ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

KHAI THÁC NGỮ LIỆU VĂN BẢN NÂNG CAO

BÁO CÁO ĐỒ ÁN THỰC HÀNH

Xây dựng Chatbot hỏi đáp Du lịch và Ẩm thực Việt Nam

Thành viên:
21120157 - Lê Phạm Hoàng Trung
21120081 - Phạm Thái Huy
21120355 - Nguyễn Anh Tú
21120279 - Lê Trần Minh Khuê

Khoa: Công nghệ Thông tin

Thành phố Hồ Chí Minh - 2025

Lời cảm ơn

Chúng em xin chân thành cảm ơn TS. Nguyễn Trường Sơn, TS. Nguyễn Tiến Huy, những giảng viên đã dày công truyền đạt kiến thức và hướng dẫn chúng em trong quá trình học và thực hiện đồ án này.

Chúng em xin gửi lời cảm ơn đến khoa Công nghệ thông tin, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia TPHCM đã tạo điều kiện cho chúng em học tập và hoàn thành đồ án này.

Chúng em đã cố gắng vận dụng những kiến thức đã học được để hoàn thành đồ án. Nhưng do kiến thức hạn chế và chưa có nhiều kinh nghiệm nên khó tránh khỏi những thiếu sót trong quá trình nghiên cứu và trình bày.

Rất kính mong sự góp ý của quý Thầy để bài báo cáo của chúng em được hoàn thiện hơn.

Chúng em trân trọng cảm ơn sự quan tâm giúp đỡ của các Thầy trong suốt môn học này.

Xin trân trọng cảm ơn!

TP. Hồ Chí Minh, ngày 15 tháng 02 năm 2025

NHÓM SINH VIÊN THỰC HIỆN

Đại diện

Lê Phạm Hoàng Trung

Mục lục

1 Tổng quan đồ án

Trong thời đại số hóa, việc khai thác và truy xuất thông tin từ các nguồn dữ liệu văn bản lớn đóng vai trò quan trọng trong nhiều lĩnh vực, đặc biệt là du lịch. Để giúp người dùng dễ dàng tìm kiếm thông tin về các địa điểm du lịch tại 63 tỉnh thành Việt Nam, đồ án này phát triển một chatbot hỏi đáp sử dụng kỹ thuật Retrieval-Augmented Generation (RAG).

Chatbot có khả năng truy xuất thông tin từ tập dữ liệu chứa thông tin du lịch và tạo ra câu trả lời phù hợp với câu hỏi của người dùng. Để thực nghiệm phương pháp RAG, nhóm đã lựa chọn sử dụng các tài liệu du lịch lưu trữ dưới dạng file PDF, qua đó đánh giá hiệu quả của mô hình trong việc tìm kiếm và tổng hợp thông tin từ dữ liệu không có cấu trúc.

Mục tiêu của đồ án không chỉ là xây dựng một hệ thống hỏi đáp thông minh mà còn là ứng dụng và đánh giá kỹ thuật RAG trong việc khai thác dữ liệu văn bản – một trong những nội dung quan trọng của môn Khai thác dữ liệu văn bản nâng cao.

2 Bộ dữ liệu

2.1 Mô tả

Tập dữ liệu của đồ án bao gồm 63 tập tin PDF, mỗi tập tin tương ứng với một tỉnh thành trên toàn quốc. Mỗi tập tin PDF chứa thông tin chi tiết về 10 địa điểm du lịch và 5 món ăn/đặc sản đặc trưng của tỉnh đó. Dữ liệu được thu thập một cách có hệ thống, phản ánh sự đa dạng và phong phú của văn hóa, di sản và ẩm thực địa phương trên phạm vi cả nước.

2.1.1 Thông tin về địa điểm du lịch

Mỗi địa điểm du lịch được ghi nhận đầy đủ các thông tin sau:

- **Địa chỉ:** Vị trí của địa điểm, giúp người dùng có thể định vị hoặc tìm được đường đi đến địa điểm đó.
- **Giờ mở cửa:** Thông tin về khung giờ hoạt động của địa điểm (nếu có), hỗ trợ việc lập kế hoạch tham quan.
- **Giá vé/Chi phí:** Mức giá vé hoặc chi phí liên quan đến việc trải nghiệm tại địa điểm (nếu có), giúp du khách ước tính chi phí.
- **Phân loại điểm du lịch:** Chỉ định loại hình điểm du lịch (ví dụ: di tích lịch sử, thiên nhiên, văn hóa, giải trí,...) nhằm phân loại và đánh giá đặc trưng của từng địa điểm.
- **Nguồn thông tin:** Ghi rõ nguồn tham khảo để đảm bảo tính xác thực và minh bạch của dữ liệu.
- **Thông tin về địa điểm:** Mô tả chi tiết các đặc điểm nổi bật, lịch sử, văn hóa và các hoạt động liên quan.
- **Các thông tin khác/Cách di chuyển:** Bao gồm các thông tin bổ sung như hướng dẫn di chuyển hoặc những lưu ý đặc biệt khác (nếu có).

