BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS519 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS519.Q11

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

LEGAL DOCUMENT RETRIEVAL USING SEMANTIC EMBEDDINGS

Trần Vinh Khánh - 23520726

Tóm tắt

- Link Github của nhóm: https://github.com/bin9639/CS519.Q11.KHT
- Link YouTube video: https://youtu.be/ZVKQI5abT4M

Trần Vinh Khánh - 23520726

Giới thiệu

Bối cảnh

- Quá trình số hóa làm số lượng văn bản pháp luật điện tử (luật, nghị định, thông tư, ...) tăng nhanh.
- Người dùng có thể tiếp cận nhiều nguồn pháp lý, nhưng việc tra cứu chính xác nội dung cần tìm vẫn gặp nhiều khó khăn.

Hạn chế

- Chủ yếu dựa trên tìm kiếm từ khóa.
- Người dùng phải đọc và đối chiếu nhiều văn bản khác nhau.
- Chỉ trả về các văn bản có chứa từ khóa, không chỉ ra đúng điều khoản hoặc đoạn nội dung liên quan.

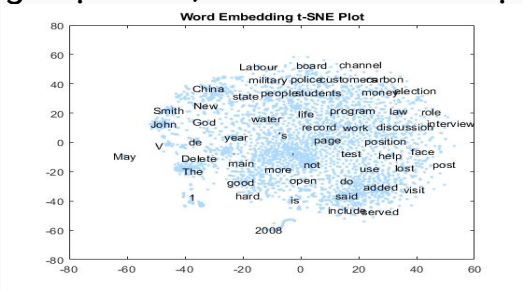
Giới thiệu

Vấn đề đặt ra

- Truy vấn của người dùng thường ở dạng ngôn ngữ tự nhiên.
- Cách diễn đạt trong truy vấn khác biệt với ngôn ngữ chuẩn mực, chặt chẽ của văn bản pháp luật.
- Kho văn bản lớn, dài và có cấu trúc phức tạp.

Bài toán nghiên cứu

- Truy vấn văn bản pháp luật dựa trên embedding ngữ nghĩa.
- Mục tiêu: từ một truy vấn ngôn ngữ tự nhiên, tìm ra các điều luật hoặc đoạn văn bản có nội dung ngữ nghĩa phù hợp nhất.



Giới thiệu

Đầu vào (Input):

- Truy vấn bằng ngôn ngữ tự nhiên

  Ví dụ: “Người lao động đơn phương chấm dứt hợp đồng trái luật thì có nghĩa vụ gì?”

  “Quy định xử phạt vi phạm nồng độ cồn khi điều khiển xe máy.”

- Tập văn bản pháp luật đã được số hóa.

Đầu ra (Output):

- Các văn bản hoặc điều khoản pháp luật liên quan nhất.
- Danh sách kết quả được xếp hạng theo mức độ tương đồng ngữ nghĩa với truy vấn.

Mục tiêu

- **Hệ thống hóa bài toán** truy vấn văn bản pháp luật dựa trên embedding ngữ nghĩa, làm rõ đầu vào, đầu ra và yêu cầu của bài toán trong bối cảnh dữ liệu pháp luật tiếng Việt.
- **Khảo sát và phân tích** các hướng tiếp cận tiêu biểu cho bài toán truy vấn văn bản, bao gồm:
  - Phương pháp tìm kiếm dựa trên từ khóa truyền thống.
  - Phương pháp truy vấn dựa trên embedding ngữ nghĩa.
- **Thiết kế và hiện thực mô hình baseline**, trong đó:
  - Biểu diễn truy vấn và các đoạn văn bản pháp luật dưới dạng embedding.
  - Xây dựng pipeline truy vấn và xếp hạng kết quả.
  - Đảm bảo tính đơn giản trong triển khai, khả năng mở rộng và tốc độ truy xuất nhanh.
- **Đánh giá** hiệu quả của phương pháp đề xuất dựa trên bộ dữ liệu **ALQAC LEGAL DATASETS**

