THIẾT KẾ VÀ THU THẬP DATASET ĐA LĨNH VỰC CHO PHÂN TÍCH XU HƯỚNG GIẢI TRÍ

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 bùng nổ của các nền tảng nội dung số đã làm thay đổi mạnh mẽ cách con người tiêu thụ giải trí. Tuy nhiên, phần lớn các nghiên cứu hiện nay chỉ tập trung vào một lĩnh vực hoặc một nền tảng riêng lẻ, dẫn đến hiện tượng **bias domain** trong phân tích xu hướng.

Báo cáo này đề xuất một **chiến lược thiết kế và thu thập dataset quy mô lớn (>10.000 bản ghi)**, mang tính **đa lĩnh vực giải trí**, nhằm phục vụ cho việc phân tích xu hướng một cách khách quan và có khả năng kiểm định khoa học. Phương pháp tiếp cận tập trung vào **ít nền tảng nhưng nhiều category**, với hai nền tảng lõi là **YouTube và Spotify**, kết hợp các nguồn phản ánh hành vi và xã hội như TikTok, Google Trends và Reddit.

Dataset thu thập được sẽ là nền tảng cho các phân tích EDA, kiểm định giả thuyết và xây dựng mô hình trong các báo cáo tiếp theo.

1. GIỚI THIỆU

Xu hướng giải trí số (Digital Entertainment Trends) ngày càng chịu ảnh hưởng mạnh từ thuật toán đề xuất, hành vi người dùng và các yếu tố xã hội. Để nghiên cứu các xu hướng này một cách đáng tin cậy, việc xây dựng **dataset đa lĩnh vực, đủ lớn và cân bằng** là yêu cầu bắt buộc.

Trong bối cảnh đó, báo cáo này tập trung vào:

Thiết kế dataset **không lệch domain**

Thu thập dữ liệu từ **API công khai**

Tạo nền móng cho các phân tích thống kê và mô hình học máy

2. MỤC TIÊU NGHIÊN CỨU

Các mục tiêu chính của nghiên cứu bao gồm:

Thu thập **dataset >10.000 bản ghi** thuộc nhiều lĩnh vực giải trí khác nhau

Đảm bảo **cân bằng category** (âm nhạc, hài hước, video giải trí, meme, podcast, v.v.)

Phân tích xu hướng giải trí dựa trên **metadata định lượng** thay vì cảm tính

3. PHƯƠNG PHÁP THU THẬP DỮ LIỆU

3.1. Nguyên tắc thiết kế dataset

Dataset được thiết kế dựa trên các nguyên tắc sau:

Quy mô: Tối thiểu 10.000 bản ghi

Công cụ: Python, API công khai, SQL

Xử lý: Làm sạch, chuẩn hóa và thống nhất metadata

Khả năng mở rộng: Phục vụ cho EDA, kiểm định giả thuyết và modeling

3.2. Chiến lược thu thập theo nền tảng

Dataset được chia thành **ba tầng dữ liệu** nhằm đảm bảo vừa đủ sâu vừa đủ rộng:

Tầng dữ liệu	Tỷ lệ	Vai trò
Core Platforms	70%	Phản ánh xu hướng giải trí chính
Behavioral Platforms	15%	Phản ánh hành vi và viral
Extended Platforms	15%	Bổ sung góc nhìn cộng đồng

3.2.1. Nền tảng lõi – Core Platforms (70%)

YouTube Data API

- **Lý do lựa chọn:**
  YouTube cung cấp danh sách Trending đa category như Music, Comedy, Entertainment, Gaming, Film & Animation.
- **Quy mô dữ liệu:**
  Khoảng 200 video/ngày/quốc gia.
  10 quốc gia × 30 ngày → ~60.000 video (chọn lọc 30.000–50.000).
- **Dữ liệu thu thập:**
  video_id, category, views, likes, comments, publish_time

Spotify Web API

- **Lý do lựa chọn:**
  Spotify là nền tảng chuyên sâu về âm nhạc và podcast, phản ánh rõ xu hướng tiêu thụ âm nhạc toàn cầu.
- **Quy mô dữ liệu:**
  Top 50 + Viral 50/ngày/quốc gia.
  20 quốc gia × 30 ngày → ~60.000 track (chọn lọc 30.000–50.000).
- **Dữ liệu thu thập:**
  track_id, artist, popularity, tempo, energy, release_date, country

3.2.2. Nền tảng phản ánh hành vi – Behavioral Platforms (15%)

TikTok (Hashtag & Sound)

- **Vai trò:** Phản ánh các trend viral ngắn hạn như meme, challenge, video hài hước
- **Quy mô dữ liệu:**
  ~200 hashtag/sound × 50–100 video → >10.000 bản ghi
- **Dữ liệu thu thập:**
  hashtag, content_type, views, growth_rate, date

Google Trends

- **Vai trò:** Xác nhận mức độ quan tâm của xã hội đối với các xu hướng giải trí

- **Dữ liệu sử dụng:** Volume tìm kiếm theo thời gian (dữ liệu phụ trợ)

3.2.3. Nền tảng mở rộng – Extended Platforms (15%)

Threads (Meta Platforms)

- **Lý do lựa chọn:**
  Threads là nền tảng mạng xã hội dạng thảo luận theo chuỗi (thread-based), phản ánh **phản ứng xã hội tức thời** của người dùng đối với các xu hướng giải trí như video viral, âm nhạc, meme và drama. Nền tảng này cho phép quan sát rõ cách cộng đồng thảo luận, đồng thuận hoặc phản đối một xu hướng.