2.1.2 Thông tin về Món ăn/Đặc sản

Mỗi món ăn hoặc đặc sản được mô tả với các thông tin sau:

- **Mô tả món ăn:** Trình bày chi tiết về thành phần, hương vị và phương pháp chế biến, qua đó phản ánh nét đặc trưng ẩm thực của tỉnh.
- **Giá tham khảo:** Cung cấp mức giá ước tính (nếu có), giúp người dùng có cái nhìn sơ bộ về chi phí trải nghiệm ẩm thực.
- **Địa chỉ tham khảo:** Thông tin về địa điểm hoặc cơ sở ẩm thực nơi món ăn được phục vụ, hỗ trợ việc xác định vị trí thưởng thức món ăn.

2.1.3 Tính Ứng dụng của Dữ liệu

Cấu trúc dữ liệu đồng nhất giữa các tỉnh thành cho phép hệ thống dễ dàng xử lý, truy xuất và tổng hợp thông tin. Việc phân chia dữ liệu theo các danh mục rõ ràng không chỉ hỗ trợ quá trình tìm kiếm và phân tích mà còn tạo nền tảng vững chắc cho việc ứng dụng các kỹ thuật khai thác dữ liệu hiện đại, đặc biệt là trong khung cảnh triển khai mô hình Retrieval-Augmented Generation (RAG).

Dữ liệu này đóng vai trò quan trọng trong việc kiểm chứng hiệu quả của các phương pháp truy xuất và tổng hợp thông tin, qua đó góp phần nâng cao độ chính xác và khả năng đáp ứng của chatbot trong việc cung cấp thông tin du lịch và ẩm thực cho người dùng.

2.2 Thu Thập Dữ Liệu

Nhóm xây dựng một mini bot nhằm mục tiêu thu thập và xử lý thông tin liên quan đến du lịch và ẩm thực đặc sản của các tỉnh thành Việt Nam. Nhóm hi vọng minibot này sẽ hỗ trợ nhóm tiết kiệm thời gian thu thập thông tin du lịch ở các tỉnh thành. Quy trình thực hiện như sau:

2.2.1 Thu Thập URL Từ Kết Quả Tìm Kiếm Google

- **Xây dựng truy vấn tìm kiếm:** Sử dụng từ khóa kết hợp với tên các tỉnh thành đã được định nghĩa trong một từ điển. (ví dụ: "điểm du lịch" hay "ẩm thực đặc sản" + "Tên tỉnh")
- **Gọi API tìm kiếm:** Thông qua Google Custom Search API, các kết quả tìm kiếm được thu thập về dưới dạng dữ liệu JSON, trong đó chứa tiêu đề và liên kết của các trang web.
- **Lưu trữ kết quả:** Thông tin thu thập được được tổ chức và lưu trữ dưới dạng các tập tin JSON và CSV để phục vụ cho bước xử lý tiếp theo. (tất cả các đường dẫn của các tỉnh được lưu trong cùng 1 tập tin)

2.2.2 Trích Xuất Nội Dung Trang Web

- **Tải nội dung trang web:** Áp dụng kỹ thuật xử lý bất đồng bộ với thư viện `aiohttp` để gửi các yêu cầu HTTP và tải về nội dung HTML từ các URL đã thu thập.
- **Phân tích cú pháp HTML:** Sử dụng thư viện `BeautifulSoup` nhằm trích xuất các nội dung chính (chẳng hạn như các thẻ `h1`, `h2`, `h3`, `p`) từ mã nguồn HTML.
- **Lưu trữ nội dung:** Các nội dung trích xuất được lưu dưới dạng các tập tin văn bản, với tên tập tin được định danh dựa trên tên tỉnh thành và số thứ tự. Mỗi đường dẫn sẽ được lưu về 1 tập tin pdf, có đánh số thứ tự riêng.

2.2.3 Điều Phối Quy Trình và Quản Lý Tác Vụ

- **Xử lý bất đồng bộ:** Sử dụng thư viện `asyncio` kết hợp với `nest_asyncio` để đồng bộ hoá vòng lặp sự kiện, cho phép chạy song song các tác vụ tải về và xử lý nội dung, giúp giảm thời gian chờ.

- **Quản lý tốc độ gửi yêu cầu:** Áp dụng cơ chế trì hoãn giữa các lần gửi yêu cầu để tránh bị máy chủ từ chối dịch vụ do gửi yêu cầu quá nhanh.
- **Xử lý lỗi:** Bao gồm các cơ chế kiểm soát ngoại lệ nhằm đảm bảo quá trình thu thập và xử lý dữ liệu được thực hiện một cách ổn định, ngay cả khi gặp phải lỗi trong quá trình truy cập URL.

