|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**  **TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**  **KHOA CÔNG NGHỆ THÔNG TIN**    **Đồ án**  **Khoa học dữ liệu ứng dụng**  Dự án hướng ứng dụng  Ứng dụng tóm tắt podcast  **LỚP: CQ2021/21**  Giảng viên:  Lê Ngọc Thành  Lê Nhựt Nam  Võ Nam Thục Đoan  **Sinh viên thực hiện:**   |  |  | | --- | --- | | **MSSV** | **Tên thành viên** | | 21120348 | Nguyễn Trần Trình | | 21120441 | Dương Huỳnh Anh Duy | | 21120464 | Nguyễn Quốc Hưng |   Học kỳ II – Năm học 2024-2025 |

**Mục lục**

[I. Tổng quan dự án 3](#_Toc200543198)

[II. Giới thiệu 3](#_Toc200543199)

[III. Kiến trúc hệ thống 3](#_Toc200543200)

[1. Tầng thu thập và xử lý dữ liệu 3](#_Toc200543201)

[2. Tầng xử lý ngôn ngữ tự nhiên 3](#_Toc200543202)

[3. Tầng giao diện người dùng 4](#_Toc200543203)

[IV. Phương pháp luận và triển khai 4](#_Toc200543204)

[1. Thu thập và tiền xử lý dữ liệu 4](#_Toc200543205)

[2. Triển khai mô hình tóm tắt 5](#_Toc200543206)

[a) Lựa chọn và khởi tạo mô hình 5](#_Toc200543207)

[b) Xử lý văn bản đầu vào và chiến lược tóm tắt 6](#_Toc200543208)

[c) Thực thi tóm tắt và quản lý yêu cầu 6](#_Toc200543209)

[3. Giao diện người dùng 7](#_Toc200543210)

[V. Kết quả và đánh giá 7](#_Toc200543211)

[1. Chỉ số đánh giá 7](#_Toc200543212)

[a) Chuẩn bị đánh giá 7](#_Toc200543213)

[b) Similarity 7](#_Toc200543214)

[c) F1-Score 7](#_Toc200543215)

[d) MSE, BCE 8](#_Toc200543216)

[e) ROUGE 8](#_Toc200543217)

[f) BLEU 8](#_Toc200543218)

[2. Đánh giá chất lượng 8](#_Toc200543219)

[3. Demo ứng dụng 8](#_Toc200543220)

[VI. Thách thức và giải pháp 8](#_Toc200543221)

[1. Thách thức 8](#_Toc200543222)

[2. Giải pháp 8](#_Toc200543223)

[VII. Kết luận 8](#_Toc200543224)

[VIII. Hướng phát triển tương lai 8](#_Toc200543225)

[IX. Tài liệu tham khảo 8](#_Toc200543226)

# Tổng quan dự án

**Tên dự án**: Podcast Summarization App.

**Thời gian thực hiện**: 12 tuần.

**Thành viên nhóm**:

* 21120348 - Nguyễn Trần Trình.
* 21120441 - Dương Huỳnh Anh Duy (Nhóm trưởng).
* 21120464 - Nguyễn Quốc Hưng.

**Mục tiêu chính**: Phát triển ứng dụng sử dụng AI để tự động tóm tắt nội dung podcast, giúp người dùng nắm bắt thông tin chính nhanh chóng mà không cần nghe toàn bộ nội dung.

# Giới thiệu

Podcast đã trở thành phương tiện phổ biến để tiếp nhận thông tin và giải trí. Tuy nhiên, thời lượng dài của nhiều podcast gây khó khăn cho người dùng muốn nắm bắt nội dung chính một cách nhanh chóng. Dự án này giải quyết thách thức đó bằng cách phát triển một ứng dụng sử dụng kỹ thuật NLP và học máy để tạo bản tóm tắt tự động từ nội dung podcast.

# Kiến trúc hệ thống

Hệ thống được xây dựng với kiến trúc 3 tầng chính.

