GDGOC Hackathon Vietnam 2025

Hồ Sơ Dự Án

Tên Đội: WinToWin

Thành viên	Vai trò
Phạm Nguyên Hải Long	Project Manager
Võ Thành Nghĩa	Technical Lead
Dương Quang Thắng	Technical Lead
Hà Tuấn Anh	Member
Nguyễn Quốc Thắng	Member

1. Thông Tin Chung

1.1 Tên Dự Án

- Dự án mang tên **NoteUS**.

1.2 Mô Tả Ngắn Gọn

NoteUS là một nền tảng AI giúp người dùng khai thác tri thức từ tài liệu và video một cách hiệu quả. Hệ thống không chỉ hỗ trợ phân tích, trích xuất thông tin mà còn đảm bảo tính minh bạch và chính xác trong từng câu trả lời.

Tính năng nổi bật:

- **Tạo mindmap & cheatsheet** giúp tóm tắt nội dung tài liệu trực quan.
- **Tóm tắt & trích xuất thông tin từ video YouTube**, tiết kiệm thời gian nắm bắt nội dung.
- **Bảo mật dữ liệu** với cơ chế xác thực và lọc nội dung nhạy cảm.
- **Hỗ trợ API OpenAI** cùng khả năng trích dẫn tài liệu nguồn, đảm bảo minh bạch.
- **Hỏi đáp trên video do người dùng cung cấp**, mang lại sự linh hoạt trong tiếp cận thông tin.

1.3 NoteUS và “Responsible AI”

- **Bảo vệ tính công bằng:** Tất cả các đối tượng khách hàng đều có thể sử dụng sản phẩm. AI điều chỉnh phản hồi nếu được yêu cầu xử lý lên tập dữ liệu mang ý nghĩa nội dung phân biệt.
- **Minh bạch:** Mọi câu trả lời đều đi kèm trích dẫn nguồn rõ ràng.
- **Bảo mật:**
    - Cơ chế xóa file tự động và lọc dữ liệu nhạy cảm giúp bảo vệ quyền riêng tư của người dùng.
    - Hệ thống tiền xử lý dữ liệu nhạy cảm trước khi gửi đi bên phía AI xử lý và tổng hợp thông tin.
- **An toàn cho người dùng và môi trường:**
    - Hệ thống cho người dùng đánh giá phản hồi từ AI và chọn phản hồi đáng tin cậy.
    - Hệ thống tính hợp tính năng caching để giảm chi phí gọi API lên AI nhằm giảm khả năng gây tác động lên môi trường từ kiến trúc hạ tầng của AI.

2. Vấn Đề và Giải Pháp

2.1 Đặt Vấn Đề

Hiện nay, người dùng gặp nhiều thách thức trong việc tìm kiếm, tổng hợp và xác thực thông tin từ nhiều nguồn khác nhau, đặc biệt là khi làm việc với tài liệu học thuật, báo cáo nghiên cứu hoặc các nội dung chuyên môn. Một số vấn đề phổ biến bao gồm:

1. **Khó khăn trong tổng hợp thông tin:**
    - Khi phải làm việc với nhiều tài liệu, người dùng thường mất nhiều thời gian đọc, chọn lọc và liên kết nội dung quan trọng.
    - Các tài liệu có thể có cách diễn đạt khác nhau, gây khó khăn trong việc thống nhất thông tin.
2. **Nguy cơ sai lệch và thiếu minh bạch trong dữ liệu:**
    - Khi tham khảo nhiều nguồn, người dùng có thể gặp khó khăn trong việc xác minh độ chính xác của thông tin.
    - Việc không có trích dẫn rõ ràng hoặc sử dụng dữ liệu từ các nguồn không đáng tin cậy có thể dẫn đến sai lệch thông tin.
3. **Lo ngại về bảo mật và rò rỉ dữ liệu:**
    - Một số tài liệu có chứa thông tin nhạy cảm (dữ liệu cá nhân, tài khoản, thông tin tài chính), nếu không được bảo vệ có thể bị lộ ra ngoài.
    - Việc lưu trữ tài liệu trên các nền tảng trực tuyến không có cơ chế bảo mật có thể làm tăng nguy cơ rò rỉ dữ liệu.
4. **Khó khăn trong xử lý thông tin từ video:**

- Nội dung trong video thường dài và không dễ dàng tóm tắt nhanh chóng.
- Người dùng cần một công cụ hỗ trợ trích xuất thông tin quan trọng từ video mà không phải xem toàn bộ.

NoteUS ra đời nhằm giải quyết những vấn đề này bằng cách cung cấp một hệ thống hỏi đáp AI thông minh, giúp người dùng trích xuất, tổng hợp và xác thực thông tin từ tài liệu và video một cách nhanh chóng, chính xác và bảo mật.

2.2 Giải Pháp

1. Hệ thống tạo Mindmap & Cheat Sheet tự động từ tài liệu đầu vào

Hệ thống sẽ giúp bạn chuyển đổi tài liệu thô thành **mindmap** (bản đồ tư duy) và **cheatsheet** (bảng ghi nhớ) một cách tự động. Điều này giúp bạn dễ dàng phân tích, tổng hợp thông tin từ nhiều nguồn khác nhau mà không cần mất hàng giờ đồng hồ để sắp xếp nội dung.