- **Quy mô dữ liệu:**
  Ước tính 150–200 thread/ngày.
  Thu thập trong 30 ngày → khoảng 4.500–6.000 thread.

- **Phương pháp thu thập:**
  Thu thập dữ liệu công khai dựa trên topic, keyword và mức độ tương tác, sau đó chuẩn hóa để phù hợp với cấu trúc dataset chung.

- **Dữ liệu thu thập:**
  thread_id, topic, content_type, likes, replies, reposts, timestamp

- **Vai trò trong nghiên cứu:**
  Threads đóng vai trò bổ trợ nhằm:

  - Phân tích **vòng đời xu hướng (trend lifecycle)**
  - Đánh giá **sentiment cộng đồng**
  - So sánh mức độ viral trên nền tảng nội dung và phản ứng xã hội

4. BÀI TOÁN NGHIÊN CỨU

4.1. Câu hỏi nghiên cứu tổng quát

Những yếu tố nào (thời điểm đăng tải, mức độ tương tác, tiêu đề, phản hồi tiêu cực) ảnh hưởng đáng kể đến khả năng một nội dung lọt vào danh sách Trending trên YouTube và Spotify?

4.2. Biến nghiên cứu

Biến độc lập:

Giờ đăng tải

Ngày trong tuần

Từ khóa tiêu đề

Like, Comment, Dislike

Tỷ lệ comment tiêu cực (sentiment analysis)

Biến phụ thuộc:

Trạng thái Trending (0/1)

Thứ hạng Trending

5. GIẢ THUYẾT NGHIÊN CỨU

Giả thuyết 1 – Thời điểm đăng tải

H_{01}: Thời điểm đăng tải không ảnh hưởng đến khả năng Trending

H_{11}: Nội dung đăng trong khung giờ 18h–22h có xác suất Trending cao hơn có ý nghĩa thống kê

Giả thuyết 2 – Ngôn ngữ tiêu đề

H_{02}: Từ khóa giật gân không ảnh hưởng đến lượt xem trung bình

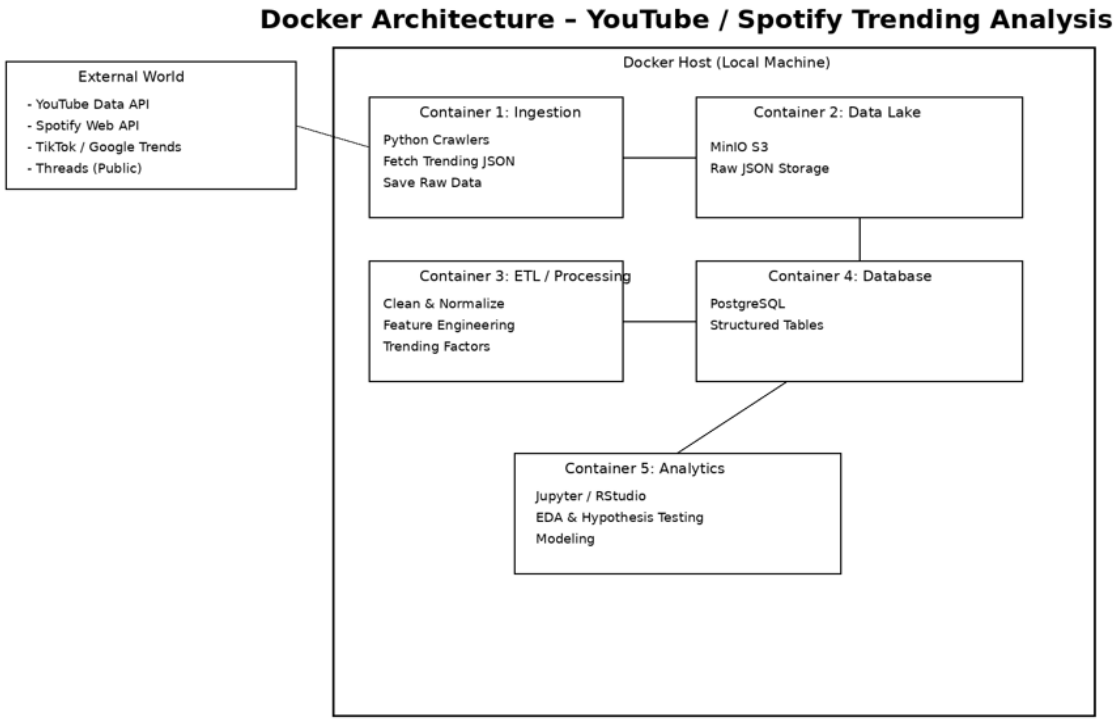
H_{12}: Tiêu đề chứa từ khóa kích thích cảm xúc có lượt xem cao hơn

Giả thuyết 3 – Tương tác tiêu cực

H_{03}: Tỷ lệ tương tác tiêu cực không ảnh hưởng đến khả năng viral

H_{13}: Tỷ lệ tương tác tiêu cực cao làm giảm khả năng xuất hiện trong Trending

6. KIẾN TRÚC HỆ THỐNG



KẾT LUẬN

Báo cáo đã trình bày chiến lược thiết kế dataset, bài toán nghiên cứu và kiến trúc hệ thống, tạo nền tảng cho các báo cáo tiếp theo trong học phần ADY201m.