## Tầng thu thập và xử lý dữ liệu

Quy trình tự động hóa podcast bao gồm hai giai đoạn chính:

* **Tải xuống**: Tự động lấy tệp âm thanh từ RSS feed, kiểm tra và bỏ qua các tập đã có, sau đó lưu trữ chúng.
* **Chuyển đổi văn bản**: Sử dụng AI (Whisper) để chuyển đổi các tệp âm thanh thành văn bản, kiểm tra và bỏ qua các bản đã chuyển đổi, sau đó lưu lại dưới dạng text.

## Tầng xử lý ngôn ngữ tự nhiên

Triển khai nhiều phương pháp tóm tắt:

* **Mô hình local**:
  + BART-large.
  + Pegasus-dm.
  + T5-base.
* **Sử dụng API**:
  + Cohere.

## Tầng giao diện người dùng

# Phương pháp luận và triển khai

Trước tiên cần đảm bảo đã cài đặt theo yêu cầu của requirements.txt.



## Thu thập và tiền xử lý dữ liệu

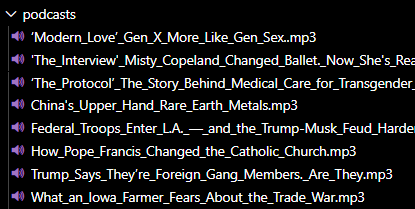


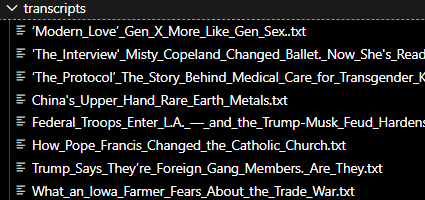
Quá trình tải podcast diễn ra tự động qua các bước chính sau:

1. **Phân tích RSS Feed**: Chương trình đọc RSS feed URL để lấy thông tin và danh sách các tập podcast gần nhất.
2. **Lặp qua từng tập**: Với mỗi tập podcast tìm thấy, chương trình trích xuất tiêu đề và đường dẫn tệp âm thanh (MP3). Tiêu đề được làm sạch để sử dụng làm tên tệp.
3. **Kiểm tra trùng lặp**: Trước khi tải, chương trình kiểm tra xem tệp MP3 của tập đó đã tồn tại trên máy tính chưa. Nếu có, tập đó sẽ bị bỏ qua để tránh tải lại.
4. **Tải và lưu tệp**: Nếu tệp chưa tồn tại, chương trình sẽ tải tệp MP3 từ đường dẫn đã lấy được. Tệp sau đó được lưu vào thư mục chỉ định với tên đã làm sạch.
5. **Xử lý lỗi**: Toàn bộ quá trình tải được giám sát để bắt và thông báo các lỗi có thể xảy ra (ví dụ: lỗi mạng, đường dẫn không hợp lệ), đảm bảo chương trình không bị dừng đột ngột.

Quá trình chuyển đổi âm thanh podcast thành văn bản được thực hiện như sau:

1. **Khởi tạo mô hình**: Tải và chuẩn bị mô hình Whisper (ví dụ: "base") để nhận diện giọng nói.
2. **Duyệt tệp MP3**: Quét qua thư mục chứa các tệp âm thanh podcast.
3. **Kiểm tra & Bỏ qua**: Nếu bản văn bản (transcript) của một tập đã tồn tại, chương trình sẽ bỏ qua tập đó.
4. **Chuyển đổi & Lưu**: Sử dụng mô hình Whisper để chuyển đổi âm thanh thành văn bản, sau đó lưu kết quả vào một tệp .txt tương ứng.
5. **Xử lý lỗi**: Bắt và thông báo các lỗi có thể xảy ra trong quá trình transcribe.





## Triển khai mô hình tóm tắt



Trọng tâm của ứng dụng, nơi các mô hình học máy được sử dụng để chuyển đổi nội dung văn bản dài (transcript của podcast) thành các bản tóm tắt ngắn gọn. Quá trình này được thiết kế để hỗ trợ cả mô hình chạy cục bộ và các dịch vụ tóm tắt dựa trên API, đồng thời quản lý hiệu quả tài nguyên và lỗi.

