**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**A blue and white sign with red text

Description automatically generated**

**TIỂU LUẬN TỐT NGHIỆP**

**XÂY DỰNG HỆ THỐNG CHATBOT THÔNG MINH HỖ TRỢ TÌM KIẾM TÀI LIỆU HỌC TẬP**

**Thành phố Hồ Chí Minh - 2024**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**A blue and white sign with red text

Description automatically generated**

**XÂY DỰNG HỆ THỐNG CHATBOT THÔNG MINH HỖ TRỢ TÌM KIẾM TÀI LIỆU HỌC TẬP**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | Ths. Ma Ngân Giang |
| **Sinh viên thực hiện:** | 47.01.104.132 – Nguyễn Phan Trà My |
| **Lớp:** | 47.01.CNTT.c |
| **Niên khóa:** | 2024 – 2025 |

**Thành phố Hồ Chí Minh - 2024**

# LỜI CẢM ƠN

Trong quá trình thực hiện và hoàn thành tiểu luận này, em xin gửi lời cảm ơn chân thành và sâu sắc đến cô ThS. Ma Ngân Giang, người đã tận tâm hướng dẫn, chỉ bảo em trong suốt quá trình nghiên cứu và viết bài. Sự tận tụy và nhiệt tình của thầy/cô không chỉ giúp em hoàn thành tốt công việc mà còn mở rộng kiến thức và khả năng tư duy của em.

Em cũng xin bày tỏ lòng biết ơn đến Ban Giám hiệu và toàn thể các thầy cô tại Khoa Công nghệ thông tin, những người đã tạo điều kiện và cung cấp nguồn tài liệu học tập quý báu, giúp em hoàn thiện bài tiểu luận này. Sự hỗ trợ của các thầy cô là động lực lớn để em nỗ lực không ngừng trong quá trình học tập và nghiên cứu.

Cuối cùng, em xin gửi lời cảm ơn chân thành đến gia đình, bạn bè, những người luôn bên cạnh, động viên và chia sẻ trong suốt quá trình thực hiện bài tiểu luận. Sự ủng hộ và khích lệ của mọi người là nguồn động viên to lớn để em vượt qua những khó khăn và thử thách.

Mặc dù đã hết sức cố gắng, tiểu luận chắc chắn không tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý từ thầy/cô và bạn đọc để hoàn thiện bài nghiên cứu của mình.

Em xin chân thành cảm ơn.

# MỤC LỤC

[LỜI CẢM ƠN 3](#_Toc186201041)

[MỤC LỤC 4](#_Toc186201042)

[BẢNG CÁC KÝ HIỆU, CHỮ VIẾT TẮT 6](#_Toc186201043)

[DANH MỤC CÁC BẢNG BIỂU 7](#_Toc186201044)

[DANH MỤC CÁC HÌNH VẼ 8](#_Toc186201045)

[TÓM TẮT 9](#_Toc186201046)

[CHƯƠNG 1: MỞ ĐẦU 10](#_Toc186201047)

[1.1. Lý do chọn đề tài 10](#_Toc186201048)

[1.2. Mục tiêu nghiên cứu 10](#_Toc186201049)

[1.3. Đối tượng và phạm vi nghiên cứu 10](#_Toc186201050)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 11](#_Toc186201051)

[2.1. Tổng quan về chatbot: 11](#_Toc186201052)

[2.1.1. Chatbot là gì? 11](#_Toc186201053)

[2.1.2. Chatbot hoạt động như thế nào? 11](#_Toc186201054)

[2.1.3. Ứng dụng của chatbot trong giáo dục: 12](#_Toc186201055)

[2.2. Công nghệ AI trong chatbot: 12](#_Toc186201056)

[2.3. Vector Embeddings 13](#_Toc186201057)

[2.4. Tạo tăng cường truy suất RAG: 13](#_Toc186201058)

[2.5. Tổng quan về chương trình giáo dục phổ thông 2018 14](#_Toc186201059)

[2.6. Công nghệ sử dụng: 15](#_Toc186201060)

[2.6.1. Ngôn ngữ lập trình 15](#_Toc186201061)

[CHƯƠNG 3: PHÂN TÍCH HỆ THỐNG 16](#_Toc186201062)

[3.1. Tài liệu học tập: 16](#_Toc186201063)

[3.2. Tiền xử lý dữ liệu: 16](#_Toc186201064)

[3.3. Tạo vector embeddings: 16](#_Toc186201065)

[3.4. Lưu trữ và tổ chức dữ liệu: 16](#_Toc186201066)

[3.5.1. Kiến trúc tổng thể: 17](#_Toc186201067)

[3.5.2. Các thành phần chính: 17](#_Toc186201068)

[3.5.3. Luồng hoạt động chính của hệ thống: 18](#_Toc186201069)

[CHƯƠNG 4: THIẾT KẾ HỆ THỐNG 19](#_Toc186201070)