Quy trình hoạt động của hệ thống:

1. **Phân tách nội dung**
   - Hệ thống sẽ chia nhỏ nội dung của từng tài liệu thành nhiều đoạn nhỏ (chunks) để xử lý hiệu quả hơn.
2. **Tổng hợp thông tin từ nhiều tài liệu**
   - Các đoạn nội dung từ nhiều tài liệu khác nhau sẽ được gộp chung thành một tập dữ liệu tổng hợp, không quan tâm đến thứ tự gốc.
3. **Chuyển đổi nội dung thành vector**
   - Mỗi đoạn nội dung sẽ được chuyển đổi thành vector số hóa bằng phương pháp embedding để phục vụ cho quá trình phân nhóm.
4. **Người dùng tùy chỉnh số nhóm**
   - Bạn có thể chọn số lượng nhóm mong muốn. Số nhóm này sẽ tương đương với:
     - Số nhánh của mindmap
     - Số tiêu đề chính của cheatsheet
5. **Gom nhóm nội dung theo chủ đề**
   - Hệ thống sẽ sử dụng thuật toán K-Means Clustering để tự động phân loại các đoạn nội dung có ý nghĩa tương đồng vào cùng một nhóm.
6. **Tối ưu hóa nội dung bằng AI**
   - Cuối cùng, mô hình GPT sẽ giúp định dạng và trình bày nội dung của từng nhóm một cách logic, rõ ràng, dễ hiểu để tạo ra mindmap hoặc cheatsheet hoàn chỉnh.

Kết quả nhận được

- **Mindmap:** Giúp bạn hình dung tổng quan về nội dung một cách trực quan, dễ nhớ.
- **Cheatsheet:** Cung cấp bản tóm tắt súc tích, giúp bạn nhanh chóng nắm bắt ý chính mà không cần đọc lại toàn bộ tài liệu.

2. Hệ thống Hỏi-Đáp Thông Minh từ Video YouTube

Hệ thống này sẽ giúp bạn tóm tắt & trả lời câu hỏi dựa trên nội dung của video, giúp bạn nắm bắt thông tin nhanh chóng và chính xác.

Cách hệ thống hoạt động

Bước 1: Người dùng nhập link YouTube

- Bạn chỉ cần cung cấp đường dẫn của bất kỳ video YouTube nào mà bạn muốn tìm hiểu.

Bước 2: Trích xuất nội dung video

- Nếu video có transcript (phụ đề), hệ thống sẽ tự động lấy nội dung từ đó.
- Nếu video không có transcript, hệ thống sẽ sử dụng mô hình nhận dạng giọng nói (speech-to-text) để chuyển đổi âm thanh trong video thành văn bản.

Bước 3: Xử lý & tối ưu nội dung

- Văn bản thu được sẽ được làm sạch, loại bỏ các phần dư thừa (như từ đệm, ngập ngừng...) để đảm bảo nội dung súc tích và dễ đọc.

Bước 4: Hỏi - Trả lời

- Bạn có thể đặt câu hỏi về nội dung video, và hệ thống sẽ sử dụng AI (LLM GPT) để tìm ra câu trả lời chính xác dựa trên nội dung đã trích xuất.
- Ngoài ra, bạn cũng có thể yêu cầu hệ thống tóm tắt toàn bộ video để có cái nhìn tổng quan nhanh nhất.

Kết quả nhận được

- **Tiết kiệm thời gian** – Không cần xem toàn bộ video, chỉ lấy thông tin quan trọng nhất
- **Linh hoạt & thông minh** – Hỗ trợ cả video có sẵn transcript lẫn video không có phụ đề.

- **Trợ lý thông tin hiệu quả** – Dễ dàng tra cứu kiến thức từ bất kỳ video nào trên YouTube.

4. Hệ thống Hỏi-Đáp Thông Qua API với OpenAI Dựa trên Tài Liệu Người Dùng

Hệ thống này cho phép bạn **tương tác và hỏi đáp thông minh** thông qua API của OpenAI, giúp bạn tìm kiếm thông tin trong tài liệu đã cung cấp một cách nhanh chóng và chính xác. Hệ thống sẽ sử dụng các kỹ thuật tiên tiến để xử lý và trích xuất thông tin từ tài liệu, đưa ra câu trả lời có căn cứ rõ ràng

Bước 1: Lưu trữ context từ tài liệu người dùng

- Người dùng upload các tài liệu lên hệ thống. Hệ thống sẽ trích xuất nội dung từ các file và chuyển đổi thành dữ liệu có thể đọc được.

Bước 2: Biểu diễn tài liệu dưới dạng vector embedding

- Hệ thống chuyển đổi nội dung tài liệu thành các vector embedding để có thể dễ dàng tìm kiếm và so sánh mức độ liên quan.

Bước 3: Người dùng nhập câu truy vấn

- Người dùng nhập câu hỏi vào hệ thống để tìm kiếm thông tin trong tài liệu đã cung cấp.

Bước 4: Xử lý truy vấn và tìm context liên quan

- Hệ thống chuyển câu truy vấn của người dùng thành vector embedding, sau đó so sánh với các vector embedding của tài liệu để tìm kiếm những phần có liên quan nhất.

Bước 5: Gửi dữ liệu đến OpenAI

- Hệ thống gửi phần context liên quan cùng câu truy vấn đến OpenAI để nhận phản hồi.

Bước 6: Hiển thị kết quả cho người dùng

- Hệ thống hiển thị phản hồi từ OpenAI kèm theo phần nội dung tài liệu đã được sử dụng để tạo ra phản hồi.