### Lựa chọn và khởi tạo mô hình

Ứng dụng hỗ trợ nhiều loại mô hình tóm tắt khác nhau, được định nghĩa trong ‘*main.py*’.

**Mô hình cục bộ (Local Models)**:

* ‘*local\_summarizer.py*’ chịu trách nhiệm tải các mô hình tóm tắt của **Hugging Face** (BART, Pegasus, T5) từ **AutoTokenizer** và **AutoModelForSeq2SeqLM**.
* Khi khởi tạo, LocalSummarizer sẽ tải tokenizer và mô hình tương ứng dựa trên model\_key được chọn (‘*bart-large*’, ‘*pegasus-dm*’, ‘*t5-base*’).
* Mô hình sẽ được chuyển sang thiết bị **cuda** (GPU) nếu có, nếu không sẽ dùng cpu.

**Mô hình dựa trên API Web (Web API Model - Cohere):**

* ‘*web\_summarizer.py*’ tích hợp với **Cohere API**.
* Một đối tượng cohere.Client được khởi tạo bằng COHERE\_API\_KEY để thiết lập kết nối với dịch vụ Cohere.
* Cohere sử dụng mô hình "command" để thực hiện tóm tắt.

### Xử lý văn bản đầu vào và chiến lược tóm tắt

Để xử lý các văn bản dài (transcript của podcast), ứng dụng áp dụng chiến lược chia nhỏ và tóm tắt theo từng phần.

**Chia nhỏ văn bản (Chunking)**:

* Trong ‘*LocalSummarizer.py*’ văn bản được mã hóa thành tokens và sau đó chia thành các đoạn nhỏ hơn, mỗi đoạn không vượt quá max\_input\_length của mô hình (1024 cho BART và Pegasus, 512 cho T5).
* Đối với Cohere API văn bản được chia thành các đoạn dựa trên số lượng ký tự (MAX\_CHUNK\_CHARS, mặc định 3000 ký tự) để tối ưu số lượng yêu cầu API và quản lý giới hạn. Các đoạn quá nhỏ có thể được gộp vào đoạn trước đó.

**Tóm tắt đa giai đoạn (Multi-pass summarization - Local Summarizer)**:

* Nếu văn bản gốc quá dài và phải chia thành nhiều chunks, ‘*LocalSummarizer.py*’ sẽ thực hiện tóm tắt hai lần.
* Giai đoạn đầu (first\_pass): Tóm tắt từng chunk riêng lẻ.
* Giai đoạn cuối (nếu cần): Các bản tóm tắt từ first\_pass được gộp lại, sau đó lại được chia thành chunks mới và tóm tắt lần nữa để tạo ra bản tóm tắt cuối cùng.

### Thực thi tóm tắt và quản lý yêu cầu

Hàm **summarize\_with\_model** trong ‘*main.py*’ là điểm điều phối chính cho việc gọi các mô hình tóm tắt.

**Tóm tắt cục bộ**:

* Hàm **summarize\_hf** trong ‘*local\_summarizer.py*’ được gọi, truyền vào văn bản, khóa mô hình, độ dài đầu vào tối đa và độ dài tóm tắt mong muốn.
* Mô hình generate được sử dụng để tạo tóm tắt với các tham số như max\_length, min\_length, num\_beams và early\_stopping.

**Tóm tắt qua API (Cohere)**:

* Hàm **summarize\_with\_cohere** trong ‘*web\_summarizer.py*’ được gọi.
* Các yêu cầu co.summarize được gửi cho từng text\_chunk với các tham số như model, length, format, temperature và additional\_command.
* Xử lý giới hạn tốc độ và thử lại: summarize\_chunk\_with\_cohere bao gồm logic thử lại khi gặp lỗi 429 (quá giới hạn tốc độ) và sử dụng time.sleep() để chờ giữa các yêu cầu, đảm bảo tuân thủ giới hạn API (5 calls/min).

**Ghi nhận kết quả**:

* Đối với mỗi lần tóm tắt, thời gian xử lý được ghi lại. Kết quả bao gồm ID của mô hình, thời gian thực hiện, bản tóm tắt và trạng thái (success hoặc failed).
* Các lỗi phát sinh trong quá trình tóm tắt (cục bộ hoặc API) đều được bắt và thông báo trong trường summary với trạng thái failed.