2.2.4 Kết quả thu được:

Dữ liệu sau khi tải và lưu trữ hoàn chỉnh sẽ nằm ở các thư mục sau:

```
food
├── texts
├── urls.csv
└── urls.json
```

```
travel
├── texts
├── urls.csv
└── urls.json
```

Trong đó:

- url.csv / url.json: tập tin lưu trữ tất cả đường dẫn đã tìm thấy về các tỉnh thành
- texts: Thư mục chứa các tập tin văn bản đã tải về từ các trang web trong tập tin url.

2.3 Tiền xử lý dữ liệu

Trong các tập tin url.*, dữ liệu mỗi tỉnh thành sẽ được thu từ khoảng 10 đường dẫn. Dữ liệu thô thu thập từ mỗi đường dẫn được lưu thành các tập tin riêng, do đó nhiều địa điểm du lịch và món ăn sẽ bị trùng. Quy trình tiền xử lý sẽ như sau:

1. Làm sạch dữ liệu: xóa thông tin thừa sau khi thu thập từ động (tiêu đề trang web, quảng cáo, tin tức liên quan,...)
2. Gộp dữ liệu: Tổng hợp các tập tin trên thành một tập tin duy nhất cho mỗi tỉnh thành.
3. Lọc dữ liệu trùng: các địa điểm du lịch hoặc món ăn trùng, tất cả sẽ được xem xét và tổng hợp thành một phiên bản đầy đủ nhất cuối cùng, chứa nhiều thông tin nhất theo theo mô tả, phần thông tin thiếu sẽ được bỏ trống.
4. Phân loại dữ liệu: Thông tin trong từng tập tin trên sẽ được phân làm 2 loại: địa điểm và món ăn.

2.4 Chuyển dữ liệu văn bản đã xử lý về dạng tập tin pdf

Các dữ liệu sau khi được tiền xử lý sẽ được chọn lọc (10 địa điểm và 5 món ăn nổi bật nhất) và đưa vào tập tin dữ liệu cuối cùng của tỉnh/thành phố. Tập tin này sẽ là một phần trong bộ dữ liệu du lịch, là nguồn thông tin để chatbot có thể tìm kiếm trong quá trình hỏi đáp.

3 Kiến trúc hệ thống

3.1 Tổng quan kiến trúc

Hệ thống chatbot được xây dựng dựa trên kiến trúc Retrieval-Augmented Generation (RAG). Kiến trúc này kết hợp khả năng truy xuất thông tin từ cơ sở dữ liệu (Retrieval) với khả năng sinh văn bản của mô hình ngôn ngữ lớn (Generation). Cụ thể, khi người dùng đặt câu hỏi, hệ thống sẽ truy xuất các đoạn văn bản liên quan từ cơ sở dữ liệu, sau đó sử dụng các đoạn văn bản này làm ngữ cảnh để tạo ra câu trả lời.

3.2 Lưu đồ hoạt động

Hình 1: Lưu đồ hoạt động của hệ thống

Lưu đồ hoạt động của hệ thống được thể hiện trong Hình 1. Bao gồm:

- User Input: Người dùng nhập câu hỏi vào giao diện.
- Hybrid Retrieval (BM25 + FAISS): Hệ thống sử dụng kết hợp hai phương pháp truy xuất thông tin: BM25Okapi và FAISS. BM25Okapi được sử dụng để tìm kiếm các đoạn văn bản dựa trên từ khóa, trong khi FAISS được sử dụng để tìm kiếm các đoạn văn bản dựa trên ngữ nghĩa.
- Reranking: Kết quả truy xuất từ BM25 và FAISS được kết hợp và sắp xếp lại bằng mô hình Cross-Encoder để chọn ra các đoạn văn bản phù hợp nhất.
- LLM Prompting: Các đoạn văn bản được chọn sẽ được đưa vào prompt cùng với câu hỏi của người dùng. Prompt này được thiết kế theo dạng few-shot learning và chain-of-thought (CoT) để hướng dẫn LLM tạo ra câu trả lời chi tiết và có logic.
- AI Response: Mô hình LLM sẽ tạo ra câu trả lời dựa trên prompt và ngữ cảnh được cung cấp. Câu trả lời này sẽ được hiển thị cho người dùng.

4 Các kỹ thuật cải thiện

4.1 Hybrid Retrieval (BM25 + FAISS)

Trong hệ thống chatbot này, việc truy xuất thông tin hiệu quả là yếu tố then chốt để đảm bảo chatbot có thể cung cấp câu trả lời chính xác và phù hợp. Thay vì chỉ sử dụng một phương pháp duy nhất, chúng em đã kết hợp hai kỹ thuật mạnh mẽ là BM25 và FAISS để tận dụng tối đa ưu điểm của mỗi loại.