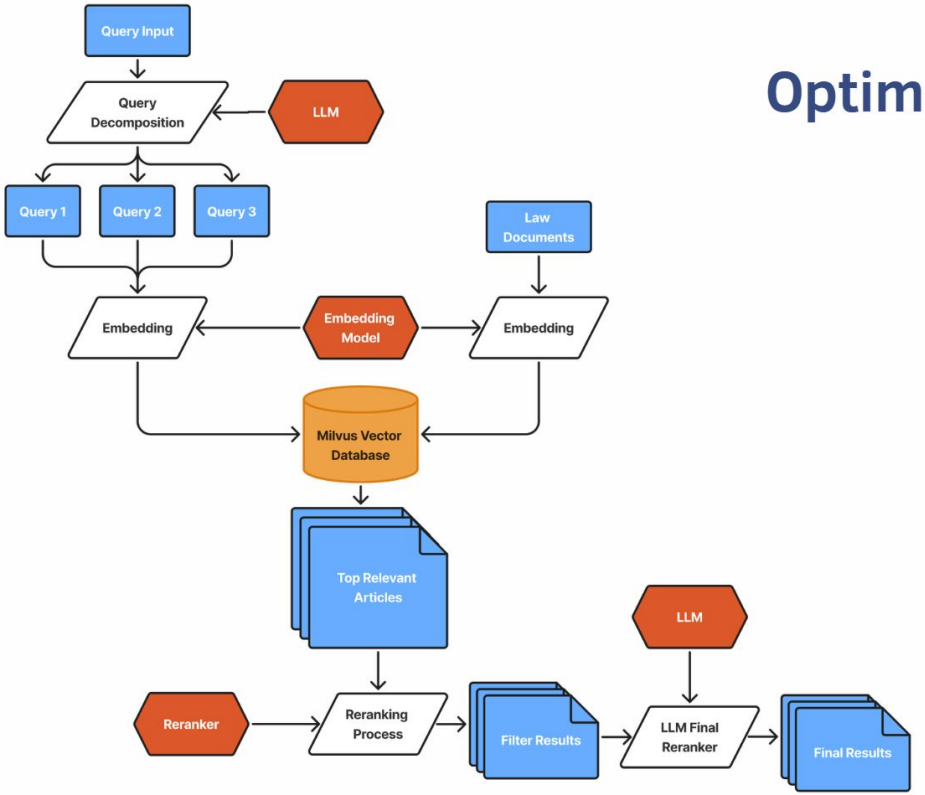
Nội dung và Phương pháp

- **Khảo sát các phương pháp truy vấn văn bản**
  a. Khảo sát hai hướng tiếp cận chính là truy vấn dựa trên từ khóa và Truy vấn dựa trên embedding ngữ nghĩa.
- **Lựa chọn hướng tiếp cận**
  a. Dựa trên phân tích ưu – nhược điểm, đề tài lựa chọn **phương pháp truy vấn dựa trên embedding ngữ nghĩa** làm hướng nghiên cứu chính.
  b. Mục tiêu là cân bằng giữa độ chính xác truy vấn và chi phí tính toán.
- **Chuẩn bị và xử lý dữ liệu**
  a. Thu thập văn bản pháp luật từ các nguồn công khai.
  b. Chia nhỏ văn bản thành các đơn vị phù hợp cho truy vấn.
  c. Loại bỏ các đoạn trùng lặp hoặc có nội dung tương đồng cao.
  d. Mã hóa toàn bộ các đoạn văn thành **vector embedding**, sử dụng BGE-M3 model.
  e. Lưu trữ các embedding trong **không gian vector** để hỗ trợ truy vấn nhanh.

Nội dung và Phương pháp

- **Xử lý truy vấn người dùng**
  a. Biểu diễn câu truy vấn bằng mô hình ngôn ngữ dưới dạng embedding.
  b. Thực hiện truy vấn trong không gian vector để tìm **top-k đoạn văn bản** có độ tương đồng ngữ nghĩa cao nhất.
  c. Một văn bản pháp luật được xem là phù hợp nếu chứa các đoạn văn tương ứng với các thành phần trong truy vấn.
- **Tổng hợp và xếp hạng kết quả**
  a. Tính điểm cho mỗi văn bản ứng viên dựa trên mức độ tương đồng ngữ nghĩa giữa truy vấn và các đoạn văn liên quan.
  b. Xếp hạng kết quả theo thứ tự giảm dần của điểm số.
  c. Trả về danh sách các văn bản phù hợp nhất cho người dùng.
- **Tiêu chí đánh giá**
  a. **Precision@K**: đánh giá chất lượng các kết quả truy vấn được trả về.
  b. **Query response time**: đánh giá thời gian phản hồi của hệ thống khi làm việc với kho văn bản quy mô lớn.

Tổng quan hệ thống



Optimal Method

Kết quả dự kiến

- Xây dựng hệ thống **truy vấn văn bản pháp luật dựa trên embedding ngữ nghĩa** theo hướng hai giai đoạn.
- Tạo **kho văn bản pháp luật đã tiền xử lý**, mỗi đoạn được biểu diễn bằng vector embedding và lưu trữ trong không gian vector.
- Cho phép **truy vấn ngôn ngữ tự nhiên**, trả về top-k đoạn văn bản pháp luật phù hợp theo độ tương đồng ngữ nghĩa.
- Xây dựng cơ chế **xếp hạng kết quả** dựa trên độ tương đồng embedding.
- Đánh giá hệ thống bằng **Precision@K** và **query response time**.

Tài liệu tham khảo

- [1]. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean:
  Efficient Estimation of Word Representations in Vector Space.
  ICLR 2013.
- [2]. Jeffrey Pennington, Richard Socher, Christopher D. Manning:
  GloVe: Global Vectors for Word Representation.
  EMNLP 2014.
- [3]. Nils Reimers, Iryna Gurevych:
  Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
  EMNLP 2019.
- [4]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin:
  Attention Is All You Need.
  NeurIPS 2017.
- [5]. Thang Nguyen, Huy Nguyen, Minh Nguyen, et al.:
  PhoBERT: Pre-trained Language Models for Vietnamese.
  EMNLP 2020.
- [6]. Stephen E. Robertson, Hugo Zaragoza:
  The Probabilistic Relevance Framework: BM25 and Beyond.
  Foundations and Trends in Information Retrieval 2009.
- [7]. Xiaofei He, Deng Cai, Ji-Rong Wen, Wei-Ying Ma:
  Learning a Unified Subspace for Face Recognition.
  IEEE Transactions on Pattern Analysis and Machine Intelligence 2005.

BÁO CÁO CỦA CÁC NHÓM

Môn học: CS519 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS519.Q11

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

UIT-HCM