[4.1. Kiến trúc hệ thống: 19](#_Toc186201071)

[4.1.1. Sơ đồ usecase: 19](#_Toc186201072)

[4.2. Thiết kế cơ sở dữ liệu: 19](#_Toc186201073)

[4.3. Thiết kế API: Thiết kế các endpoint API và tương tác với các thành phần khác. 20](#_Toc186201074)

[4.4. Tích hợp GPT-4o và LlamaIndex: Mô tả quá trình tích hợp các công nghệ AI. 20](#_Toc186201075)

[CHƯƠNG 5: KIỂM THỬ, ĐÁNH GIÁ HỆ THỐNG VÀ KẾT LUẬN 21](#_Toc186201076)

[5.1. Kiểm thử chức năng: Phương pháp kiểm thử đơn vị và tích hợp. 21](#_Toc186201077)

[5.2. Kiểm thử hiệu năng: Thử nghiệm về hiệu suất và tốc độ của hệ thống. 21](#_Toc186201078)

[5.3. Đánh giá kết quả: Mức độ chính xác và hiệu quả của hệ thống chatbot. 21](#_Toc186201079)

[5.4. Kết luận: Tổng kết những kết quả chính đạt được. 21](#_Toc186201080)

[5.5. Hướng phát triển trong tương lai: Các đề xuất cải tiến và phát triển sau này. 21](#_Toc186201081)

[KẾT LUẬN 24](#_Toc186201082)

[DANH MỤC TÀI LIỆU THAM KHẢO 25](#_Toc186201083)

[PHỤ LỤC 26](#_Toc186201084)

# BẢNG CÁC KÝ HIỆU, CHỮ VIẾT TẮT

**Viết tắt Tiếng Anh Tiếng Việt**

NLP Neuro-Linguistic Programming Lập trình ngôn ngữ tư duy

ML Neuro-Linguistic Programming Lập trình ngôn ngữ tư duy

NLU Neuro-Linguistic Programming Lập trình ngôn ngữ tư duy

RAG Neuro-Linguistic Programming Lập trình ngôn ngữ tư duy

FAQ Neuro-Linguistic Programming Lập trình ngôn ngữ tư duy

OCR Optical Character Recognition Nhận dạng kí tự quang học

# DANH MỤC CÁC BẢNG BIỂU

# DANH MỤC CÁC HÌNH VẼ

# TÓM TẮT

Dự án "Xây dựng hệ thống chatbot thông minh hỗ trợ tìm kiếm tài liệu học tập" tập trung phát triển một công cụ chatbot sử dụng mô hình Retrieval Augmented Generation (RAG) kết hợp với GPT-4o để hỗ trợ tìm kiếm tài liệu học tập từ chương trình giáo dục phổ thông 2018. Hệ thống được thiết kế nhằm cải thiện việc truy xuất thông tin học tập, giúp người dùng có thể tìm kiếm và truy cập tài liệu nhanh chóng và hiệu quả. Các tính năng chính bao gồm tìm kiếm tài liệu theo từ khóa, phân loại theo chủ đề, và đề xuất tài liệu dựa trên lịch sử tìm kiếm.

# MỞ ĐẦU

## Lý do chọn đề tài

Trong thời đại công nghệ số, việc truy cập thông tin học tập một cách nhanh chóng và hiệu quả là một nhu cầu thiết yếu đối với học sinh, sinh viên, và giáo viên. Tuy nhiên, việc tìm kiếm tài liệu học tập phù hợp giữa hàng loạt thông tin có thể gặp nhiều khó khăn, đặc biệt khi người dùng cần tìm kiếm theo một chủ đề cụ thể hoặc theo yêu cầu riêng biệt.

Sự phát triển của trí tuệ nhân tạo (AI) và các hệ thống chatbot thông minh đã mở ra những tiềm năng mới trong việc hỗ trợ tìm kiếm và truy xuất thông tin học tập một cách dễ dàng. Một hệ thống chatbot không chỉ giúp người dùng tìm kiếm nhanh chóng, mà còn có khả năng hiểu được các câu hỏi phức tạp và đưa ra các gợi ý tài liệu liên quan, từ đó nâng cao hiệu quả học tập và giảng dạy.

Vì những lý do trên, đề tài "Xây dựng hệ thống chatbot thông minh hỗ trợ tìm kiếm tài liệu học tập" được chọn với mục tiêu xây dựng một công cụ hỗ trợ hiệu quả cho học sinh, sinh viên, giáo viên trong việc tìm kiếm và truy cập các tài liệu học tập.

## Mục tiêu nghiên cứu

Phát triển một hệ thống chatbot thông minh có khả năng hỗ trợ người dùng trong việc tìm kiếm và truy cập tài liệu học tập một cách nhanh chóng và hiệu quả.

