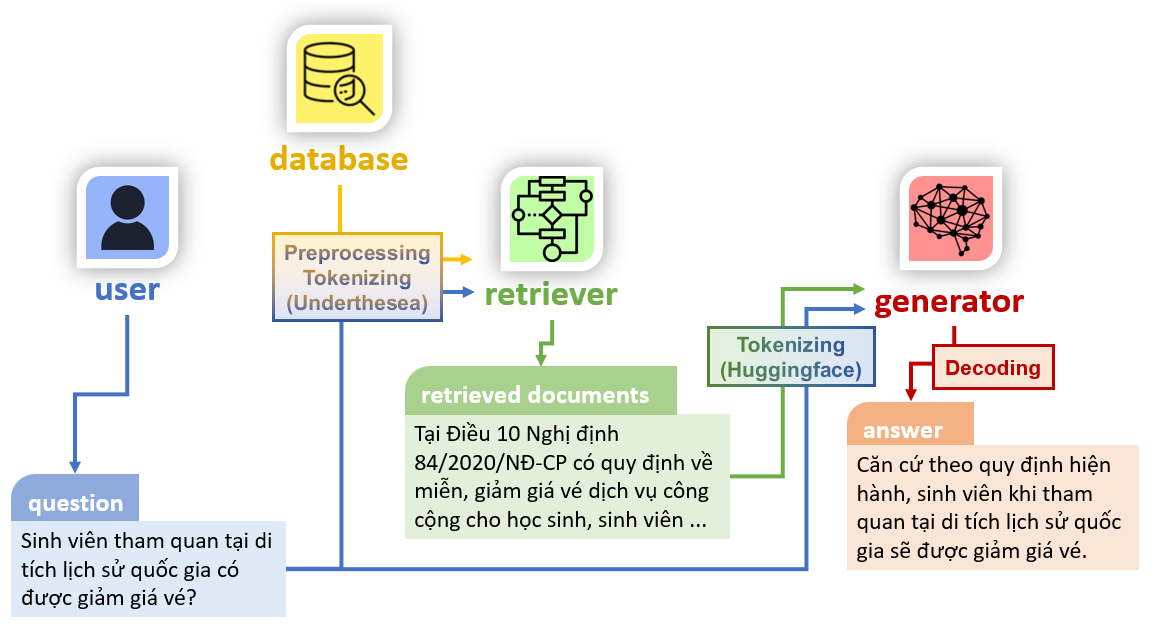
**BÁO CÁO TIỀN DỰ ÁN**

|  |  |
| --- | --- |
| **Thành viên:**  Phan Trung Thuận  Nguyễn Quang Tùng  Trần Trăng Sáng  Ung Khánh Như |  |

# I. Bài toán:

Xây dựng hệ thống hỏi đáp (question-answering system) về luật giáo dục đại học Việt Nam sử dụng kiến trúc retrieval augmentation generation (RAG).

# II. Thiết kế giải pháp:



Hình 1: Kiến trúc tổng quát của hệ thống

Mô tả hệ thống: khi người dùng nhập vào câu hỏi, hệ thống sẽ xử lý câu hỏi đó và tìm tài liệu phù hợp, dựa vào tài liệu đó, hệ thống sẽ sinh ra câu trả lời phù hợp với câu hỏi.