- BM25 (Best Matching 25): Đây là một thuật toán tìm kiếm dựa trên từ khóa, được đánh giá cao về khả năng xác định mức độ liên quan giữa truy vấn của người dùng và các tài liệu dựa trên tần suất xuất hiện của từ. BM25 đặc biệt hiệu quả khi người dùng sử dụng các từ khóa rõ ràng và cụ thể.
- FAISS (Facebook AI Similarity Search): Ngược lại, FAISS là một thư viện được thiết kế để tìm kiếm các vector biểu diễn ngữ nghĩa của văn bản một cách nhanh chóng và hiệu quả. FAISS cho phép chatbot tìm kiếm các đoạn văn bản có ý nghĩa tương tự với câu hỏi của người dùng, ngay cả khi không có từ khóa nào khớp chính xác.

Bằng cách kết hợp BM25 và FAISS, hệ thống có thể tận dụng lợi thế của cả hai phương pháp:

- **Độ phủ:** BM25 giúp đảm bảo rằng chatbot không bỏ lỡ bất kỳ tài liệu nào chứa các từ khóa quan trọng trong câu hỏi.
- **Độ chính xác:** FAISS giúp chatbot tìm ra các đoạn văn bản có ý nghĩa phù hợp nhất, ngay cả khi câu hỏi được diễn đạt theo nhiều cách khác nhau.

4.2 Reranking với Cross-Encoder

Sau khi nhận được kết quả từ quá trình truy xuất hỗn hợp (BM25 + FAISS), chúng em sử dụng một mô hình Cross-Encoder để sắp xếp lại các đoạn văn bản theo mức độ liên quan thực sự với câu hỏi.

- Cross-Encoder: Khác với các mô hình bi-encoder (chỉ tạo ra embedding cho câu hỏi và văn bản một cách riêng biệt), Cross-Encoder xem xét cả câu hỏi và văn bản cùng một lúc để đánh giá mức độ liên quan. Điều này giúp Cross-Encoder nắm bắt được các mối quan hệ phức tạp giữa câu hỏi và văn bản, từ đó đưa ra đánh giá chính xác hơn.
- Việc sử dụng Cross-Encoder để reranking giúp hệ thống chọn ra các đoạn văn bản phù hợp nhất để đưa vào prompt cho LLM, từ đó cải thiện đáng kể chất lượng của câu trả lời.

4.3 Prompting

Prompt engineering là một yếu tố quan trọng để khai thác tối đa khả năng của LLM. Trong đồ án này, nhóm chúng em khảo sát các kỹ thuật prompting bao gồm:

- One-shot learning kết hợp cùng Chain-of-Thought:

```
1    prompt_template = """
2    ### Instruction:
3    You are an AI assistant. Provide a detailed answer based on the given
     contexts.
4    Use structured information and reasoning to generate a complete
     response.
5
```

```
    ### Contexts:
    {context}

    ### Question:
    {question}

    ### Example Response:
    **Q:** What is the best time to visit HaLong Bay?
    **A:** The best time to visit HaLong Bay is from **October to April**
    when the weather is cool and dry. Avoid June to August due to typhoons.

    ### Answer:
    """
```

- Few-shot learning kết hợp cùng Chain-of-Thought:

```
    prompt_template = """
    ### Instruction:
    You are an AI assistant. Provide a detailed answer based on the given
    contexts.
    Use structured information and reasoning to generate a complete
    response.

    ### Contexts:
    {context}

    ### Question:
    {question}

    ### Example Response:

    **Q:** What is the best time to visit HaLong Bay?
    **A:** The best time to visit HaLong Bay is from **October to April**
    when the weather is cool and dry. Avoid June to August due to typhoons.
    During this period, you can expect pleasant temperatures, less rainfall
    , and calmer seas, making it ideal for cruising and exploring the bay.

    **Q:** What are some must-try dishes in Hanoi?
    **A:** Hanoi boasts a rich culinary heritage. Some must-try dishes
    include **Pho** (noodle soup), **Bun Cha** (vermicelli noodles with
    grilled pork), **Cha Ca** (turmeric fish with dill), and **Banh Mi** (
    Vietnamese sandwich). Each dish offers a unique flavor profile and
    represents a different aspect of Hanoi's street food culture. Don't
    forget to try the local coffee as well!

    **Q:** Where can I find the best beaches in Vietnam?
```

```
21      **A:** Vietnam has a long coastline with numerous beautiful beaches.
        Some of the most popular destinations include **Phu Quoc** (known for
        its pristine beaches and clear waters), **Nha Trang** (offering a
        vibrant beach city atmosphere), **Da Nang** (with its stunning coastline
         and modern infrastructure), and **Mui Ne** (famous for its sand dunes
        and windsurfing).  The "best" beach depends on your preferences -
        whether you're looking for relaxation, water sports, or a lively
        atmosphere.
22
23      **Q:** What are some popular tourist attractions in Hoi An?
24      **A:** Hoi An is a charming ancient town with many attractions.  Key
        sights include the **Japanese Covered Bridge**, **Fujian Assembly Hall
        **, **Tan Ky Old House**, and the **Central Market**.  Exploring the
        narrow streets lined with tailor shops and enjoying the colorful
        lanterns at night are also essential Hoi An experiences.  Consider
        taking a cooking class to learn about local cuisine.
25
26      **Q:**  How do I travel from Hanoi to Sapa?
27      **A:** There are several ways to travel from Hanoi to Sapa.  The most
        common options are by **overnight train** (takes about 8-9 hours), **bus
        ** (takes about 5-6 hours), or **private car** (takes about 4-5 hours).
         The train offers a more comfortable experience, while the bus is a more
         budget-friendly option. A private car provides the most flexibility.
        Booking in advance is highly recommended, especially during peak season.
28
29
30      ### Answer:
31      """
32
```