## Đối tượng và phạm vi nghiên cứu

* Đối tượng: Sinh viên, giáo viên, và những người có nhu cầu tìm kiếm tài liệu học tập.
* Phạm vi: Hệ thống sẽ tập trung vào tài liệu học tập của chương trình giáo dục phổ thông 2018

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Tổng quan về chatbot:

### Chatbot là gì?

Chatbot là chương trình máy tính mô phỏng lại các đoạn hội thoại của con người, văn bản hoặc lời nói, cho phép con người tương tác với các thiết bị kỹ thuật số như thể họ đang giao tiếp với một người thật. Chatbot có thể đơn giản như các chương trình cơ bản trả lời một truy vấn đơn giản bằng một phản hồi một dòng hoặc phức tạp như trợ lý kỹ thuật số học hỏi và phát triển để cung cấp mức độ cá nhân hóa ngày càng tăng khi chúng thu thập và xử lý thông tin.

### Chatbot hoạt động như thế nào?

Với sự thúc đẩy của AI, các quy tắc tự động, xử lý ngôn ngữ tự nhiên (NLP) và học máy (ML) thì chatbot xử lý dữ liệu để đưa ra phản hồi cho mọi loại yêu cầu.

Có 2 loại chatbot chính:

* **Task-oriented (declarative) chatbots (Chatbot khai báo)** những chương trình đơn lẻ tập trung vào thực hiện một chức năng. Bằng cách cách sử dụng các quy tắc, NLP và một chút ít ML, Task-oriented (declarative) chatbots tạo ra các phản hồi tự động nhưng mang tính hội thoại, trả lời các câu hỏi hoặc yêu cầu từ người dùng. Các tương tác với chatbot định hướng nhiệm vụ thường rất cụ thể và có cấu trúc rõ ràng, phù hợp nhất với các chức năng hỗ trợ và dịch vụ. Một cách dễ hiểu, đây có thể được xem như phiên bản tương tác nâng cao của các câu hỏi thường gặp (FAQ) nhưng được tích hợp khả năng phản hồi linh hoạt hơn.
* **Data-driven and predictive (conversational) chatbots (Chatbot dự đoán)** thường được xem là trợ lý ảo. Đây loại chatbot tiên tiến với khả năng tương tác và cá nhân hóa cao hơn nhiều so với các chatbot chỉ thực hiện nhiệm vụ đơn giản. Các chatbot này sử dụng các công nghệ hiện đại như hiểu ngôn ngữ tự nhiên (NLU), xử lý ngôn ngữ tự nhiên (NLP) và học máy (ML) để không ngừng học hỏi và cải thiện trong quá trình hoạt động. Nhờ vậy, chúng có khả năng nhận biết ngữ cảnh, áp dụng trí tuệ dự đoán và phân tích để cá nhân hóa trải nghiệm người dùng dựa trên hồ sơ và hành vi trước đây. Chatbot này không chỉ hiểu sâu về ngữ cảnh mà còn có thể học hỏi sở thích của người dùng theo thời gian, từ đó đưa ra các gợi ý phù hợp hoặc thậm chí dự đoán nhu cầu trong tương lai. Ngoài ra, chúng còn có thể chủ động khởi xướng cuộc trò chuyện, chẳng hạn như nhắc nhở, gợi ý hoặc cung cấp thông tin hữu ích.

### Ứng dụng của chatbot trong giáo dục:

Chatbot hiện đang được ứng dụng rộng rãi trong nhiều lĩnh vực và trong số đó không thể thiếu giáo dục. Nhờ khả năng hỗ trợ học tập thông minh, tương tác tự nhiên và cung cấp thông tin một cách nhanh chóng mà chatbot được ứng dụng rộng rãi trong lĩnh vực giáo dục.

Ứng dụng chatbot vào giáo dục thì nổi bật với tính năng giải đáp câu hỏi trong học tập. thông qua tích công nghệ xử lý ngôn ngữ tự nhiên (NLP) và trí tuệ nhân tạo, chatbot có thể hiểu được ngữ cảnh của các câu hỏi từ học sinh và sinh viên, từ đó đưa ra câu trả lời chính xác và chi tiết. Không chỉ dừng lại ở việc trả lời các câu hỏi đơn giản, chatbot còn có khả năng giải thích các vấn đề phức tạp, cung cấp ví dụ minh họa và hướng dẫn từng bước để người học hiểu rõ hơn.

## Công nghệ AI trong chatbot:

GPT-4o là mô hình biến đổi được đào tạo trước đa phương thức và đa ngôn ngữ được phát triển bởi OpenAI, một tổ chức nổi tiếng trong lĩnh vực nghiên cứu trí tuệ nhân tạo. Chữ “o” trong GPT-4o là viết tắt của omni và nhấn mạnh rằng GPT-4o là một mô hình AI đa phương thức với khả năng âm thanh và hình ảnh.

