Dữ liệu phân tích:

1. Biểu đồ phân tán thể hiện mối quan hệ giữa ba chỉ số.
Biểu đồ này sẽ được đính kèm dưới dạng byte:
`{scatter_fig.to_image()}`
2. Bảng ma trận tương quan giữa các chỉ số (hệ số Pearson).
Dưới đây là bảng thống kê thể hiện các thông tin này dưới dạng LaTeX:
`{get_correlation_matrix(select_columns(df, METRICS)).to_latex()}`
3. Thông tin bổ sung:
 - Hệ số tương quan từ 0.7-1.0: tương quan mạnh
 - Hệ số tương quan từ 0.3-0.7: tương quan trung bình
 - Hệ số tương quan từ 0.0-0.3: tương quan yếu""""

Mục tiêu cần đạt được (tùy báo cáo giữa kỳ)



Tích hợp các webapp thành 1 website duy nhất

- Giúp người dùng không cần chuyển đổi giữa các trang web



Viết 1 homepage để giới thiệu về dự án, hướng dẫn người dùng sử dụng

- **Ví dụ:** Nếu người dùng muốn xem thống kê về 1 TikToker nào đó thì họ phải làm gì? (*Nêu rõ các bước hướng dẫn*)



Dùng AI rút ra insight trong mỗi dashboard

- Insight sẽ được hiển thị bên dưới mỗi dashboard



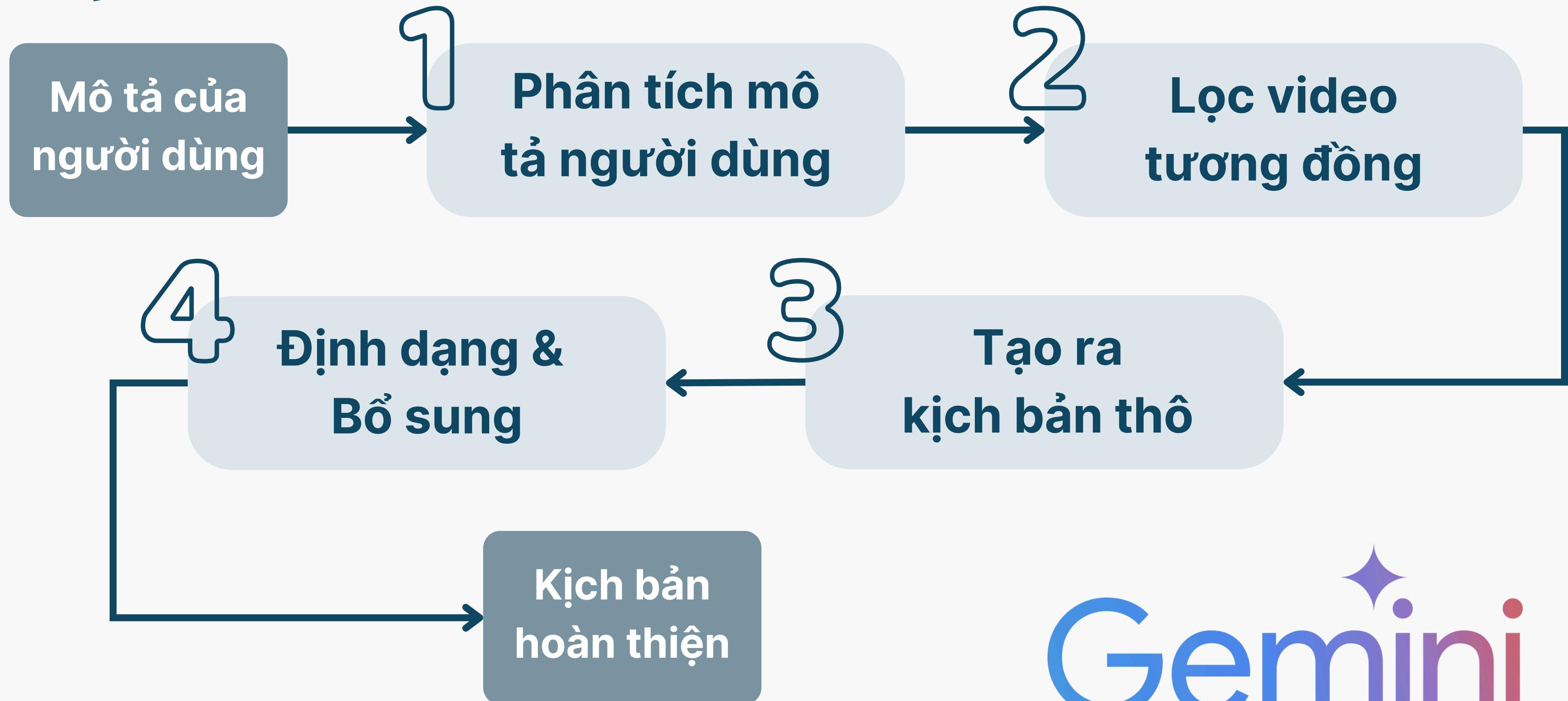
Đưa insight về nội dung video vào prompt viết kịch bản