Kết quả nhận được

- **Truy xuất thông tin chính xác** – Hệ thống giúp người dùng nhanh chóng tìm thấy thông tin cần thiết từ tài liệu đã cung cấp.

- **Hỗ trợ tìm kiếm ngữ nghĩa** – Nhờ vào embedding, hệ thống có thể tìm kiếm thông tin theo ý nghĩa thay vì chỉ dựa trên từ khóa đơn thuần.
- **Cung cấp nguồn trích dẫn** – Đảm bảo rằng phản hồi từ AI có căn cứ rõ ràng từ tài liệu gốc, giúp tăng độ tin cậy của câu trả lời.

5. Hệ thống Trích Dẫn Từ Tài Liệu Để Trả Lời Câu Hỏi

Hệ thống này không chỉ trả lời câu hỏi mà còn **trích dẫn nguồn gốc** của thông tin từ tài liệu bạn cung cấp, giúp bạn dễ dàng kiểm tra lại nguồn tham khảo một cách chính xác.

Cách hệ thống hoạt động

Bước 1: Người dùng cung cấp tài liệu

- Bạn upload tài liệu dưới dạng văn bản hoặc các file có nội dung mà bạn muốn trích dẫn.

Bước 2: Chuyển đổi tài liệu thành vector embedding

- Hệ thống sẽ chuyển đổi toàn bộ nội dung của tài liệu thành vector embedding. Đây là quá trình mã hóa văn bản thành các vector số, giúp máy tính dễ dàng so sánh và phân tích nội dung.

Bước 3: Phân tích câu hỏi và tìm kiếm context phù hợp

- Khi bạn đặt câu hỏi, hệ thống sẽ tạo ra vector câu trả lời và tính toán độ tương đồng (cosine similarity) giữa vector câu trả lời và các vector embedding trong tài liệu.

Bước 4: Trích dẫn phần nội dung phù hợp

- Các vector embedding trong tài liệu có độ tương đồng cao nhất với câu trả lời sẽ được chọn làm nguồn trích dẫn. Điều này đảm bảo rằng câu trả lời không chỉ chính xác mà còn có cơ sở rõ ràng từ tài liệu gốc.

Bước 5: Trả lời có trích dẫn rõ ràng

- Hệ thống sẽ kết hợp GPT để tạo ra câu trả lời mạch lạc và đính kèm trích dẫn từ tài liệu, giúp người dùng xác minh nguồn thông tin dễ dàng.

Kết quả nhận được

- **Trả lời chính xác, có nguồn gốc rõ ràng** – Đảm bảo rằng bạn luôn nhận được câu trả lời từ các tài liệu đáng tin cậy.
- **Tự động trích dẫn** – Không cần phải tìm kiếm thủ công, hệ thống sẽ tự động trích dẫn phần nội dung liên quan nhất.
- **Tiết kiệm thời gian** – Hệ thống giúp bạn tìm kiếm thông tin nhanh chóng và hiệu quả từ tài liệu đã cung cấp.

6. Hệ thống Hỏi-Đáp Trên Video Cá Nhân với Fine-Tuning

Với hệ thống này, bạn có thể **tương tác với video cá nhân** của mình theo cách hoàn toàn mới. Bằng cách upload video, hệ thống sẽ giúp bạn **trích xuất và phân tích nội dung âm thanh** từ video, và trả lời các câu hỏi của bạn về nội dung video một cách chính xác, đặc biệt hỗ trợ giọng nói tiếng Việt với nhiều đặc trưng vùng miền.

Cách hệ thống hoạt động

Bước 1: Người dùng upload video

- Bạn chỉ cần tải lên bất kỳ video nào bạn muốn làm việc. Không cần phải lo lắng về định dạng, hệ thống sẽ xử lý tất cả.

Bước 2: Trích xuất dữ liệu âm thanh từ video

- Hệ thống sẽ tách phần âm thanh của video và chuẩn bị dữ liệu cho các bước xử lý tiếp theo.

Bước 3: Chuyển giọng nói thành văn bản

- Sử dụng mô hình speech-to-text (chuyển giọng nói thành văn bản), hệ thống sẽ nhận dạng và chuyển đổi nội dung âm thanh thành văn bản.
- Điểm đặc biệt: Hệ thống có khả năng nhận diện giọng nói tiếng Việt, bao gồm các đặc trưng vùng miền Bắc, Trung, Nam và cả giọng địa phương, điều mà nhiều mô hình trước đây gặp khó khăn.

Bước 4: Tạo mô hình hỏi-đáp fine-tuned

- Sau khi chuyển đổi thành văn bản, hệ thống sẽ sử dụng mô hình GPT đã được fine-tuned để hiểu ngữ cảnh của video và trả lời câu hỏi của bạn.
- Bạn có thể hỏi bất kỳ câu hỏi nào về nội dung video và hệ thống sẽ trả lời dựa trên thông tin đã phân tích.

Bước 5: Trả lời câu hỏi chính xác

- Hệ thống sẽ đưa ra câu trả lời dựa trên nội dung video, kèm theo các phân tích ngữ nghĩa sâu sắc, giúp bạn hiểu rõ hơn về video mà bạn đã upload.