- One-shot learning kết hợp cùng Tree-of-Thought:

```
1       prompt_template = """
2       ### Instructions:
3       You are an AI assistant. Use *Tree of Thought (ToT)* reasoning to
        analyze multiple perspectives before generating a complete response.
4       Each thought branch should:
5       - Identify a unique approach to answering the question.
6       - Reason step by step.
7       - Evaluate logical consistency.
8       - Combine the best insights to form a well-structured response.
9       - Translate the answer into the language of the question.
10
11      ### Context:
12      {context}
13
14      ### Question:
```

```
    {question}

    ### Example Tree of Thought:

    *Q:* How to plan a travel itinerary and enjoy local cuisine in Hanoi?

    * - Based on historical data & traveler reviews:*
    A typical 3-day itinerary in Hanoi includes exploring the Old Quarter, Hoan Kiem Lake, and famous landmarks such as the Ho Chi Minh Mausoleum. Visitors should also experience traditional water puppet shows and take a cyclo tour.

    * - Analyzing local food specialties:*
    Hanoi is famous for dishes like *Pho, Bun Cha, and Egg Coffee*. A food tour covering local street vendors and hidden gems is highly recommended for an authentic experience.

    * - Considering budget & travel season:*
    The best time to visit is *autumn (September November) and spring (March April)* when the weather is cool. Travelers on a budget can explore street food stalls and local homestays to optimize costs.

    *A:*
    For a complete Hanoi travel experience, *explore historical sites*, *enjoy street food tours*, and *visit in autumn or spring* for the best weather.

    ---

    ### Context:
    {context}

    ### Question:
    {question}

    ### Answer:
    ---
    """
```

- Few-shot learning kết hợp cùng Tree-of-Thought:

```
    prompt_template = """
    ### Instructions:
    You are an AI assistant. Use *Tree of Thought (ToT)* reasoning to analyze multiple perspectives before generating a complete response.
    Each thought branch should:
```

```
- Identify a unique approach to answering the question.
- Reason step by step.
- Evaluate logical consistency.
- Combine the best insights to form a well-structured response.
- Translate the answer into the language of the question.

### Context:
{context}

### Question:
{question}

### Example Tree of Thought:

*Q:* How to plan a travel itinerary and enjoy local cuisine in Hanoi?

* - Based on historical data & traveler reviews:*
A typical 3-day itinerary in Hanoi includes exploring the Old Quarter, Hoan Kiem Lake, and famous landmarks such as the Ho Chi Minh Mausoleum. Visitors should also experience traditional water puppet shows and take a cyclo tour.

* - Analyzing local food specialties:*
Hanoi is famous for dishes like *Pho, Bun Cha, and Egg Coffee*. A food tour covering local street vendors and hidden gems is highly recommended for an authentic experience.

* - Considering budget & travel season:*
The best time to visit is *autumn (September November) and spring (March April)* when the weather is cool. Travelers on a budget can explore street food stalls and local homestays to optimize costs.

*A:*
For a complete Hanoi travel experience, *explore historical sites*, *enjoy street food tours*, and *visit in autumn or spring* for the best weather.

---

*Q:* What are the most interesting attractions to visit in Hai Phong?

* - Based on popular tourist destinations:*
Hai Phong is known for *Do Son Beach, Cat Ba Island, and Lan Ha Bay*, offering beautiful coastal scenery and various water activities.

* - Analyzing cultural and historical significance:*
```

```
 Historical sites such as *Trang Kenh relic site, Hang Kenh Communal House, and Du Hang Pagoda* provide insight into the city's rich heritage.

 * - Considering travel experience & adventure:*
 Cat Ba National Park is perfect for nature lovers, while Do Son Casino attracts visitors looking for entertainment. The *Buffalo Fighting Festival in Do Son (September)* is a must-see cultural event.

 *A:*
 Hai Phongs top attractions include *Cat Ba Island, Do Son Beach, and Trang Kenh relic site*. For adventure seekers, Cat Ba National Park is ideal.

 ---

 *Q:* What are the must-try specialty dishes in Da Nang?

 *- Researching the city's culinary highlights:*
 Da Nang is famous for *Mi Quang (turmeric-infused noodles with shrimp and pork), Bun Cha Ca (fish cake noodle soup), and Banh Xeo (crispy Vietnamese pancakes)*.

 *- Understanding local dining culture:*
 Street food stalls and traditional restaurants offer the best authentic flavors. *Han Market and Con Market* are great spots to try multiple dishes at affordable prices.

 *- Considering seasonal specialties:*
 Seafood is a must-try in Da Nang, with fresh catches like *grilled stingray, squid, and clams* available throughout the year.

 *A:*
 The must-try dishes in Da Nang include *Mi Quang, Bun Cha Ca, and Banh Xeo*. For an authentic experience, visit local markets and seafood restaurants.

 ---

 *Q:* When is the best time to travel to Can Tho?

 * - Considering customer reviews:*
 Can Tho is famous for the Cai Rang floating market, fruit orchards, and its waterways. The peak fruit season is from June to August.

 * - Weather conditions:*
```

```
    Can Tho has two seasons: the rainy season (May to November) and the dry
    season (December to April). The dry season is suitable for easy travel
    and visiting the floating markets.

    * - Economic factors & experiences:*
    Summer (June to August) offers plenty of fruits and lively activities,
    but it is also the peak tourist season. If you want to avoid crowds,
    November to December is a good choice.

    *A:*
    The ideal time to travel to Can Tho is *December to April* to enjoy the
    dry weather and ease of movement. If you want to experience the fruit
    season, you can visit in *June to August*.

    ---

    ### Answer:
    ---
    """