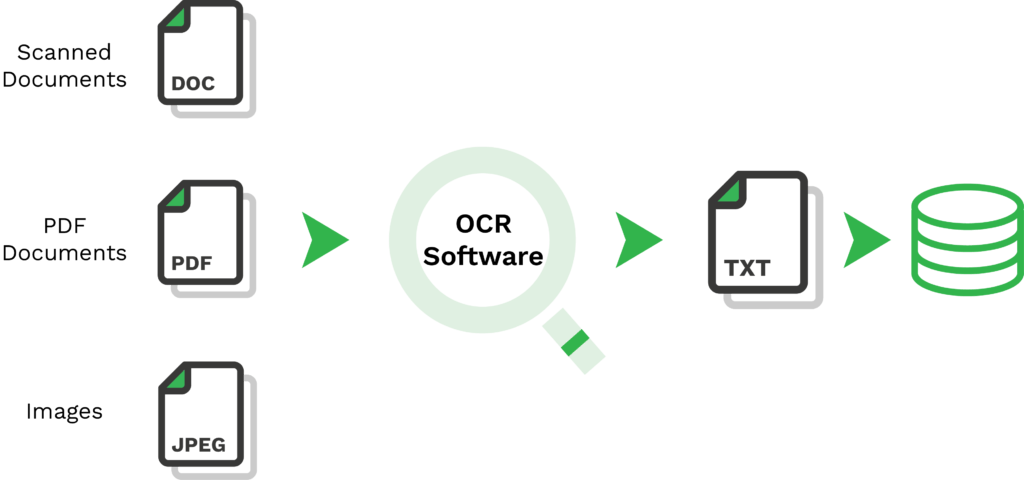
GPT-4o có chức năng trả lời câu hỏi, phân tích và tóm tắt tài liệu, phân tích dữ liệu, phân tích tình cảm, …

Nhờ kết hợp các công nghệ tiên tiến như xử lý ngôn ngữ tự nhiên (NLP), học sâu (Deep Learning), và học tăng cường (Reinforcement Learning), GPT-4o có khả năng hiểu sâu ngữ cảnh của các cuộc hội thoại mà model GPT-4o có thể đưa ra những phản hồi có tính chính xác cao, phù hợp và có giá trị.

## Nhận dạng ký tự quang học OCR - *Optical Character Recognition*

OCR là quá trình chuyển đổi kí tự văn bản trong hình ảnh thành định dạng văn bản có thể đọc được bằng máy.

OCR hoạt động bằng cách cho máy quét đọc tài liệu và chuyển đổi chúng thành dữ liệu nhị phân, phần mềm OCR sẽ phân tích hình ảnh được quét để phân vùng sáng thành nền của văn bản còn vùng tối là các kí tự văn bản. Sau đó phần mềm OCR sẽ làm sạch hình ảnh và loại bỏ lỗi để chuẩn bị cho việc đọc văn bản. Tiếp đến là Sử dụng thuật toán So khớp mẫu (**Pattern matching**) hoặc Trích xuất đặc điểm (**Feature extraction**) để nhận dạng văn bản. Cuối cùng hệ thống sẽ chuyển đổi dữ liệu văn bản thành một tập tin được lưu trên máy tính.



Hình 2.1- Cách hoạt động của chương trình OCR

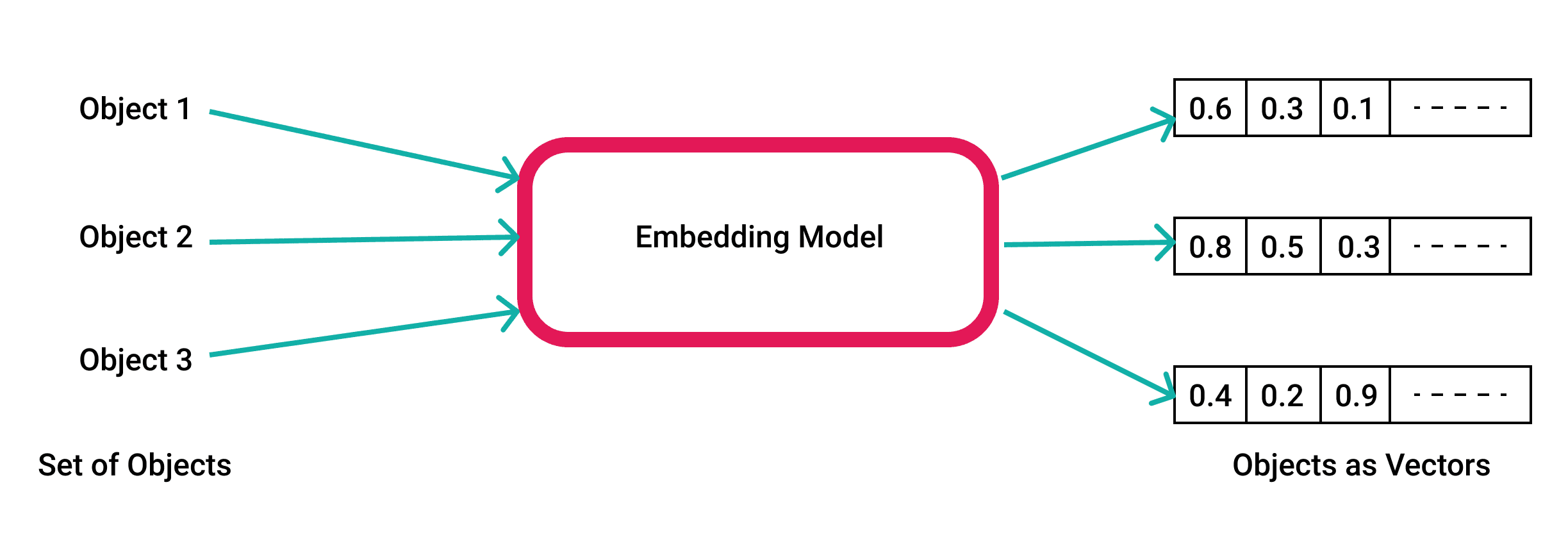
## Vector Databases và Vector Embeddings:

Vector Databases là hệ thống được thiết kế chuyên sử dụng để lưu trữ và truy xuất dữ liệu vector.

Vector embeddings là một kỹ thuật trong học máy và xử lý ngôn ngữ tự nhiên (NLP) nhằm chuyển đổi các đối tượng (như từ, câu, hay tài liệu) thành các vector số (dãy số). Mục tiêu của việc sử dụng vector embeddings là biểu diễn các đối tượng này trong một không gian vector có nghĩa, giúp các thuật toán máy học dễ dàng hơn trong việc xử lý và so sánh.

Vector embeddings, từ bao hàm, câu và danh mục tài liệu đều là nền tảng cơ bản trong NLP cho các nhiệm vụ như phân tích tình cảm và tóm tắt tài liệu. Có các kĩ thuật như Word2Vec, GloVe, BERT và các dịch vụ như OpenAI cung cấp nhiều cách tiếp cận khác nhau để tạo ra các vector embeddings, nâng cao khả năng ứng dụng của chúng trong nhiều nhiệm vụ trong xử lý ngôn ngữ tự nhiên.

Vector Embeddings thường được ứng dụng trong dịch thuật, chatbot, phân tích dữ liệu cho nghiên cứu thị trường và hệ thống đề xuất trong thương mại điện tử. Chúng đóng vai trò trong giải quyết các thách thức như khả năng mở rộng hoặc nắm bắt sở thích phức tạp của người dùng và có thể rất hiệu quả trong việc tạo điều kiện học tập năng động để thích ứng với các xu hướng thay đổi.



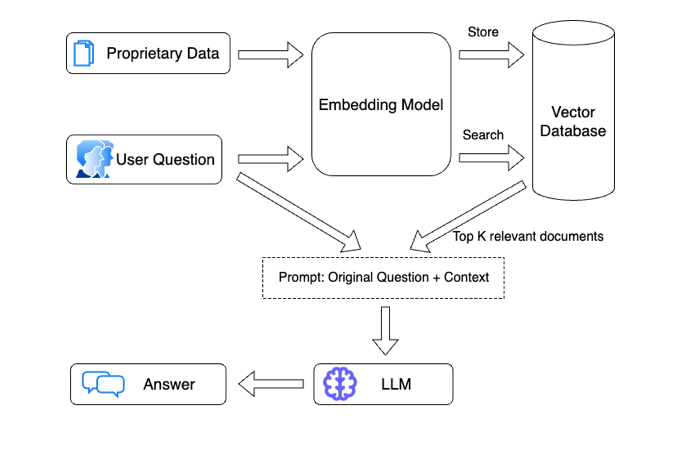
Hình 2.2- Quy Trình Biểu Diễn Đối Tượng Dưới Dạng Vector Bằng Mô Hình Embedding

## Tạo tăng cường truy suất RAG:

RAG là quá trình tối ưu hóa đầu ra của một mô hình ngôn ngữ lớn, vì vậy nó tham khảo một cơ sở kiến thức có thẩm quyền bên ngoài các nguồn dữ liệu đào tạo của nó trước khi tạo phản hồi. Mô hình ngôn ngữ lớn (LLM) được đào tạo trên khối lượng dữ liệu khổng lồ và có sử dụng hàng tỷ tham số để tạo ra đầu ra ban đầu cho các nhiệm vụ như trả lời câu hỏi, dịch ngôn ngữ và hoàn thành câu. RAG mở rộng các khả năng vốn đã mạnh mẽ của LLM đến các miền cụ thể hoặc cơ sở kiến thức nội bộ của tổ chức, tất cả mà không cần đào tạo lại mô hình. Đây là một cách tiếp cận hiệu quả về chi phí để cải thiện đầu ra LLM, để nó vẫn phù hợp, chính xác và hữu ích trong nhiều bối cảnh khác nhau.