Kết quả nhận được

- **Tương tác trực quan với video cá nhân** – Giúp bạn dễ dàng hỏi đáp và khám phá nội dung video của mình.
- **Chuyển giọng nói tiếng Việt chuẩn xác** – Hỗ trợ giọng nói từ nhiều vùng miền (Bắc, Trung, Nam) và giọng địa phương, vượt qua khó khăn mà các mô hình trước gặp phải.
- **Mô hình fine-tuned** – Hệ thống được tối ưu hóa để hiểu tốt hơn ngữ cảnh và trả lời câu hỏi chính xác hơn về video của bạn.

7. Hệ thống Bảo Mật Thông Tin Nhạy Cảm từ Tài Liệu Người Dùng

Với mục tiêu bảo vệ dữ liệu cá nhân và thông tin nhạy cảm, hệ thống sẽ giúp bạn xử lý và lưu trữ tài liệu một cách an toàn, đồng thời đảm bảo rằng mọi thông tin nhạy cảm sẽ được mã hóa trước khi lưu vào cơ sở dữ liệu.

Cách hệ thống bảo mật hoạt động

Bước 1: Trích xuất dữ liệu từ tài liệu người dùng

- Người dùng upload tài liệu và hệ thống sẽ trích xuất nội dung từ các file mà bạn cung cấp. Dữ liệu này sẽ được chuyển thành dạng mà phần mềm có thể đọc và lưu trữ.

Bước 2: Áp dụng mô hình Nhận Dạng Thực Thể (NER)

- Hệ thống sẽ sử dụng mô hình Named Entity Recognition (NER) để nhận dạng các thực thể nhạy cảm trong tài liệu, như tên riêng, địa danh, tổ chức, v.v.
- Các thông tin nhạy cảm này sẽ được mã hóa lại thành các ký tự chung chung, ví dụ: "person_1", "place_2", "organization_3" để bảo vệ danh tính và thông tin cá nhân.

Bước 3: Mã hóa và lưu trữ an toàn

- Sau khi các thông tin nhạy cảm đã được mã hóa, dữ liệu sẽ được lưu trữ an toàn trong cơ sở dữ liệu mà không làm lộ bất kỳ thông tin nhạy cảm nào.

- Đồng thời, file gốc của người dùng cũng sẽ được lưu trữ, nhưng đã được xử lý sao cho không có bất kỳ thông tin cá nhân nào bị lộ.

Bước 4: Quy trình bảo mật toàn diện

- Hệ thống sẽ đảm bảo rằng chỉ các thông tin không nhạy cảm được sử dụng trong quá trình phân tích và xử lý, đồng thời bảo vệ tuyệt đối tính riêng tư của người dùng.

Kết quả nhận được

- **Bảo vệ thông tin nhạy cảm** – Hệ thống đảm bảo rằng mọi thông tin cá nhân, địa chỉ, tổ chức đều được mã hóa và bảo vệ trước khi lưu trữ.
- **Bảo mật toàn diện** – Chỉ các dữ liệu không nhạy cảm được sử dụng cho mục đích xử lý, đảm bảo sự riêng tư tuyệt đối cho người dùng.
- **Mã hóa thông minh** – Mô hình NER giúp nhận diện và thay thế thông tin nhạy cảm chính xác, mang lại sự bảo mật cao cho tài liệu của người dùng.

3. Công Nghệ và Nền Tảng

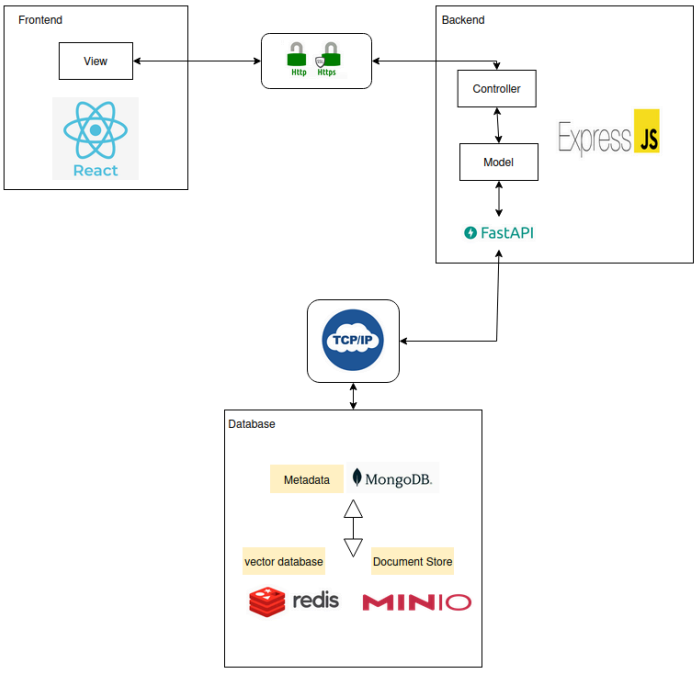
3.1 Công Nghệ

Thành phần	Ngôn ngữ	Framework
Upload & Preprocessing	Python	Fast API
Vector Database & Indexing	Python	Fast API
Question-Answering (RAG)	Python	Fast API
Mindmap & Cheatsheet	Python	Fast API
API Gateway	Node.js	Express.js
Authentication & User Management	Node.js	Express.js

