THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/lFfXpxQkwzQ*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/khoatdds/CS2205.CH183/blob/main/CS2205_CH183_Slide.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Trần Đăng Khoa
- MSSV: 230101009

- Lớp: CS2205.CH183
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 3
- Số câu hỏi QT của cả nhóm: 3
- Link Github: https://github.com/mynameuit/CS2205.CH183/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

HỆ THỐNG HỎI ĐÁP PHÁP LUẬT VIỆT NAM

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

VIETNAM LEGAL QA SYSTEM

TÓM TẮT

Đề tài "Hệ thống hỏi đáp Pháp luật Việt Nam" hướng đến việc xây dựng một công cụ thông minh, hỗ trợ người dùng tra cứu và nắm bắt các quy định pháp luật một cách nhanh chóng, chính xác và minh bạch. Nền tảng của hệ thống là phương pháp RAG (Retrieval Augmented Generation) [1], kết hợp ưu điểm của hệ thống truy xuất thông tin và mô hình tạo sinh văn bản dựa trên deep learning. Nhờ đó, hệ thống không chỉ truy xuất được các văn bản pháp luật đa dạng như luật, nghị định, thông tư mà còn tạo ra các câu trả lời phù hợp với ngữ cảnh của truy vấn, đồng thời kèm theo những dẫn chứng.

Trong quá trình phát triển, đề tài sẽ tập trung xây dựng một bộ dữ liệu chuyên ngành pháp luật Việt Nam được làm mới và tinh gọn, cùng với tiêu chuẩn đánh giá hiệu suất (benchmark) cụ thể. Điều này giúp hệ thống đảm bảo tính đầy đủ, cập nhật và chính xác của các thông tin được cung cấp. Ngoài ra, đề tài cũng nghiên cứu và triển khai các kỹ thuật tối ưu nhằm tự động cập nhật dữ liệu pháp lý từ nhiều nguồn khác nhau, từ đó nâng cao hiệu quả và khả năng mở rộng của hệ thống.

Hệ thống hướng đến việc phục vụ đa dạng đối tượng người dùng, bao gồm cá nhân, doanh nghiệp và các chuyên gia pháp lý, với mục tiêu giảm bớt khó khăn trong việc tìm kiếm và hiểu rõ các quy định pháp luật. Thông qua việc đánh giá lại và sắp xếp các kết quả trả về dựa trên mức độ liên quan, hệ thống giúp người dùng dễ dàng tiếp cận những thông tin pháp lý phù hợp nhất với nhu cầu của họ.

Tóm lại, đề tài "Hệ thống hỏi đáp Pháp luật Việt Nam" không chỉ mang lại giá trị ứng dụng cao trong việc hỗ trợ tra cứu thông tin pháp lý mà còn đóng góp vào sự phát triển của công nghệ trí tuệ nhân tạo trong lĩnh vực pháp luật. Với mục tiêu cung cấp một giải pháp thông minh, linh hoạt và hiệu quả, đề tài hứa hẹn sẽ tạo ra bước tiến mới trong việc ứng dụng AI vào quản lý và phân tích thông tin pháp lý, góp phần nâng cao trải nghiệm người dùng trong thời đại số.

GIỚI THIỆU

Trong bối cảnh chuyển đổi số và sự phát triển nhanh chóng của công nghệ trí tuệ nhân tạo, việc tra cứu và áp dụng các quy định pháp luật đang trở nên ngày càng quan trọng đối với cá nhân, doanh nghiệp và tổ chức pháp lý. Các bộ luật hiện hành của Việt Nam liên tục được cập nhật để phản ánh các yêu cầu mới của xã hội, nhưng đồng thời cũng gây ra nhiều khó khăn cho người dùng do nguồn dữ liệu phân tán, thông tin phức tạp và khó đồng bộ. Chính vì vậy, đề tài "Hệ thống hỏi đáp Pháp luật Việt Nam" ra đời nhằm cung cấp một giải pháp thông minh, giúp người dùng nhanh chóng truy xuất, phân tích và nắm bắt các quy định pháp luật một cách chính xác và hiệu quả.

Hệ thống hoạt động theo cơ chế kết hợp giữa kỹ thuật truy xuất thông tin (Retrieval) và mô hình tạo sinh văn bản (Generation) – hay còn gọi là phương pháp RAG (Retrieval Augmented Generation). Cụ thể, đầu vào (input) của hệ thống là các câu hỏi hoặc từ khóa liên quan đến pháp luật do người dùng nhập vào. Sau đó, hệ thống sẽ truy xuất các văn bản pháp lý từ kho dữ liệu gồm các văn bản luật, nghị định, thông tư, quyết định, v.v. và xử lý thông tin qua mô hình tạo sinh văn bản để đưa ra câu trả lời chi tiết, phù hợp với ngữ cảnh và yêu cầu của truy vấn. Đầu ra (output) là các câu trả lời văn bản hoàn chỉnh, kèm theo các thông tin tham khảo cần thiết, giúp người dùng có được cái nhìn toàn diện về vấn đề pháp lý đang quan tâm. Hệ thống đề xuất sẽ khắc phục được những hạn chế hiện tại của các hệ thống hiện đã có như tốc độ, hiệu suất, vấn đề ảo giác (hallucination) của các mô hình ngôn ngữ lớn.