Mô hình RAG trải qua các giai đoạn sau:

* Tạo dữ liệu bên ngoài: Dữ liệu bên ngoài là dữ liệu mới bên ngoài của tập dữ liệu đào tạo ban đầu của LLM. Dữ liệu có thể tồn tại ở nhiều định dạng khác nhau như tệp, bản ghi cơ sở dữ liệu hoặc văn bản dạng dài. Một kỹ thuật AI khác, được gọi là mô hình ngôn ngữ nhúng, chuyển đổi dữ liệu thành các dạng trình bày số và lưu trữ nó trong cơ sở dữ liệu véc-tơ. Quá trình này tạo ra một thư viện kiến thức mà các mô hình AI tạo sinh có thể hiểu được.
* Truy xuất thông tin liên quan: Truy vấn người dùng được chuyển đổi thành dạng trình bày véc-tơ và được mang đi so sánh với cơ sở dữ liệu véc-tơ. Sự liên quan được tính toán và thiết lập bằng các phép tính và dạng trình bày véc-tơ toán học.
* Tăng cường lời nhắc LLM: mô hình RAG tăng cường đầu vào của người dùng (hoặc lời nhắc) bằng cách thêm dữ liệu truy xuất liên quan trong ngữ cảnh. Bước này sử dụng các kỹ thuật tạo câu lệnh để giao tiếp hiệu quả với LLM. Với lời nhắc tăng cường, các mô hình ngôn ngữ lớn có thể tạo ra câu trả lời chính xác cho các truy vấn của người dùng



Hình 2.3- Cách mô hình RAG hoạt động

## Tổng quan về chương trình giáo dục phổ thông 2018

Đây là chương trình giáo dục mới nhằm đáp ứng yêu cầu đổi mới văn bản và toàn diện giáo dục với trọng tâm là phát triển phẩm chất năng lực.

Chương trình giáo dục phổ thông cụ thể hoá mục tiêu giáo dục phổ thông, giúp học sinh làm chủ kiến thức phổ thông, biết vận dụng hiệu quả kiến thức, kĩ năng đã học vào đời sống và tự học suốt đời, có định hướng lựa chọn nghề nghiệp phù hợp, biết xây dựng và phát triển hài hoà các mối quan hệ xã hội, có cá tính, nhân cách và đời sống tâm hồn phong phú, nhờ đó có được cuộc sống có ý nghĩa và đóng góp tích cực vào sự phát triển của đất nước và nhân loại.

## Công nghệ sử dụng:

### LlamaIndex:

Là một khung điều phối giúp tích hợp dữ liệu với các mô hình AI để trả lời câu hỏi cảu người dùng

Ưu điểm:

* Hỗ trợ nhiều nguồn dữ liệu (cơ sở dữ liệu, văn bản, API, dữ liệu phi cấu trúc
* Tìm kiếm thông tin hiệu quả dựa trên độ tương đồng vector.
* Tự động hóa tiền xử lý và tối ưu lưu trữ dữ liệu.



Hình 2.4- LlamaIndex logo

### Python:

Python là ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, máy học (ML) và khoa học dữ liệu. Đây là ngôn ngữ lập trình dễ học, có thể chạy trên nhiều nền tảng khác nhau nên đây là ngôn ngữ lập trình chính cho backend. Với thư viện phong phú giúp hỗ trợ các công cụ xử lý ngôn ngữ tự nhiên (NLP), học máy (ML) và trí tuệ nhân tạo như spaCy, OpenAI, LlamaIndex, FastAPI, SQLAlchemy, Pandas, NumPy, …

Ưu điểm

* Dễ học và dễ sử dụng
* Thư viện phong phú
* Ứng dụng được trong nhiều lĩnh vực



Hình 2.5- Python logo

### React:

React là thư viện JavaScript mã nguồn mở, được dùng để xây dựng giao diện cho người dùng.

Ưu điểm:

* Nhanh và dễ học
* Tính năng linh hoạt
* Phát triển ứng dụng với cung cụ mạnh mẽ

A blue and black symbol

Description automatically generated

Hình 2.6- React logo

### PostgreSQL:

PostgreSQL là hệ thống quản trị cơ sở dữ liệu quan hệ và đối tượng (object-relational database management system) miễn phíu và nguồn mở tiến tiến nhất hiện nay.

Ưu điểm:

* Có các tính năng và tiện ích mở rộng phông phú
* Giao tiếp với máy chủ cơ sở dữ liệu bằng cách sử dụng các đối tượng trong mã của chúng.