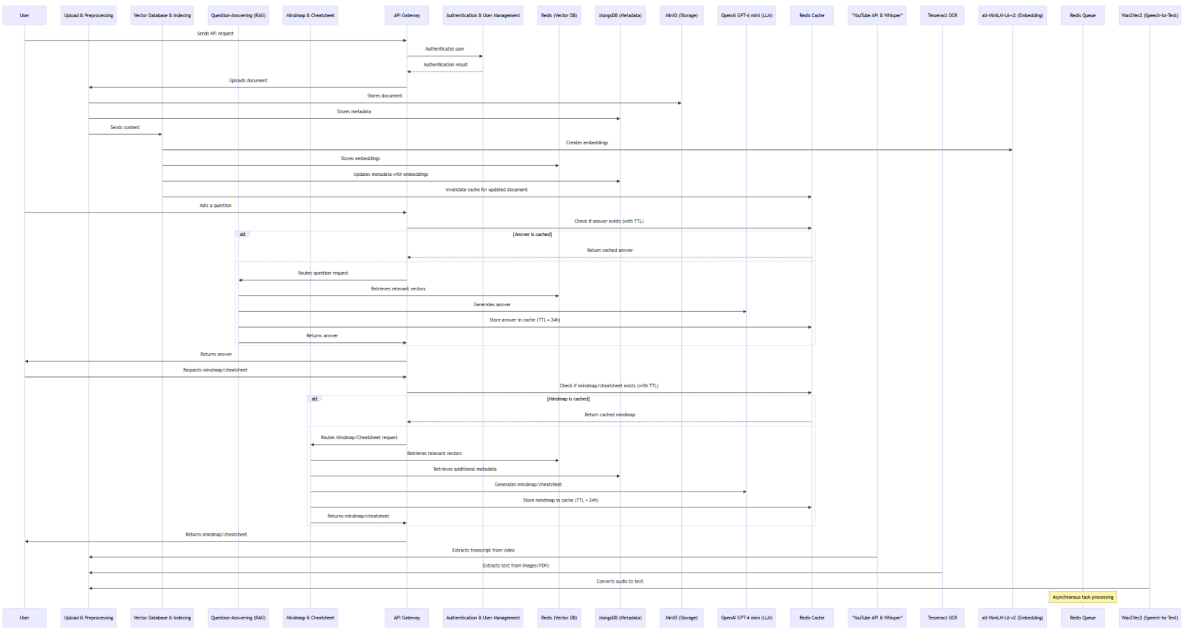
Thành phần	Công cụ & Công nghệ

Vector Database	Redis
Database (Metadata)	MongoDB
Storage	MinIO
LLM API	OpenAI GPT-4 mini
YouTube Transcript	YouTube API + Whisper
Containerization & Orchestration	Docker, Ngrok
OCR	Tesseract OCR
Model Embedding	all-MiniLM-L6-v2
Queue & Message Broker	Redis Queue
Speech-to-Text Model	Wav2Vec2

3.3 Pipeline hoạt động của hệ thống



Cách các thành phần trong hệ thống tương tác với nhau



Sequence Diagram thể hiện quy trình thao tác các chức năng của người dùng trong hệ thống Sequence Diagram-1.png

Luồng xử lý dữ liệu

Bước 1: Người dùng tải lên tài liệu

User gửi request tải tài liệu lên API Gateway.
API Gateway xác thực người dùng bằng cách kiểm tra với **Authentication & User Management**.
Upload & Preprocessing xử lý tài liệu, bao gồm:

- Lưu tài liệu vào **MinIO (Storage)**.
- Trích xuất metadata và lưu vào **MongoDB (Metadata)**.
- Gửi nội dung đến **Vector Database & Indexing** để tạo embedding.
  - **Vector Database & Indexing thực hiện số hóa tài liệu**, gồm:
- **Tạo embedding với MiniLM-L6-v2** (mô hình AI để mã hóa văn bản thành vector).
- **Lưu embedding vào Redis (Vector DB)** để truy vấn nhanh.
- **Cập nhật lại metadata trong MongoDB** để ghi nhận thông tin mới.
  - **Cache Invalidation:**
- Nếu tài liệu đã tồn tại trước đó, **cache cũ sẽ bị xóa** để tránh lỗi dữ liệu cũ.

Mục tiêu: Chuẩn bị dữ liệu cho các truy vấn AI, tối ưu hóa lưu trữ.

Bước 2: Người dùng đặt câu hỏi về tài liệu

User gửi câu hỏi lên API Gateway.
API Gateway kiểm tra cache trong **Redis Cache**:

- Nếu câu hỏi đã từng được hỏi trước đó **và cache chưa hết hạn**, hệ thống **lấy ngay kết quả từ cache**, giúp tăng tốc độ phản hồi.
- Nếu không có trong cache, tiếp tục bước 3.
  - **API Gateway gửi truy vấn đến Question-Answering (RAG).**
  - **Question-Answering (RAG) thực hiện các bước sau:**
- **Truy xuất dữ liệu từ Redis (Vector DB)** để lấy các đoạn văn bản liên quan.
- **Gửi truy vấn và dữ liệu đến OpenAI GPT-4 mini** để sinh câu trả lời.
  - **Kết quả từ GPT-4 được lưu vào cache** với TTL = **24 giờ** để giảm tải trong các truy vấn sau.
  - **API Gateway trả kết quả về cho người dùng.**

Mục tiêu: Tăng tốc xử lý câu hỏi với caching, giảm số lần gọi OpenAI GPT-4 để tiết kiệm chi phí.