- Giúp nội dung được tạo ra phù hợp với thị hiếu của người xem hơn

05

SCRIPTWRITING ASSISTANT WEBAPP

Pipeline tạo Kịch bản



Gemini

0

Trích xuất đặc trưng nội dung

Đầu tiên, dùng Gemini gán nhãn tự do cho 100 video ngẫu nhiên

- Không giới hạn số lượng
- Không định trước các nhãn

Sau khi thu thập đủ dữ liệu:

- Thống kê tần suất xuất hiện các nhãn
- Chọn ra các nhãn phổ biến nhất cho mỗi trường
- Tiến hành tinh chỉnh thủ công để xây dựng tập nhãn ổn định

0

Trích xuất đặc trưng nội dung

Gán nhãn theo một response schema chuẩn hoá, gồm các trường:

- Thể loại nội dung
- Cách xây dựng nội dung
- Kiểu mở đầu video
- Giọng điệu của người nói
- Nhịp độ lời thoại
- Kiểu kêu gọi hành động
- Phong cách tổng thể của video
- Nhóm khán giả mục tiêu

```
{  
    "type": "object",  
    "description": "Gán nhãn nội dung video TikTok ẩm thực và trích xuất các đặc điểm nổi bật từ transcript.",  
    "required": ["categories"],  
    "properties": {  
        "categories": {  
            "type": "string",  
            "description": "Thể loại chính của video. Nếu kết quả là 'Không liên quan ẩm thực' thì trả về duy nhất trường 'categories', không cần trích xuất các trường khác",  
            "enum": [  
                "Review quán ăn",  
                "Review sản phẩm ăn uống",  
                "Review món ăn",  
                "Mukbang",  
                "Nấu ăn",  
                "Không liên quan ẩm thực"  
            ]  
        },  
        ...  
    }  
}
```

1

Phân tích mô tả người dùng

MODEL: Gemini-2.0-flash

INPUT: Mô tả tự nhiên từ người dùng

OUTPUT: JSON gồm

- Các đặc điểm nội dung mong muốn
- Thời lượng video (nếu có)

```
{  
    "type": "object",  
    "description": "Các trường dùng để gán nhãn (label). Gán nhãn dựa trên nội dung miêu tả video mà người dùng cung cấp.",  
    "required": ["categories"],  
    "properties": {  
        "categories": {  
            "type": "string",  
            "description": "Thể loại chính của video. Nếu kết quả là 'Không liên quan ẩm thực' thì trả về duy nhất trường 'categories', không cần trích xuất các trường khác",  
            "enum": [  
                "Review quán ăn",  
                "Review sản phẩm ăn uống",  
                "Review món ăn",  
                "Mukbang",  
                "Nấu ăn",  
                "Không liên quan ẩm thực"  
            ]  
        },  
        ...  
        "duration": {  
            "type": "integer",  
            "description": "Thời lượng mong muốn của video tính theo giây. Ví dụ: 180 (tương đương 3 phút)"  
        }  
    }  
}
```

1

Phân tích mô tả người dùng

Mô tả của người dùng:

“Tôi muốn làm video **quảng bá cho sản phẩm** bánh tráng chấm phô mai của nhà tôi,
cách nói chuyện gần gũi, có **hướng dẫn** cách ăn, **giọng điệu từ tốn**.”

Annotate mô tả:

- “categories”: “Review sản phẩm ăn uống”
- “structure_style”: [“Hướng dẫn”]
- “tone_of_voice”: [“Chân thành”, “Thân thiện”]
- “content_style”: [“Đời thường”]
- “pacing”: [“Chậm”]

2

Lọc video tương đồng



Mục tiêu: Tìm ít nhất 20 video tương đồng trong tập dữ liệu

Bước 1: Lọc đầy đủ tất cả trường

→ Chỉ tìm được 6 video

Bước 2: Lần lượt *nới lỏng* một vài nhãn có *độ ưu tiên thấp* (ví dụ “content_style”)

→ Tăng lên 14 video

Bước 3: Giữ lại “categories” + “structure_style” (chính yếu)

→ Đạt 32 video

→ Sử dụng nhóm này làm ví dụ đầu vào cho mô hình sinh kịch bản

Annotate mô tả:

- “categories”: “Review sản phẩm ăn uống”
- “structure_style”: [“Hướng dẫn”]
- “tone_of_voice”: [“Chân thành”, “Thân thiện”]
- “content_style”: [“Đời thường”]
- “pacing”: [“Chậm”]

3

Tạo ra kịch bản thô

MODEL: Gemini-2.0-flash

INPUT: Mô tả người dùng + 20 video tương đồng
+ số từ giới hạn

OUTPUT: Đoạn kịch bản thô (plain text)

Quy trình:

- Hệ thống sử dụng 20 video mẫu có **điểm cao** để làm ví dụ minh họa.
- Gemini viết một đoạn văn tự nhiên (kịch bản) theo nội dung mô tả.
- **Giới hạn độ dài** dựa trên:
 - Thời lượng mong muốn của người dùng (nếu có)
 - **Số từ = thời lượng trung bình × tốc độ nói trung bình** của các video mẫu

$$\text{Score} = \frac{\text{Lượt xem}}{(\text{Số ngày đăng} + 1)^{0.7}}$$

3

Tạo ra kịch bản thô

MODEL: Gemini-2.0-flash

INPUT: Mô tả người dùng + 20 video tương đồng
+ số từ giới hạn

OUTPUT: Đoạn kịch bản thô (plain text)

Prompt:

Bạn là chuyên gia viết kịch bản video TikTok trong lĩnh vực ẩm thực.

Hãy viết một kịch bản video TikTok dạng **lời thoại tự nhiên** dựa trên **các transcript mẫu bên dưới** và **mô tả món ăn từ người dùng**.

Yêu cầu:

- Kịch bản có độ dài khoảng **{word_count}** từ, không được chênh lệch quá 100 từ.
- Viết theo dạng **lời thoại tự nhiên**, như thể đang nói trong video TikTok.
- **Không chia phần**, **không thêm tiêu đề**, **không mở ngoặc giải thích** hoặc mô tả bối cảnh.
- **Không chèn chú thích** như (cảnh quay), (hình ảnh), (âm thanh).
- Sử dụng **giọng điệu**, cách nói, tốc độ, hook và CTA tương tự các transcript mẫu.
- Ưu tiên sử dụng cụm từ đồi thường, dễ viral như "Trời ơi ngon gì đâu luôn á", "ăn là ghiền", v.v.

Mô tả từ người dùng:

{user_input}

Các transcript mẫu để tham khảo phong cách viết:

{transcript_sample_text}

3

Tạo ra kịch bản thô

MODEL: Gemini-2.0-flash

INPUT: Mô tả người dùng + 20 video tương đồng
+ số từ giới hạn

OUTPUT: Đoạn kịch bản thô (plain text)

Kịch bản thô:

Trời ơi tin được không mọi người ơi! Tui mới khám phá ra một cái quán bánh mì vỉa hè mà nó ngon dữ thần luôn á! Thiệt sự là kiểu ngon mà xỉu ngang xỉu dọc á.

Hôm nay tui dẫn mọi người đi ăn bánh mì mà kiểu bánh mì ổ bánh mì quốc dân luôn á. Ai mà không biết ăn bánh mì vỉa hè thì chắc chắn là chưa phải dân Sài Gòn rồi. Mà cái quán này nha, tui nói thiệt, nó nằm lọt thỏm trong hẻm thõi, vậy mà lúc nào cũng đông nghẹt khách.

Để tui quay cận cảnh cho mọi người coi nè. Đó, thấy ổ bánh mì vàng ruộm chưa? Nhìn thõi là muốn xỉu liền. Mà giá thì thõi khỏi bàn luôn, rẻ như cho á. Thời buổi này kiếm đâu ra đồ ăn vừa ngon vừa rẻ như vầy nữa.

...

4

Định dạng & Bổ sung

MODEL: Gemini-2.0-flash

INPUT: Kịch bản thô + 20 đoạn mô tả video mẫu + hashtag phổ biến + thời lượng

OUTPUT: JSON có cấu trúc gồm các phần: *main_content*, *duration*, *setting*, *characters*, *video_description*

Quy trình:

- Sinh mô tả video và hashtag dựa trên **mô tả mẫu và hashtag phổ biến**.
- Bổ sung thông tin về thời lượng dự kiến, bối cảnh, nhân vật.
- Chia kịch bản thô thành các đoạn có nhãn (tên, miêu tả về cảnh quay, lời thoại, mốc thời gian).
- Cuối cùng, hệ thống tính lại **mốc thời gian** thủ công để đảm bảo độ chính xác.

4

Định dạng & Bổ sung

Prompt:

MODEL: Gemini-2.0-flash

INPUT: Kịch bản thô + 20 đoạn mô tả video mẫu + hashtag phổ biến + thời lượng

OUTPUT: JSON có cấu trúc gồm các phần: *main_content*, *duration*, *setting*, *characters*, *video_description*

Bạn là chuyên gia viết kịch bản TikTok âm thực.