|  |  |  |
| --- | --- | --- |
|  | Vai trò | Mục tiêu/Yêu cầu |
| User | Nhập vào câu hỏi tiếng Việt liên quan đến các vấn đề giáo dục đại học | Nhắm đến nhóm người dùng là sinh viên, giảng viên đại học và cao đẳng trên toàn quốc có nhu cầu tìm hiểu về luật giáo dục đại học |
| Question | Là lời truy vấn để hệ thống có thể tìm kiếm tài liệu và sinh câu trả lời phù hợp | Câu hỏi rõ ràng, liên quan đến luật giáo dục đại học, không được lang mang, khó hiểu |
| Preprocessing | Xử lý các từ ngữ dư thừa, dấu câu để hỗ trợ Retriever tìm kiếm tài liệu chính xác hơn |  |
| Tokenizing | Tách các từ ra thành các token để giảm độ dài sequence, tăng hiệu suất tính toán của mô hình | Độ dài sequence tối đa là 4000 tokens |
| Retriever | Hệ thống thuật toán tìm kiếm tài liệu phù hợp từ bộ dữ liệu dựa trên câu hỏi được đặt ra | Cài đặt TFIDF với độ chính xác Acc@3 đạt 50% |
| Retrieved documents | Tài liệu đã được Retriever tìm thấy và phù hợp với câu hỏi | Chọn tối đa 3 tài liệu từ bộ dữ liệu để hỗ trợ trả lời câu hỏi |
| Generator | Mô hình ngôn ngữ cỡ vừa | Scale down pretrain model để phù hợp với lượng tài nguyên giới hạn. BLEU đạt 0.5 |
| Decoding | Chuyển các token được tạo ra từ Generator thành câu trả lời |  |
| Answer | Câu trả lời hoàn chỉnh cho câu hỏi |  |

Bảng 1: Mô tả chi tiết mục tiêu và vai trò của các thành phần

# III. Xây dựng kế hoạch:

Kế hoạch chi tiết được cả nhóm đồng thuận như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Nội dung công việc | Thời gian thực hiện | Tiến độ |
| 1 | Chuẩn bị dữ liệu |  |  |
| 1.1 | - Crawl dữ liệu từ trang web thuvienphapluat.vn | 25/8 – 26/8 | Đạt tiến độ |
| 1.2 | - Xử lý và chọn lọc dữ liệu thô | 27/8 – 28/8 | Sớm tiến độ (đã hoàn thành trong ngày 27/8) |
| 2 | Xây dựng và đánh giá Retriever |  |  |
| 2.1 | - Xây dựng | 29/8 – 30/8 | Đạt tiến độ |
| 2.2 | - Đánh giá | 30/8 – 31/8 | Đạt tiến độ |
| 3 | Xây dựng, huấn luyện và đánh giá Generator |  |  |
| 3.1 | - Xây dựng | 1/9 – 2/9 | Trễ tiến độ (đến hết ngày 3/9 mới hoàn tất) |
| 3.2 | - Huấn luyện | 2/9 – 10/9 | Trễ tiến độ (đến hết ngày 11/9 mới hoàn tất) |
| 3.3 | - Đánh giá | 11/9 – 12/9 | Trễ tiến độ |
| 4 | Kết hợp lại thành một hệ thống hoàn chỉnh |  |  |
| 4.1 | - Kết hợp | 12/9 – 13/9 | Đạt tiến độ |
| 4.2 | - Demo | 13/9 – 14/9 | Đạt tiến độ |
| 5 | Viết báo cáo và trình bày | 14/9 – 15/9 | Trễ tiến độ |
| 6 | Cải tiến (nếu còn thời gian) | Không có |  |

Bảng 2: Chi tiết quá trình thực hiện dự án

Rủi ro:

Về kỹ thuật: với lượng kiến thức chưa hoàn chỉnh và tài nguyên hạn hẹp, yêu cầu đặt ra cho hệ thống có thể không đáp ứng được về kích thước mô hình, độ lớn của bộ dữ liệu, thời gian huấn luyện đặc biệt về mặt hiệu suất.

Về con người: các thành viên trong nhóm phải tận tâm, kiên trì, ham học hỏi và phải dành rất nhiều thời gian cho dự án mới có thể đạt được hiệu quả cao.

# IV. Một số nguyên tắc làm việc nhóm:

* Tất cả các thành viên đều tham gia đóng góp trong mọi giai đoạn phát triển của dự án.
* Tất cả các thành viên đều liên lạc thường xuyên là zalo nhóm.
* Nếu không thể hoàn thành được công việc cần chuyển giao cho thành viên khác trước deadline ít nhất 1 ngày.
* Code cần phải rõ ràng, dễ đọc, dễ hiểu.
* Ưu tiên sử dụng những thư viện có sẵn.