Bước 3: Người dùng yêu cầu tạo Mindmap hoặc Cheatsheet

User gửi yêu cầu tạo Mindmap/Cheatsheet lên API Gateway.
API Gateway kiểm tra cache trong **Redis Cache**:

- Nếu đã có kết quả trong cache **và TTL chưa hết hạn**, trả về ngay.
- Nếu không có trong cache, tiếp tục bước 3.
  - **API Gateway gửi yêu cầu đến Mindmap & Cheatsheet.**
  - **Mindmap & Cheatsheet thực hiện các bước sau:**
- **Truy xuất dữ liệu từ Redis (Vector DB)** để lấy nội dung cần thiết.
- **Lấy thêm thông tin metadata từ MongoDB** để đảm bảo nội dung đủ chính xác.
- **Gửi yêu cầu đến OpenAI GPT-4** để sinh Mindmap/Cheatsheet.
  - **Kết quả được lưu vào Redis Cache** với TTL = **24 giờ**.
  - **API Gateway trả kết quả về User.**

Mục tiêu: Giảm số lần sinh Mindmap bằng AI, tối ưu hiệu suất với caching.

Bước 4: Xử lý nội dung từ YouTube, OCR, Speech-to-Text

Người dùng tải lên video/audio/hình ảnh.
Hệ thống tự động trích xuất nội dung:

- **YouTube API & Whisper:** Lấy transcript từ video.
- **Tesseract OCR:** Trích xuất văn bản từ ảnh/PDF.
- **Wav2Vec2:** Chuyển đổi giọng nói thành văn bản.
  - **Dữ liệu trích xuất được gửi đến Upload & Preprocessing** để xử lý giống như tài liệu văn bản thông thường.
  - **Các nội dung này cũng được lưu trữ vào hệ thống caching nếu cần thiết.**

Mục tiêu: Hỗ trợ đa dạng nguồn dữ liệu, giúp người dùng khai thác thông tin từ video và ảnh.

Cơ chế caching & tối ưu hóa hiệu suất

Caching với TTL (Time-To-Live)

- **Câu trả lời từ GPT-4:** Cache trong **24 giờ** để giảm tải API.
- **Mindmap & Cheatsheet:** Cache trong **24 giờ** để tránh tính toán lại.

Cache Invalidation (Xóa cache khi dữ liệu thay đổi)

- Khi **người dùng cập nhật hoặc thay thế tài liệu**, cache cũ sẽ bị **xóa tự động** để đảm bảo dữ liệu mới luôn chính xác.

Lưu trữ embedding trong Redis (Vector DB)

- Giúp tìm kiếm thông tin nhanh hơn so với lưu trực tiếp trong database.

Xử lý bất đồng bộ với Redis Queue

- **Các tác vụ lớn (ví dụ: trích xuất video, sinh Mindmap)** sẽ được xử lý bất đồng bộ để không làm chậm hệ thống.

3.3 Ưu Điểm Cạnh Tranh

1. Kiến trúc hệ thống

Hệ thống được chia thành nhiều services riêng biệt, mỗi dịch vụ đảm nhận một chức năng cụ thể để tối ưu hóa hiệu suất và khả năng mở rộng:

- **Upload & Preprocessing**: Xử lý và tiền xử lý dữ liệu, giúp trích xuất thông tin từ tài liệu người dùng.
- **Vector Database & Indexing**: Lưu trữ và lập chỉ mục vector cho các nội dung từ tài liệu, hỗ trợ tìm kiếm ngữ nghĩa hiệu quả.

- **Question-Answering (RAG)**: Ứng dụng Retrieval-Augmented Generation (RAG) để cung cấp câu trả lời chính xác dựa trên tài liệu đã cung cấp.
- **Mindmap & Cheatsheet**: Tạo sơ đồ tư duy và bảng ghi nhớ từ tài liệu, hỗ trợ việc học tập và tổng hợp thông tin.
- **API Gateway**: Điều phối yêu cầu giữa các dịch vụ, đảm bảo tính nhất quán và an toàn dữ liệu.
- **Authentication & User Management**: Quản lý người dùng và xác thực, đảm bảo an toàn truy cập hệ thống.

Toàn bộ các thành phần đều sử dụng **FastAPI** (Python) hoặc **Express.js** (Node.js), giúp đảm bảo tốc độ xử lý cao và dễ dàng triển khai.

2. Hệ thống lưu trữ và xử lý dữ liệu

- **Vector Database**: Redis – hỗ trợ tìm kiếm văn bản nhờ khả năng xử lý embedding hiệu quả.
- **Database (Metadata)**: MongoDB – lưu trữ metadata của tài liệu, cho phép truy vấn nhanh và linh hoạt.
- **Storage**: MinIO – cung cấp giải pháp lưu trữ đối tượng, giúp lưu trữ tài liệu người dùng một cách hiệu quả.
- **Queue & Message Broker**: Redis Queue – đảm bảo xử lý yêu cầu không đồng bộ, giúp hệ thống hoạt động mượt mà ngay cả khi tải cao.

3. Trí tuệ nhân tạo và Xử lý ngôn ngữ tự nhiên

- **LLM API**: OpenAI GPT-4 mini – cung cấp khả năng tạo phản hồi chất lượng cao, hỗ trợ hỏi đáp thông minh và tạo nội dung từ tài liệu.
- **YouTube Transcript**: Kết hợp **YouTube API** và **Whisper** để trích xuất nội dung từ video, hỗ trợ người dùng tra cứu và tóm tắt thông tin.
- **Model Embedding**: all-MiniLM-L6-v2 – giúp mã hóa nội dung tài liệu thành vector để phục vụ truy vấn và tìm kiếm chính xác hơn.