```

Nhược điểm của việc promt bằng tiếng Anh là câu trả lời đôi khi sẽ trộn lẫn tiếng Anh vào đó. Chúng em đã thử hướng tiếp cận viết promt hướng dẫn bằng tiếng Việt nhưng việc này chỉ khiến mô hình hoạt động tệ hơn. Mô hình chỉ tạo sinh ra được các câu trả lời gồm các ký tự chữ cái ngẫu nhiên thay vì từng từ hoàn chỉnh.

4.4 Tối ưu mô hình LLM

Để đảm bảo chatbot hoạt động mượt mà trên các thiết bị có tài nguyên hạn chế, nhóm chúng em đã tối ưu mô hình Llama 7B bằng kỹ thuật quantization 4-bit (BitsAndBytesConfig).

Quantization là một kỹ thuật giảm độ chính xác của các tham số trong mô hình, từ đó giảm kích thước của mô hình và yêu cầu về bộ nhớ. Quantization 4-bit cho phép giảm kích thước mô hình xuống còn một phần tư mà vẫn giữ được hiệu năng tương đối. Việc tối ưu mô hình LLM giúp chatbot có thể chạy hiệu quả trên các thiết bị có cấu hình không quá mạnh, đồng thời giảm thời gian xử lý và tăng tốc độ phản hồi của chatbot.

5 Quy trình thực thi

- **Tải tài liệu:** Tất cả các file PDF chứa thông tin về du lịch và ẩm thực Việt Nam được lưu trữ trong thư mục **data**. Hệ thống sẽ tự động tải và xử lý các file này.
- **Tạo kho dữ liệu FAISS & BM25:** Văn bản từ các file PDF được chia thành các đoạn nhỏ (chunk) bằng RecursiveCharacterTextSplitter. Sau đó, các đoạn văn bản này được nhúng (embedding)

bằng mô hình SentenceTransformer và được lưu trữ trong kho dữ liệu FAISS. Đồng thời, chỉ mục BM25 cũng được xây dựng trên các đoạn văn bản này.