## Giao diện người dùng

# Kết quả và đánh giá

## Chỉ số đánh giá

### Chuẩn bị đánh giá

Cho dữ liệu chạy qua tất cả mô hình cho ra kết quả cuối cùng sau đó dùng kết quả này để đánh giá từng mô hình.

### Similarity

Đánh giá độ tương đồng của văn bản sau tóm tắt so với nguyên bản, cho thấy phần nội dung có bị thay đổi hay không còn nội dung chính như nguyên bản.

Các mô hình sẽ sử dụng Similarity để dùng làm đơn vị đo độ chính xác của văn bản tóm tắt.

### F1-Score

Dùng định nghĩa F1-score cho vào đây và nó dùng Similarity như tham số.

### MSE, BCE

Cho định nghĩa và nó dùng Similarity.

### ROUGE

Thằng này cũng dùng Similarity nhưng chuyên dùng để đo so sánh text.

### BLEU

Thằng này cũng Similarity và chuyên dùng để so sánh text.

## Đánh giá chất lượng

## Demo ứng dụng

# Thách thức và giải pháp

|  |  |
| --- | --- |
| **Thách thức** | **Giải pháp** |
| Xử lý văn bản đầu vào quá dài. | Chia văn bản gốc thành các đoạn nhỏ hơn (chunks) trước khi đưa vào mô hình tóm tắt.  Đối với các mô hình cục bộ, có thể thực hiện tóm tắt nhiều lần (multi-pass summarization) nếu có nhiều đoạn. |
| Giới hạn đầu vào của mô hình tóm tắt. | Các mô hình cục bộ được cấu hình với max\_input khác nhau.  Chức năng chunk\_text đảm bảo các đoạn văn bản phù hợp với giới hạn này.  Đối với Cohere API, có MAX\_CHUNK\_CHARS để chia nhỏ văn bản trước khi gửi yêu cầu. |
| Tải lại hoặc xử lý lại các tệp đã có. | Khi tải podcast: Kiểm tra os.path.exists(mp3\_path) để bỏ qua các tập đã tồn tại.  Khi chuyển đổi văn bản: Kiểm tra os.path.exists(text\_path) để bỏ qua các bản ghi (transcript) đã có. |
| Lỗi trong quá trình tải hoặc chuyển đổi/tóm tắt. | Sử dụng khối try...except để bắt và xử lý các ngoại lệ (Exception) có thể xảy ra trong quá trình tải podcast, chuyển đổi âm thanh thành văn bản và tóm tắt văn bản. Thông báo lỗi cụ thể được in ra để hỗ trợ gỡ lỗi. |
| Giới hạn tốc độ (Rate Limiting) của API bên ngoài (Cohere). | Triển khai logic thử lại (retry) và tạm dừng (time.sleep) giữa các yêu cầu API Cohere (RATE\_LIMIT\_SECONDS) khi nhận mã lỗi 429 (quá giới hạn). |
| Quản lý thời gian xử lý cho mỗi tác vụ tóm tắt. | Áp dụng giới hạn thời gian (TIMEOUT) cho mỗi model khi tóm tắt. Nếu quá thời gian, một TimeoutError sẽ được kích hoạt. |
| Khả năng tương thích và lựa chọn mô hình. | Định nghĩa rõ ràng các SUPPORTED\_MODELS (cục bộ và web) với đường dẫn và cấu hình tương ứng. Cung cấp hàm get\_available\_models và validate\_models để đảm bảo người dùng chọn đúng mô hình được hỗ trợ. |
| Lưu trữ và tổ chức kết quả. | Các thư mục TRANSCRIPT\_DIR và OUTPUT\_DIR được tạo tự động nếu chưa tồn tại. Kết quả tóm tắt được lưu vào một tệp JSON (summaries.json) với cấu trúc rõ ràng, bao gồm tên tệp gốc, thông tin mô hình, thời gian và bản tóm tắt/trạng thái. |