4. Bảo mật và triển khai hệ thống

- **Containerization & Orchestration**: Docker & Ngrok – giúp đóng gói và triển khai hệ thống linh hoạt, đảm bảo khả năng mở rộng dễ dàng.
- **OCR (Nhận dạng ký tự quang học)**: Tesseract OCR – hỗ trợ trích xuất văn bản từ hình ảnh và tài liệu scan, giúp mở rộng khả năng xử lý dữ liệu.

4. Đối Tượng Sử Dụng

Học sinh, sinh viên, giáo viên: Sử dụng **NoteUS** để tạo flashcard, mindmap và cheatsheet phục vụ học tập. Giúp tóm tắt nội dung từ nhiều tài liệu khác nhau để tiết kiệm thời gian.

Phóng viên, nhà nghiên cứu: Hỗ trợ phân tích thông tin từ nhiều tài liệu, trích dẫn chính xác nội dung và phát hiện đạo văn. Giúp phóng viên tóm tắt tin tức từ video và tài liệu một cách nhanh chóng, đảm bảo tính bảo mật.

Chuyên gia xử lý dữ liệu: Cung cấp công cụ tự động trích xuất, tổng hợp và tổ chức dữ liệu từ nhiều nguồn, hỗ trợ trong việc phân tích thông tin một cách hiệu quả.

5. Tính Khả Thi

5.1 Kế Hoạch Phát Triển

Giai đoạn (RUP)	*Sprint*	*Mục tiêu*	*Bắt đầu*	*Kết thúc*
Khởi tạo	Sprint 1	Lập kế hoạch và nghiên cứu	06/01/2025	12/01/2025
		TẾT		
Phát triển	Sprint 2	Chi tiết hóa tính năng, thiết kế cơ sở dữ liệu và kiến trúc hệ thống	17/02/2025	23/02/2025
Xây dựng	Sprint 3	Xây dựng nền tảng cho hệ thống	17/02/2025	23/02/2025
	Sprint 4	Triển khai các chức năng cơ bản	10/03/2025	16/03/2025
		Triển khai các chức năng chi tiết	17/03/2025	23/03/2025
	Sprint 5	Hoàn thiện tính năng	24/03/2025	30/03/2025
Phát hành	Sprint 6	Kiểm thử tích hợp	31/03/2025	06/04/2025
		Triển khai nâng cao và kiểm thử hệ thống	07/04/2025	13/04/2025
	Sprint 7	Hoàn thiện và chuẩn bị phát hành	14/04/2025	20/04/2025

Kế Hoạch Giai Đoạn và Lặp

1. Khởi tạo (Inception)

Sprint 1

- **Thời gian:** 06/01/2025 - 12/01/2025
- **Mục tiêu chung:** Lập kế hoạch và nghiên cứu
- **Mục tiêu cụ thể:**
    - Thành lập nhóm và xác định vai trò thành viên.
    - Xây dựng ý tưởng cốt lõi về hệ thống quản lý notebook và tổng hợp thông tin thành mindmap/cheatsheet.
    - Nghiên cứu kiến trúc hệ thống, xác định công nghệ cần sử dụng.
    - Phân tích yêu cầu hệ thống, bao gồm các API chính như Upload & Preprocessing, Question-Answering, Mindmap & Cheatsheet.

2. Làm rõ yêu cầu (Elaboration)

Sprint 2

- **Thời gian:** 10/02/2025 - 16/02/2025
- **Mục tiêu chung:** Chi tiết hóa tính năng, thiết kế cơ sở dữ liệu và kiến trúc hệ thống

- **Mục tiêu cụ thể:**
    - Thiết kế giao diện demo của hệ thống (UI/UX cho quản lý notebook và mindmap/cheatsheet).
    - Thiết kế sơ đồ UML cho hệ thống backend (MVC , API Gateway).
    - Chi tiết hóa danh sách API endpoints và phương thức giao tiếp giữa các MVC.
    - Xây dựng cấu trúc thư mục của dự án trên GitHub theo mô hình MVC.

3. Xây dựng (Construction)

Sprint 3

- **Thời gian:** 17/02/2025 - 02/03/2025
- **Mục tiêu chung:** Triển khai các tính năng cơ bản và kiểm thử đơn vị
- **Mục tiêu cụ thể:**
    - Triển khai Service Upload & Preprocessing: xử lý file PDF, DOC, TXT, OCR ảnh, transcript YouTube.
    - Người dùng có thể tải lên tài liệu và xem nội dung trên giao diện web.
    - Lưu trữ tài liệu và metadata
    - Tích hợp công nghệ để xử lý OCR và indexing dưới nền.
    - Kiểm thử đơn vị cho các service cơ bản.

Sprint 4

- **Thời gian:** 03/03/2025 - 16/03/2025
- **Mục tiêu chung:** Xây dựng nền tảng chính cho hệ thống.
- **Mục tiêu cụ thể:**
    - Triển khai Service Vector Database & Indexing: chuyển đổi tài liệu thành vector embedding.
    - Tích hợp truy vấn vector database vào hệ thống backend.
    - Triển khai Service Question-Answering (RAG**)**: nhận câu hỏi, truy xuất dữ liệu từ vector database, gọi OpenAI API để sinh câu trả lời.
    - Xây dựng giao diện người dùng cho tính năng tìm kiếm thông tin trong tài liệu đã tải lên.

