THIẾT KẾ VÀ THU THẬP DATASET

PHÂN TÍCH XU HƯỚNG NỘI DUNG TRÊN TIKTOK VÀ YOUTUBE

Học phần: ADY201m – AI, Data Science with Python & SQL
Sinh viên thực hiện:

- QE200009 – Dương Thị Mỹ Tâm
- QE200083 – Trần Vân Khánh

TÓM TẮT

Sự phát triển mạnh mẽ của các nền tảng video trực tuyến đã làm thay đổi cách người dùng tiếp cận và tiêu thụ nội dung giải trí. Trong đó, **TikTok** và **YouTube** là hai nền tảng có ảnh hưởng lớn nhất, đại diện cho hai dạng nội dung khác nhau: video ngắn mang tính lan truyền nhanh và video dài mang tính hệ thống, bền vững.

Báo cáo này trình bày **đề xuất thiết kế và thu thập dataset chỉ dựa trên hai nguồn dữ liệu thực tế đã thu được là TikTok và YouTube**, nhằm phục vụ cho việc phân tích xu hướng nội dung số. Dataset được xây dựng từ dữ liệu công khai, tập trung vào metadata định lượng (thời điểm đăng tải, mức độ tương tác, hashtag, tiêu đề), tạo nền tảng cho các bước phân tích EDA, kiểm định giả thuyết và mô hình hóa trong các báo cáo tiếp theo.

1. GIỚI THIỆU

Xu hướng nội dung số (Digital Content Trends) chịu ảnh hưởng mạnh mẽ từ thuật toán đề xuất, hành vi người dùng và tốc độ lan truyền trên mạng xã hội. TikTok và YouTube hiện là hai nền tảng đại diện rõ rệt cho hai cơ chế hình thành xu hướng:

- **TikTok:** Xu hướng hình thành nhanh, vòng đời ngắn, phụ thuộc mạnh vào hashtag, âm thanh và tương tác sớm.
- **YouTube:** Xu hướng ổn định hơn, phụ thuộc vào thời điểm đăng tải, tiêu đề, mức độ tương tác tích lũy và thuật toán Trending.

Do giới hạn thực tế của việc thu thập dữ liệu, nghiên cứu này **chủ động thu hẹp phạm vi** chỉ tập trung vào hai nền tảng trên, nhằm đảm bảo chất lượng dữ liệu, khả năng kiểm soát pipeline và tính khả thi của dự án.

2. MỤC TIÊU NGHIÊN CỨU

Các mục tiêu chính của báo cáo bao gồm:

- Thu thập và xây dựng dataset từ **TikTok và YouTube** dựa trên dữ liệu công khai.
- Chuẩn hóa dữ liệu về cùng một cấu trúc metadata để so sánh chéo giữa hai nền tảng.
- Phân tích mối quan hệ giữa **thời điểm đăng tải – nội dung – mức độ tương tác – khả năng viral/trending**.
- Đặt nền móng cho các phân tích thống kê, EDA và mô hình dự đoán xu hướng trong các báo cáo tiếp theo.

3. PHƯƠNG PHÁP THU THẬP DỮ LIỆU

3.1. Nguyên tắc thiết kế dataset

Dataset được xây dựng dựa trên các nguyên tắc:

- **Dữ liệu thực tế:** Chỉ sử dụng dữ liệu đã thu thập được từ TikTok và YouTube.
- **Định lượng hóa:** Ưu tiên các biến có thể đo lường (likes, comments, views, publish time).
- **Chuẩn hóa:** Đồng nhất kiểu dữ liệu, múi giờ, định dạng thời gian.
- **Khả năng mở rộng:** Dễ dàng bổ sung dữ liệu theo thời gian.

3.2. Nguồn dữ liệu và cách thu thập

3.2.1. TikTok (Hashtag-based Crawling)

- **Phương pháp:** Sử dụng Python + Playwright để crawl dữ liệu công khai từ trang tìm kiếm hashtag.
- **Đơn vị dữ liệu:** Video TikTok theo hashtag.
- **Quy mô:** Hàng nghìn video từ nhiều hashtag trending và phổ biến.
- **Trường dữ liệu chính:**
    - hashtag
    - caption
    - publish_time
    - likes

- comments
- is_trending (0/1 – suy diễn)
- has_clickbait (0/1 – dựa trên từ khóa và ký tự đặc biệt)

TikTok được sử dụng để phản ánh **hành vi viral ngắn hạn** và tác động của nội dung giật gân, hashtag và tương tác sớm.

3.2.2. YouTube (Trending-based Dataset)

- **Phương pháp:** Sử dụng YouTube Data API để thu thập danh sách video Trending.
- **Đơn vị dữ liệu:** Video YouTube trong mục Trending.
- **Quy mô:** Hàng nghìn video theo thời gian và khu vực.
- **Trường dữ liệu chính:**
  - video_id
  - title
  - category
  - publish_time
  - views
  - likes
  - comments

YouTube được sử dụng để phản ánh **xu hướng ổn định và dài hạn**, nơi yếu tố thời điểm đăng tải và tiêu đề đóng vai trò quan trọng.

4. BÀI TOÁN NGHIÊN CỨU

4.1. Câu hỏi nghiên cứu tổng quát

Những yếu tố nào ảnh hưởng đáng kể đến khả năng một nội dung trở nên viral hoặc xuất hiện trong danh sách Trending trên TikTok và YouTube?