Hình 2.7- PostgreSQL logo

### Các công nghệ khác:

* Thư viện Tesseract OCR (pytesseract): Đây là công cụ OCR mạnh mẽ hỗ trợ nhiều loại ngôn ngữ bao gồm cả tiếng Việt.
* OpenAI Embedding API (text-embedding-ada-002): Mô hình nhúng giúp chuyển đổi văn bản thành vector.
* FastAPI là 1 web framework hiện đại, hiệu năng cao để xây dựng APIs voiwsw Python.

# PHÂN TÍCH HỆ THỐNG

## Phân tích yêu cầu:

### Yêu cầu chức năng:

“Hệ thống chatbot thông minh hỗ trợ tìm kiếm tài liệu học tập” gồm có các chức năng:

* Tương tác thông minh: Chatbot có khả năng hiểu và phản hồi các câu hỏi của người dùng liên quan đến tài liệu học tập.
* Tìm kiếm tài: Hỗ trợ người dùng tìm kiếm tài liệu dựa trên từ khóa, danh mục hoặc chủ đề.
* Gợi ý tài liệu: Gợi ý tài liệu liên quan dựa trên lịch sử tìm kiếm và sở thích của người dùng.
* Cung cấp thông tin chi tiết: Cung cấp thông tin chi tiết về tài liệu, bao gồm tên, tác giả, mô tả.
* Hỗ trợ nhiều định dạng tài liệu: Có khả năng tìm kiếm tài liệu ở nhiều định dạng khác nhau như PDF, DOCX, hình ảnh

### Chức năng phụ:

* Tạo phiên trò chuyện mới: cho phép người dùng bắt đầu 1 phiên trò chuyện mới với chatbot và có đặt tên cho mỗi phiên trò chuyện.
* Lưu trữ phiên trò chuyện: Lưu lại nội dung của phiên trò truyện trước khi tạo phiên mới và hiển thị danh sách các phiên ở sidebar.
* Chuyển đổi giữa các phiên trò chuyện:P người dùng có thể chuyển đổi qua lại giữa các phiên trò chuyện với chatbot.

## Đặc tả hệ thống:

### Đặc tả chức năng chính:

* Tương tác thông minh:
* Chatbot hiểu và phản hồi các câu hỏi của người dùng liên quan đến tài liệu học tập.
* Cung cấp câu trả lời dựa trên dữ liệu trong hệ thống, không lấy dữ liệu bên ngoài tài liệu được cho.
* Tìm kiếm tài liệu:
* Tìm kiếm tài liệu theo từ khóa, chủ đề, hoặc danh mục.
* Gợi ý tài liệu dựa trên lịch sử tìm kiếm và sở thích cá nhân.
* Cá nhân hóa:
* Lưu trữ lịch sử tìm kiếm và câu hỏi của người dùng để cải thiện trải nghiệm.

### Đặc tả chức năng phụ:

* Tạo phiên trò chuyện mới (Thêm chức năng):
* Cho phép người dùng tạo một phiên trò chuyện mới bất kỳ lúc nào.
* Tự động lưu trữ nội dung của phiên trò chuyện trước đó.
* Đặt tên cho phiên trò chuyện dựa trên câu hỏi đầu tiên hoặc do người dùng nhập.

Tài liệu học tập bao gồm tài liệu học tập trong chương trình giáo dục phổ thông 2018 như sách giáo khoa, tài liệu tham khảo theo định dạng pdf, docx, hình ảnh

## Tiền xử lý dữ liệu:

Để dữ liệu có thể được dùng 1 cách hiệu quả, quá trình xử lí dữ lệu gồm:

* Văn bản từ PDF và DOCX được chuyển đổi thành dạng text để phân tích.
* Hình ảnh được xử lý bằng OCR (Optical Character Recognition) để trích xuất nội dung.

## Tạo vector embeddings:

Vector embeddings là kỹ thuật chính giúp biểu diễn văn bản trong không gian số, phục vụ cho việc tìm kiếm và so sánh thông tin.

* Mô hình sử dụng: Sử dụng OpenAI Embedding API hoặc các mô hình tương tự như SentenceTransformers.
* Quy trình tạo embeddings: Mỗi đoạn văn bản được chuyển đổi thành một vector số có ý nghĩa. Sau đó lưu trữ các vector trong cơ sở dữ liệu vector để dễ dàng tìm kiếm.

## Lưu trữ và tổ chức dữ liệu:

Lưu những dữ liệu văn bản vào trong cơ sở dữ liệu PostgreSQL

Cấu trúc lưu trữ bao gồm:

* Thư mục data/: Chứa các tệp nguồn.
* Cơ sở dữ liệu: Bao gồm bảng documents (thông tin tài liệu) và embeddings (vector)

## Thiết kế hệ thống:

Kiến trúc tổng thể:

Gồm có 3 lớp chính:

1. Lớp giao diện (FrontEnd):

* Chịu trách nhiệm hiển thị và tương tác với người dùng thông qua giao diện trò chuyện.
* Được phát triển bằng ReactJS kết hợp với TailwindCSS để tạo nên giao diện đơn giản, dễ sử dụng nhưng vẫn hiện đại và thân thiện.