# Kết luận

# Hướng phát triển tương lai

## Hỗ trợ đa ngôn ngữ và mô hình nâng cao

* **Thách thức hiện tại**: Mô hình Whisper "base" có thể hoạt động tốt cho nhiều ngôn ngữ nhưng các mô hình lớn hơn hoặc chuyên biệt hơn có thể cho độ chính xác cao hơn. Quá trình tóm tắt hiện đang yêu cầu rõ ràng "Tóm tắt bằng tiếng Việt" cho Cohere và các mô hình Hugging Face có thể cần tinh chỉnh hoặc các mô hình đã được huấn luyện cho ngôn ngữ cụ thể.
* **Hướng phát triển**:
  + Tích hợp các mô hình Whisper lớn hơn (ví dụ: "large", "large-v2") hoặc các mô hình chuyển đổi giọng nói chuyên biệt cho từng ngôn ngữ để cải thiện độ chính xác.
  + Tự động phát hiện ngôn ngữ của podcast và chọn mô hình chuyển đổi giọng nói và tóm tắt phù hợp.
  + Hỗ trợ nhiều API tóm tắt khác ngoài Cohere (ví dụ: OpenAI GPT, Google Gemini, Anthropic Claude) để cung cấp sự lựa chọn đa dạng và khả năng xử lý tốt hơn cho các ngôn ngữ khác.

## Cải thiện chất lượng tóm tắt

* **Thách thức hiện tại**: Tóm tắt nhiều đoạn (multi-pass summarization) có thể chưa tối ưu. Chất lượng tóm tắt có thể thay đổi tùy thuộc vào mô hình và độ dài của văn bản gốc.
* **Hướng phát triển**:
  + Triển khai các kỹ thuật tóm tắt phân cấp (hierarchical summarization) hoặc tóm tắt theo cụm (cluster-based summarization) để xử lý các podcast rất dài một cách hiệu quả hơn.
  + Thêm tùy chọn cho người dùng để điều chỉnh các tham số tóm tắt chi tiết hơn (ví dụ: mức độ trừu tượng, độ dài câu, từ khóa quan trọng).
  + Kết hợp các kỹ thuật trích xuất thông tin (Information Extraction) để nhận diện các thực thể, chủ đề chính, và điểm thời gian quan trọng trong podcast, giúp tóm tắt tập trung hơn.

## Tối ưu hóa hiệu suất và chi phí

* **Thách thức hiện tại**: Các mô hình cục bộ có thể yêu cầu tài nguyên lớn (RAM, VRAM). API bên ngoài có chi phí sử dụng và giới hạn tốc độ.
* **Hướng phát triển**:
  + Triển khai phân phối tải (load balancing) hoặc sử dụng các worker khác nhau để xử lý các tác vụ tải, chuyển đổi và tóm tắt song song.
  + Tối ưu hóa việc sử dụng GPU (ví dụ: sử dụng fp16 cho mô hình Whisper nếu phần cứng hỗ trợ).
  + Cơ chế quản lý cache thông minh hơn cho các mô hình và tokenizers để giảm thời gian khởi tạo.
  + Tích hợp các công cụ theo dõi chi phí khi sử dụng API bên ngoài.

## Tích hợp và tự động hóa cao hơn

* **Thách thức hiện tại**: Quy trình hiện tại là thủ công một phần (cần chạy script).
* **Hướng phát triển**:
  + Thiết lập lịch trình tự động (cron job hoặc dịch vụ nền) để định kỳ kiểm tra RSS feed, tải podcast mới và xử lý chúng.
  + Tích hợp với các nền tảng lưu trữ đám mây (Google Drive, Dropbox) để tự động tải lên tệp âm thanh, transcript và tóm tắt.
  + Khả năng xuất bản tóm tắt trực tiếp lên các nền tảng blog hoặc mạng xã hội.

# Tài liệu tham khảo

Lewis, M. et al. (2019). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation".

Raffel, C. et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer".

Radford, A. et al. (2022). "Robust Speech Recognition via Large-Scale Weak Supervision".

Tài liệu API Cohere: <https://docs.cohere.com>

Repository GitHub: <https://github.com/NguyenQuocHung47/ADS_Project>