4.2. Biến nghiên cứu

Biến độc lập:

- Thời điểm đăng tải (giờ, ngày trong tuần)

- Hashtag (TikTok)
- Tiêu đề / caption
- Lượt likes
- Lượt comments
- Chỉ báo clickbait

Biến phụ thuộc:

- Trạng thái viral / trending (0/1)
- Mức độ tương tác (engagement proxy)

5. GIẢ THUYẾT NGHIÊN CỨU

Giả thuyết 1 – Thời điểm đăng tải

- H_{01}: Thời điểm đăng tải không ảnh hưởng đến khả năng viral/trending.
- H_{11}: Nội dung đăng trong khung giờ buổi tối có xác suất viral/trending cao hơn.

Giả thuyết 2 – Nội dung giật gân (Clickbait)

- H_{02}: Caption/tiêu đề giật gân không ảnh hưởng đến mức độ tương tác.
- H_{12}: Caption/tiêu đề giật gân làm tăng lượt tương tác ban đầu.

Giả thuyết 3 – Tương tác tiêu cực

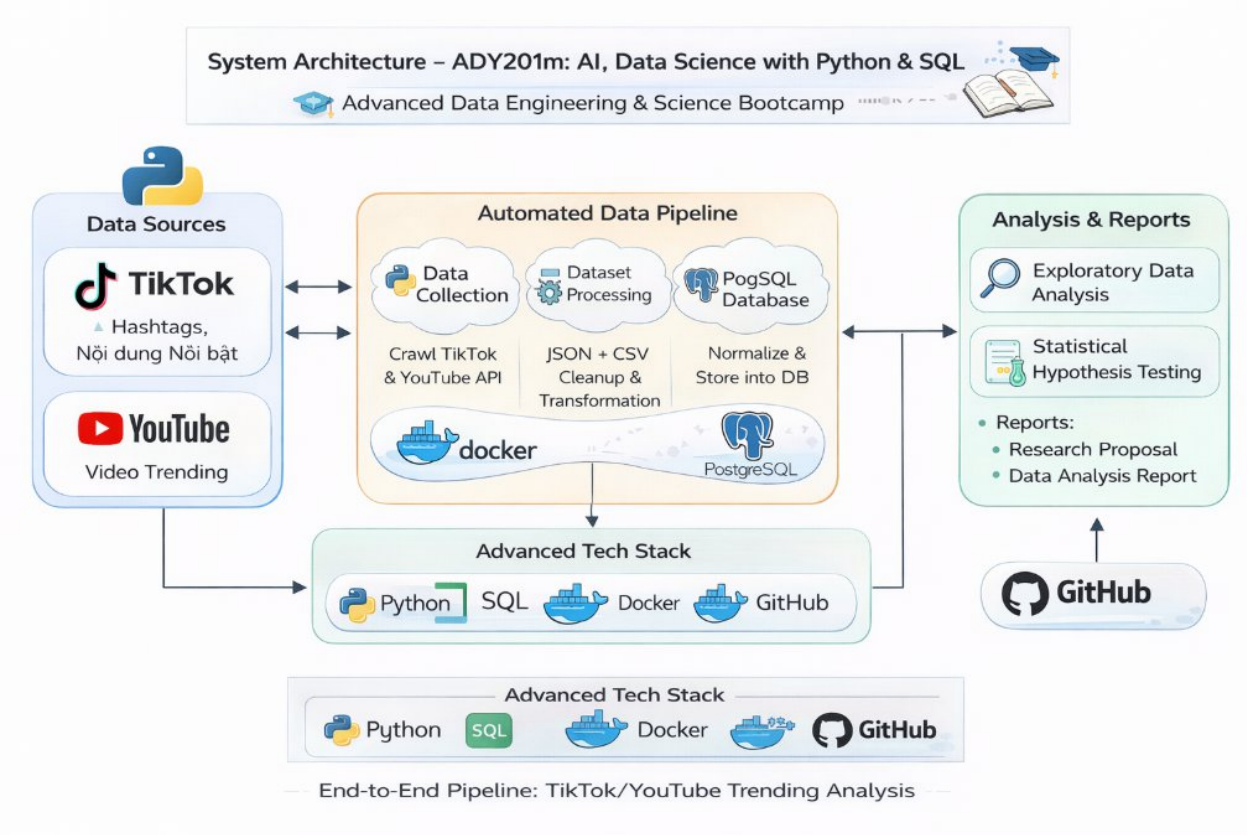
- H_{03}: Tỷ lệ tương tác thấp không ảnh hưởng đến khả năng viral.
- H_{13}: Nội dung có tương tác thấp khó duy trì trạng thái viral/trending.

6. KIẾN TRÚC HỆ THỐNG (DỰ KIẾN)

Hệ thống xử lý dữ liệu được thiết kế theo mô hình pipeline:

1. **Data Ingestion:** Python crawler / YouTube API
2. **Raw Storage:** JSON / CSV (Data Lake)
3. **Processing & Cleaning:** Chuẩn hóa, làm sạch dữ liệu
4. **Database:** SQL

5. **Analytics:** EDA, kiểm định giả thuyết, mô hình hóa



KẾT LUẬN

Báo cáo đã trình bày đề xuất **thiết kế và thu thập dataset chỉ dựa trên hai nguồn TikTok và YouTube**, phù hợp với dữ liệu thực tế đã thu thập được. Cách tiếp cận này giúp đảm bảo tính khả thi, độ tin cậy và khả năng mở rộng của nghiên cứu, đồng thời tạo nền tảng vững chắc cho các báo cáo tiếp theo trong học phần ADY201m.