1. Lớp xử lý (BackEnd):

* Xử lý các yêu cầu từ người dùng và kết nối với các thành phần khác của hệ thống.
* Được xây dựng bằng FastAPI, đảm bảo hiệu suất cao, linh hoạt và dễ tích hợp.
* Tích hợp các mô hình AI như GPT-4o và LlamaIndex để xử lý truy vấn, tìm kiếm dữ liệu và trả lời câu hỏi.

1. Lớp cơ sở dữ liệu:

* Lưu trữ tài liệu học tập, vector embeddings, và lịch sử tương tác của người dùng.
* Sử dụng PostgreSQL để đảm bảo tính bảo mật, khả năng mở rộng và hiệu quả trong xử lý dữ liệu lớn.

### 3.5.2. Các thành phần chính:

1. Module Xử lý ngôn ngữ tự nhiên (NLP):

* Tích hợp GPT-4o để xử lý và phân tích các truy vấn tự nhiên từ người dùng.
* Sử dụng vector embeddings để tìm kiếm và so sánh nội dung trong cơ sở dữ liệu.

1. Module quản lý dữ liệu:

* Lưu trữ và tổ chức tài liệu học tập theo chủ đề, loại định dạng (PDF, DOCX, hình ảnh, video).
* Cung cấp API để truy cập và cập nhật dữ liệu.

1. Module tìm kiếm thông tin:

* Áp dụng mô hình Retrieval-Augmented Generation (RAG) để truy xuất và kết hợp thông tin từ cơ sở dữ liệu trước khi sinh câu trả lời.
* Tăng cường độ chính xác và liên quan của thông tin được cung cấp.

1. Module cá nhân hóa:

* Theo dõi lịch sử tìm kiếm và tương tác của người dùng.
* Cung cấp gợi ý tài liệu dựa trên sở thích và nhu cầu học tập.

### Luồng hoạt động chính của hệ thống:

* 1. Người dùng gửi câu hỏi hoặc yêu cầu thông qua giao diện chatbot.
  2. Backend nhận yêu cầu, xử lý bằng NLP để hiểu nội dung và ý định.
  3. Hệ thống tìm kiếm thông tin trong cơ sở dữ liệu, kết hợp với GPT-4o để tạo câu trả lời.
  4. Trả lời được gửi lại frontend và hiển thị cho người dùng.

# THIẾT KẾ HỆ THỐNG

## Kiến trúc hệ thống:

### Sơ đồ usecase:

A diagram of a flowchart

Description automatically generated

Hình 4.1. Sơ đồ usecase

## Thiết kế cơ sở dữ liệu:

Sơ đồ cơ sở dữ liệu:

A screenshot of a computer

Description automatically generated

Hình 4.2. Sơ đồ cơ sở dữ liệu

## Thiết kế API: Thiết kế các endpoint API và tương tác với các thành phần khác.

## Tích hợp GPT-4o và LlamaIndex: Mô tả quá trình tích hợp các công nghệ AI.

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

# DANH MỤC TÀI LIỆU THAM KHẢO

1. https://www.coursera.org/articles/what-is-a-chatbot
2. <https://www.ibm.com/think/topics/vector-embedding#:~:text=Vector%20embeddings%20are%20numerical%20representations,(ML)%20models%20can%20process>.
3. <https://www.ibm.com/think/topics/gpt-4o>
4. <https://www.datacamp.com/blog/vector-embedding?utm_source=google&utm_medium=paid_search&utm_campaignid=19589720824&utm_adgroupid=152984013054&utm_device=c&utm_keyword=&utm_matchtype=&utm_network=g&utm_adpostion=&utm_creative=724847714572&utm_targetid=dsa-2222697810678&utm_loc_interest_ms=&utm_loc_physical_ms=9198864&utm_content=DSA~blog~Artificial-Intelligence&utm_campaign=230119_1-sea~dsa~tofu_2-b2c_3-row-p2_4-prc_5-na_6-na_7-le_8-pdsh-go_9-nb-e_10-na_11-na-dec24&gad_source=1&gclid=EAIaIQobChMIlrGE7p7DigMVl8wWBR09pgW5EAAYASAAEgInk_D_BwE>
5. <https://aws.amazon.com/what-is/retrieval-augmented-generation/>
6. https://thuvienphapluat.vn/chinh-sach-phap-luat-moi/vn/ho-tro-phap-luat/tu-van-phap-luat/59386/chuong-trinh-giao-duc-pho-thong-2018-chuong-trinh-tong-the#

# PHỤ LỤC