Hãy chuyển kịch bản dạng plain dưới đây thành format theo schema JSON sau, giữ nguyên lời thoại gốc, chia nhỏ theo từng bước nội dung như mở đầu, mô tả món, cảm nhận, CTA,...

Thông tin thêm:

- duration: {duration_text}

Top 10 hashtag được sử dụng nhiều nhất: {top_10_cat_hashtags_text}

Các đoạn mô tả video mẫu:

{desc_sample_text}

Kịch bản plain:

{plain_script}

4

Định dạng & Bổ sung

Kịch bản có cấu trúc:

```
{'characters': 'Một người review quán bánh mì vỉa hè.',  
 'duration': '2 phút 11 giây',  
 'main_content': [{  
     'time_range': '0:00-0:15',  
     'title': 'Mở đầu và giới thiệu quán bánh mì',  
     'visual_description': 'Cảnh người quay phim bắt ngờ khám phá ra quán bánh mì vỉa hè.',  
     'dialogue': 'Trời ơi tin được không mọi người ơi! Tui mới khám phá ra một cái quán bánh mì vỉa hè mà  
 nó ngon dữ thần luôn á! Thiệt sự là kiểu ngon mà xiêu ngang xiêu dọc á.'},  
     {'time_range': '0:15-0:30',  
     'title': 'Giới thiệu về bánh mì vỉa hè',  
     'visual_description': 'Hình ảnh quán bánh mì nằm trong hẻm nhỏ, đông khách.',  
     'dialogue': 'Hôm nay tui dẫn mọi người đi ăn bánh mì mà kiểu bánh mì ổ bánh mì quốc dân luôn á. Ai  
 mà không biết ăn bánh mì vỉa hè thì chắc chắn là chưa phải dân Sài Gòn rồi. Mà cái quán này nha, tui  
 nói thiệt, nó nằm lọt thỏm trong hẻm thôi, vậy mà lúc nào cũng đông nghẹt khách.'},  
     ...  
 }]
```

MODEL: Gemini-2.0-flash

INPUT: Kịch bản thô + 20 đoạn mô tả video mẫu + hashtag phổ biến + thời lượng

OUTPUT: JSON có cấu trúc gồm các phần: *main_content*, *duration*, *setting*, *characters*, *video_description*

Scriptwriting Assistant Webapp

DEMO

06

DISCUSSION & CONCLUSION

Conclusion



Tích hợp các webapp thành 1 website duy nhất

- Giúp người dùng không cần chuyển đổi giữa các trang web



Viết 1 homepage để giới thiệu về dự án, hướng dẫn người dùng sử dụng

- **Ví dụ:** Nếu người dùng muốn xem thống kê về 1 TikToker nào đó thì họ phải làm gì? (*Nêu rõ các bước hướng dẫn*)



Dùng AI rút ra insight trong mỗi dashboard

- Insight sẽ được hiển thị bên dưới mỗi dashboard



Đưa insight về nội dung video vào prompt viết kịch bản

- Giúp nội dung được tạo ra phù hợp với thị hiếu của người xem hơn

Discussion

01 **Trình bày về các dashboard phân tích user và hashtag**

- Nhóm sẽ trình bày thêm các dashboard này nếu còn dư thời gian

02 **Xây dựng công cụ nghiên cứu chủ đề và đề xuất quay video**

- Nhóm sẽ demo về các công cụ này nếu dư thời gian

03 **Trang “Giới thiệu về nhóm” (About Us)**

- Giới thiệu về các thành viên tham gia vào quá trình xây dựng sản phẩm

04 **Trang “Gợi ý để tăng hiệu suất” (Conclusion)**

- Trình bày các yếu tố tạo nên sự thành công của 1 video TikTok

07

FUTURE DEVELOPMENT

Future Development

- 01 Mở rộng quy trình phân tích cho nhiều chủ đề khác**
 - Giúp sản phẩm có thể phục vụ cho nhiều người dùng hơn
- 02 Tích hợp airflow vào quy trình khoa học dữ liệu**
 - Giúp tự động hóa quy trình: thu thập → xử lý → đẩy dữ liệu lên cloud
- 03 Dùng AI để tạo báo cáo hoàn chỉnh từ dữ liệu**
 - Tích hợp insight từ việc nhận xét biểu đồ vào báo cáo tổng quát
- 04 Thử nghiệm với API từ nhiều mô hình khác nhau**
 - Kiểm tra hiệu suất tạo kịch bản từ nhiều mô hình tiên tiến khác



Thank you