Sprint 5

- **Thời gian:** 17/03/2025 - 30/03/2025
- **Mục tiêu chung:** Hoàn thiện tính năng và kiểm thử tích hợp.
- **Mục tiêu cụ thể:**
    - Người dùng có thể đặt câu hỏi và nhận phản hồi từ hệ thống QA.
    - Triển khai Service Mindmap & Cheatsheet Generator: tổng hợp thông tin từ tài liệu đã upload để sinh mindmap và cheatsheet.
    - Kiểm tra API Gateway và đảm bảo các API hoạt động đúng.
    - Kiểm thử hiệu năng hệ thống với tải lớn.

4. Các giai đoạn phát hành

Phát hành Alpha

- **Thời gian:** 31/03/2025 - 13/04/2025
- **Mục đích:** Kiểm tra các chức năng cốt lõi và xác định lỗi lớn.
- **Phạm vi:**
    - Đăng ký người dùng, tải tài liệu, xử lý OCR, truy vấn thông tin, sinh câu trả lời.
    - Demo giao diện với các tính năng cơ bản.
- **Đối tượng:** Nội bộ nhóm phát triển, nhóm thử nghiệm giới hạn.
- **Hạn chế:** Chưa có đầy đủ tính năng như mindmap nâng cao, tối ưu hiệu suất chưa hoàn thiện.
- **Kết quả mong đợi:** Phát hiện lỗi lớn, cải thiện hiệu suất và tính năng.

Phát hành Beta

- **Thời gian:** 14/04/2025 - 27/04/2025
- **Mục đích:** Thử nghiệm trên phạm vi rộng hơn và thu thập phản hồi từ người dùng tiềm năng.
- **Phạm vi:**
    - Tất cả các tính năng từ bản Alpha, cộng thêm sinh mindmap nâng cao, quản lý notebook tương tự Notion.
    - Tích hợp phản hồi người dùng, tối ưu hiệu suất.
- **Đối tượng:** Nhóm người dùng mở rộng, bao gồm khách hàng tiềm năng.
- **Hạn chế:** Một số lỗi nhỏ có thể vẫn còn.
- **Kết quả mong đợi:** Cải thiện trải nghiệm người dùng, sửa lỗi cuối cùng.

Phát hành Chính Thức (Final Release)

- **Thời gian:** 28/04/2025
- **Mục đích:** Ra mắt sản phẩm ổn định với đầy đủ tính năng, tối ưu hiệu suất.
- **Phạm vi:**
    - Hoàn thiện tất cả tính năng theo kế hoạch.
    - Mindmap và cheatsheet được tối ưu hóa và có khả năng tùy chỉnh sâu.
    - Tích hợp lịch, nhắc nhở và quản lý kế hoạch.
- **Đối tượng:** Toàn bộ người dùng tiềm năng.
- **Tiêu chí phát hành:**
    - Hoàn thiện tất cả tính năng, hiệu suất mượt mà, giao diện trực quan.
    - Bảo mật hệ thống và kiểm thử toàn diện.
- **Lịch phát hành:**
    - **Ngày phát hành chính thức:** 28/04/2025
    - **Kênh phân phối:** Web trên máy tính và điện thoại di động.

5.2 Ngân Sách Dự Kiến

Khoản chi	Mô tả	Chi phí/tháng	Chi phí một lần
Chi phí nhân sự	Lương cho đội ngũ phát triển và triển khai	$10,000	-
Chi phí phần cứng và hạ tầng	Máy chủ, thiết bị mạng và phần cứng khác	$500	-
Chi phí phần mềm	Bản quyền các phần mềm cần thiết	-	$2000
Chi phí tư vấn và đào tạo	Thuê chuyên gia và đào tạo nhân viên	-	$3000
Chi phí bảo trì phần mềm	Cập nhật, sửa lỗi nâng cấp	$1000	-
Chi phí hỗ trợ kỹ thuật	Duy trì đội ngũ hỗ trợ kỹ thuật	$800	-
Chi phí sử dụng API OpenAI	15 triệu token / tháng	$1,125	-
Chi phí tinh chỉnh mô hình	1 triệu token, 3 vòng huấn luyện	-	$1,350
Chi phí xử lý dữ liệu	Xử lý dữ liệu giọng nói và chữ viết tay	-	$3000
Chi phí nghiên cứu thị trường	Thu thập và phân tích thông tin thị trường	-	$1,500
Chi phí quảng cáo	Quảng cáo trên các kênh truyền thông	$2,000	-
Chi phí tham gia sự kiện	Tham gia hội thảo, triển lãm	-	$2,500
Chi phí lập kế hoạch và giám sát	Lập kế hoạch, theo dõi tiến độ và quản lý rủi ro	$1,200	-
Chi phí hành chính	Văn phòng phẩm và thiết bị văn phòng	$300	-

6. Liên Hệ

Thành Viên	Email	SĐT	Resume
Phạm Nguyên Hải Long	hailong552004@gmail.com	0968521483	PhamNguy...
Võ Thành Nghĩa	vothanhnghia270604@gmail.com	0787983328	VoThanhNg...
Hà Tuấn Anh	hatuananh2k4@gmail.com	0913780141	HaTuanAnh...
Dương Quang Thắng	quangthangduongt@gmail.com	0945842813	DuongQuan...
Nguyễn Quốc Thắng	nguyenquocthang934@gmail.com	0393029805	NguyenQuo...