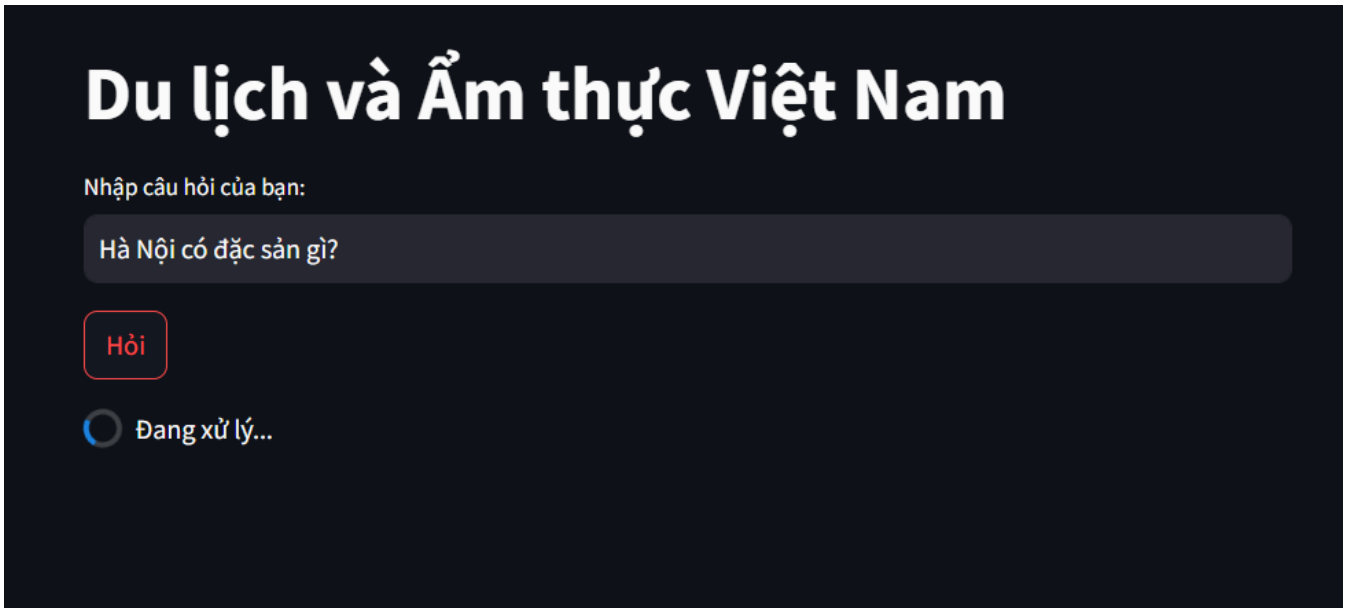
- **Truy xuất tài liệu:** Khi người dùng đặt câu hỏi, hệ thống sẽ sử dụng kết hợp BM25 và FAISS để truy xuất các đoạn văn bản liên quan. Kết quả truy xuất được sắp xếp lại bằng mô hình Cross-Encoder.
- **Sinh câu trả lời từ LLM:** Các đoạn văn bản được truy xuất sẽ được đưa vào prompt cùng với câu hỏi của người dùng. Mô hình LLM sẽ tạo ra câu trả lời dựa trên prompt này.

6 Giao diện

- Giao diện ban đầu:

Hình 2: Giao diện ban đầu

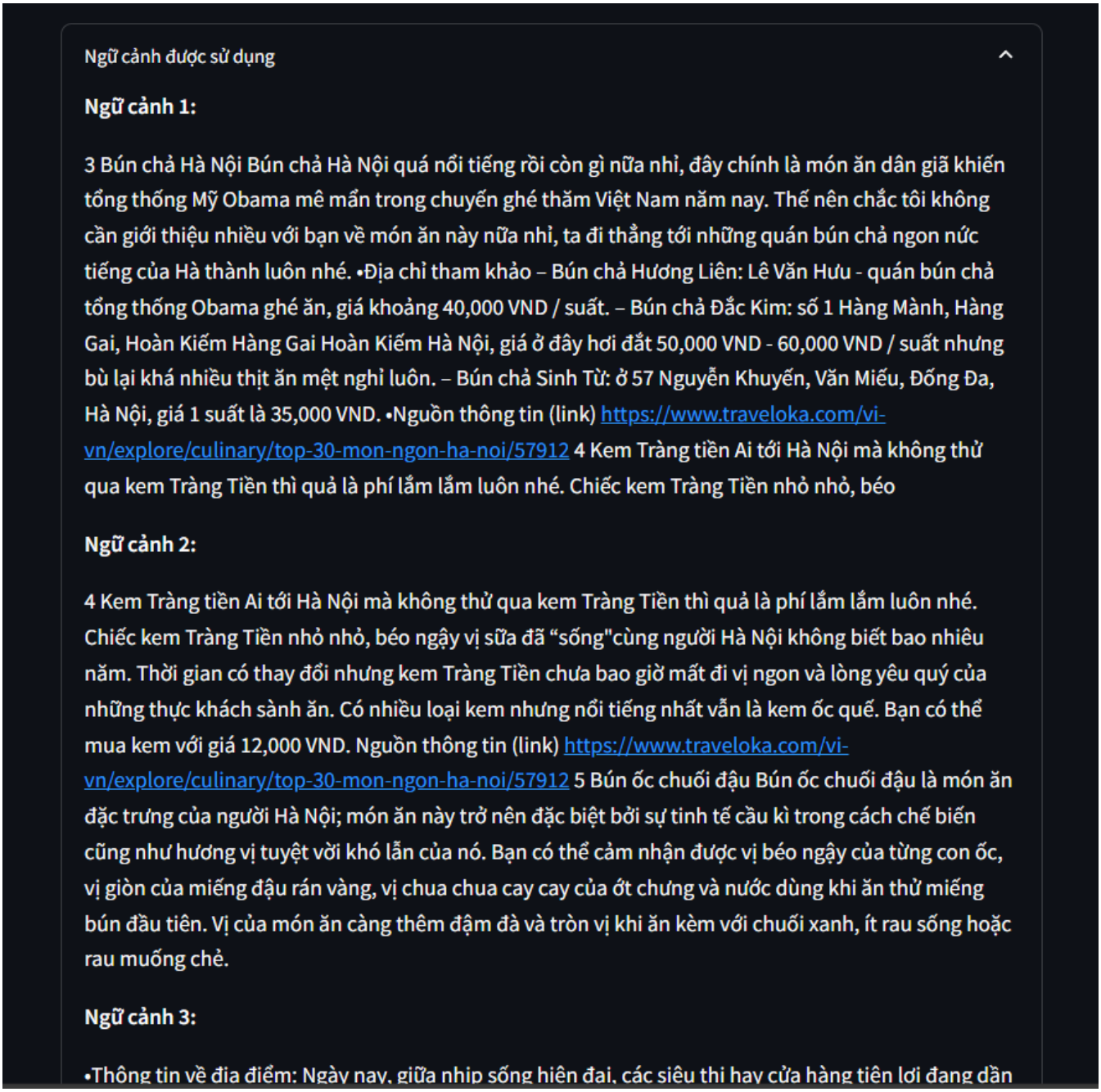
- Người dùng nhập câu hỏi:



Hình 3: Khi người dùng nhập câu hỏi

- Câu trả lời:
- Các thông tin ngữ cảnh được sử dụng:

Hình 4: Câu trả lời



Hình 5: Ngữ cảnh được sử dụng

7 Demo

Phần chạy thực tế mô hình trên giao diện của streamlit được chúng em trình bày trong clip.

Bên cạnh những trường hợp tạo sinh câu trả lời tốt thì cũng còn đó là những câu trả lời kém chất lượng do quá trình tìm kiếm thông tin kém hiệu quả. Điều này có thể do quá trình chunking kém hiệu quả, dữ liệu gây nhầm lẫn (vd: Bắc Ninh và Quảng Ninh có cùng chữ Ninh trong tên tỉnh nên sai), cách embedding và tokenize dữ liệu chưa hợp lý hoặc do mô hình không đủ mạnh do đã quantize. Chúng em không dám chắc đâu là nguyên nhân chính dẫn đến hiện tượng này.

Hình 6: Một ví dụ về hiện tượng gen câu trả lời kém chất lượng

8 Một số ý tưởng cải tiến và kết luận

Ngoài những nổ lực tối ưu quá trình promting để cải thiện thêm chất lượng câu trả lời thì nhóm em cũng có tham khảo qua một số giải pháp nhưng vẫn chưa thực thi được hết đó là:

- Pre - processing questions. Ý tưởng của nhóm là chuẩn hóa các câu hỏi đầu vào lại theo một dạng đơn giản hơn, thống nhất rằng câu hỏi chỉ gồm 2 thông tin là người dùng hỏi về du lịch hay ẩm thực và tỉnh thành nào được hỏi trước khi cho vào mô hình ngôn ngữ. Tuy nhiên trong quá trình thực hiện đồ án, chúng em xếp độ ưu tiên của giải pháp này sau việc cải thiện promting và quản lý thời gian không hiệu quả dẫn đến chưa thực hiện được ý tưởng này.

- Post - processing answers. Ý tưởng của nhóm là chuẩn hóa lại các câu trả lời sau khi mô hình trả ra để có format được đẹp và đồng nhất hơn thay vì có định dạng ngẫu nhiên như trong demo. Chúng em thử thực hiện ý tưởng này bằng hướng dẫn mô hình trả lời ra các câu hỏi có phong cách tương tự nhau nhưng lại không hiệu quả. Chúng em dự định dùng thêm một model để chuẩn hóa lại câu trả lời của mô hình chính. Chúng em xếp mức độ ưu tiên thực hiện ý tưởng này sau cả ý tưởng tiền xử lý câu hỏi trên dẫn đến việc ý tưởng này cũng chưa được thực hiện thử nghiệm.

Tóm lại, nhóm chúng em đã cố gắng cải thiện chất lượng dữ liệu cho mô hình, tối ưu quá trình promting để nâng cao hiệu suất câu trả lời tốt được đưa ra cũng như tìm cách chạy mô hình trên những phần cứng giới hạn có thể tiếp cận được. Kết quả còn chưa thật sự như ý. Vẫn còn đó những ý tưởng cải tiến chưa thể thực hiện hết. Nhưng qua đồ án này nhóm đã có thêm kinh nghiệm và kỹ năng để tối ưu cho những dự án thực tế dựa trên giải pháp GAR này.

Tài liệu

[1] Xuezhi Wang et al. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. 2023. arXiv: 2203.11171 [cs.CL]. URL: https://arxiv.org/abs/2203.11171.