MỤC TIÊU

Đề tài "Hệ thống hỏi đáp Pháp luật Việt Nam" hướng tới xây dựng một bộ dữ liệu pháp luật Việt Nam đầy đủ và hoàn chỉnh. Từ đó phát triển một mô hình truy xuất và tạo sinh thông minh thông qua việc kết hợp phương pháp RAG với mô hình ngôn ngữ lớn (LLM). Thông qua việc thực hiện các mục tiêu trên, đề tài nhằm tạo ra một công cụ hỗ trợ tra cứu pháp luật tiên tiến, góp phần nâng cao hiệu quả tiếp cận và sử dụng thông tin pháp lý, từ đó hỗ trợ quyết định và hoạt động của các cá nhân, doanh nghiệp và tổ chức pháp lý trong thời đại số.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. **Nội dung:**

Dựa vào các mục tiêu đã đặt ra, chúng tôi sẽ tiến hành thực hiện theo các nội dung sau:

- Xây dựng bộ dữ liệu pháp luật Việt Nam:
  - Tìm hiểu các nguồn dữ liệu pháp lý uy tín của Việt Nam như https://thuvienphapluat.vn/,..
  - Thu thập và làm sạch, chuẩn hóa dữ liệu.
- Phát triển hệ thống:
  - RAG - Retrieval Augmented Generation
  - Mô hình Embedding
  - Mô hình ReRanking
  - Mô hình tạo sinh văn bản
- Đánh giá mô hình: sử dụng các độ đo phù hợp để đánh giá.

2. **Phương pháp:**

Ứng với từng nội dung, chúng tôi đã đánh giá, khảo sát để lựa chọn phương pháp thực hiện phù hợp:

- RAG - Retrieval Augmented Generation là một kỹ thuật giúp nâng cao khả năng của mô hình sinh (language model generation) kết hợp với tri thức bên ngoài (external knowledge). Phương pháp này thực hiện bằng

cách truy xuất thông tin liên quan từ kho tài liệu (tri thức) và sử dụng chúng để tăng cường dữ liệu đầu vào cho quá trình sinh câu trả lời dựa trên LLM. Từ đó sẽ hạn chế được vấn đề ảo giác của LLM (hallucination), tiết kiệm chi chí và thời gian huấn luyện.

- Đối với xây dựng bộ dữ liệu pháp luật Việt Nam: chúng tôi lựa chọn https://thuvienphapluat.vn/ là nguồn dữ liệu uy tín để sử dụng. Dữ liệu thu thập sẽ bao gồm các văn bản pháp luật, nghị định, thông tư, quyết định, hiến pháp,..và các cặp câu hỏi - câu trả lời từ phần hỏi đáp pháp luật của trang web.
- Đối với xây dựng mô hình: chúng tôi sử dụng:
  - Mô hình Embedding là BGGE-M3 [2], đây là một mô hình embedding đa ngôn ngữ tiên tiến, được thiết kế đặc biệt để hỗ trợ cả tiếng Việt và các ngôn ngữ khác.
  - Mô hình ReRanking là BGE-Reranker-V2-M3 [3], một mô hình ngôn ngữ tiên tiến được phát triển nhằm tối ưu hóa quá trình xếp hạng lại các kết quả tìm kiếm.
  - Mô hình tạo sinh văn bản là Vistral-7B-chat [4], một mô hình ngôn ngữ lớn tiên tiến, được phát triển từ mô hình Mistral 7B thông qua quy trình tiếp tục tiền huấn luyện và tinh chỉnh theo hướng dẫn với dữ liệu tiếng Việt đa dạng.
- Về đánh giá hiệu suất mô hình: chúng tôi đánh giá trên 2 tiêu chí là khả năng truy vấn sử dụng độ đo Recall@k1,3,5 và khả năng trả lời câu hỏi sử dụng 2 độ đo là Answer Similarity và BLUE Score.

KẾT QUẢ MONG ĐỢI

1. Xây dựng được một bộ dữ liệu pháp luật Việt Nam hoàn chỉnh và đầy đủ nhất.
2. Xây dựng được một hệ thống hỏi đáp pháp luật Việt Nam, câu trả lời sẽ được dẫn chứng cụ thể, rõ ràng.

TÀI LIỆU THAM KHẢO

[1]. P. S. H. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks,” Neural Information Processing Systems, vol. 33, pp. 9459–9474, May 2020, [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[2]. Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Findings of the Association for Computational Linguistics: ACL 2024, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

[3]. Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. arXiv preprint arXiv:2312.15503.

[4]. Van Nguyen, C., Nguyen, T., Nguyen, Q., Nguyen, H., Plüster, B., Pham, N., ... & Nguyen, T. (2023). Vistral-7b-chat-towards a state-of-the-art large language model for vietnamese.