

# Xây dựng Hệ Thống hỏi đáp nội quy trường ĐH CNTT (UIT)

## 1. Tổng quan

Hệ thống hỏi đáp phát triển cho trường Đại học Công nghệ Thông tin, được ứng dụng để cung cấp thông tin về ngành học, phương thức xét tuyển và môi trường học tập cho học sinh và phụ huynh. Nhờ khả năng cá nhân hóa và phản hồi nhanh chóng, hệ thống không chỉ giảm tải cho đội ngũ tư vấn mà còn nâng cao trải nghiệm người dùng.

## 2. Dataset

Bộ dữ liệu UIT-R2RQA gồm các câu hỏi về tuyển sinh và các văn bản tương ứng. Bộ dữ liệu có các cột: index, context, article, document, question, extractive answer, abstractive answer, yes/no

Bộ dữ liệu được chia thành 3 tập:

- Train: 7806 mẫu
- Valid: 976 mẫu
- Test: 976 mẫu

Trong đó, tập corpus có tổng cộng 294 cặp (context, article, document) duy nhất.

Các tập train, valid và test không chỉ để tạo tập corpus mà còn sử dụng để huấn luyện và đánh giá các mô hình sentence transformer nhằm mục đích cải thiện kết quả truy vấn.

## 3. Phương pháp

### Semantic Search

Semantic Search hoạt động bằng cách chuyển đổi câu truy vấn và văn bản/tài liệu thành vector trong không gian nhiều chiều, sau đó so sánh độ tương đồng giữa các cặp vector truy vấn và vector tài liệu để tìm kết quả phù hợp nhất.

Phương pháp này sử dụng **độ tương đồng** (cosine similarity) để đo lường mức độ liên quan giữa truy vấn và dữ liệu. Các vector được tạo bởi các embedding models. Khi truy vấn, hệ thống tìm **k vector tài liệu gần nhất** với vector truy vấn trong không gian embedding, giúp truy xuất tài liệu liên quan.

Sử dụng semantic search với các model

- Vietnamese-bi-encoder
- Halong\_embedding

**Vietnamese-bi-encoder** là mô hình sentence transformer được thiết kế để tạo ra các vector embedding 768 chiều cho văn bản tiếng Việt, hỗ trợ các tác vụ như tìm kiếm ngữ nghĩa và đo

lượng độ tương đồng giữa các câu, văn bản tiếng Việt. Nó được huấn luyện trên tập dữ liệu tiếng Việt bằng mô hình base là Phobert-base-v2, do đó, Vietnamese-bi-encoder có context length là 256 tokens.

**Halong\_embedding** là mô hình sentence transformer cho tiếng Việt tập trung vào hiệu quả trong các ứng dụng như RAG (Retrieval-Augmented Generation). Halong\_embedding cũng tạo ra các vector embedding 768 chiều cho văn bản tiếng Việt. Mô hình pre-train này được huấn luyện trên khoảng 100.000 cặp câu hỏi và tài liệu liên quan, sử dụng hàm mất mát Matryoshka từ mô hình base là multilingual-e5-base, cho phép giảm kích thước vector nhưng mà không làm giảm đáng kể hiệu suất. Điều này giúp tăng tốc độ so sánh và tìm kiếm văn bản. Mô hình này có context length là 512 tokens.

Cả hai mô hình Vietnamese-bi-encoder và Halong\_embedding đều thuộc loại bi-encoder trong kiến trúc sentence transformers. Trong phương pháp bi-encoder, mỗi câu hoặc đoạn văn được encode độc lập thành vector trong không gian nhiều chiều. Điều này cho phép tính toán nhanh chóng độ tương đồng giữa các văn bản bằng cách đo độ tương tự giữa các vector.

#### **Phương pháp huấn luyện bi-encoder:**

1. Chuẩn bị dữ liệu: Thu thập các cặp câu hỏi (anchor) và văn bản có liên quan (positive).
2. Mã hóa độc lập: Sử dụng mô hình transformer để mã hóa từng câu trong cặp thành vector riêng biệt.
3. Tính toán độ tương đồng: Sử dụng các phép đo như cosine similarity để đánh giá mức độ tương đồng giữa các vector.
4. Hàm mất mát: Áp dụng hàm mất mát MultipleNegativesRankingLoss để tối ưu hóa mô hình sao cho các cặp câu liên quan có độ tương đồng cao và các cặp không liên quan có độ tương đồng thấp.

Đối với mô hình Halong\_embedding, một kỹ thuật đặc biệt được sử dụng trong quá trình huấn luyện là Matryoshka loss. Kỹ thuật này cho phép giảm kích thước vector embedding mà không làm giảm đáng kể hiệu suất, giúp tăng tốc độ so sánh và tìm kiếm văn bản.

#### **Keyword search**

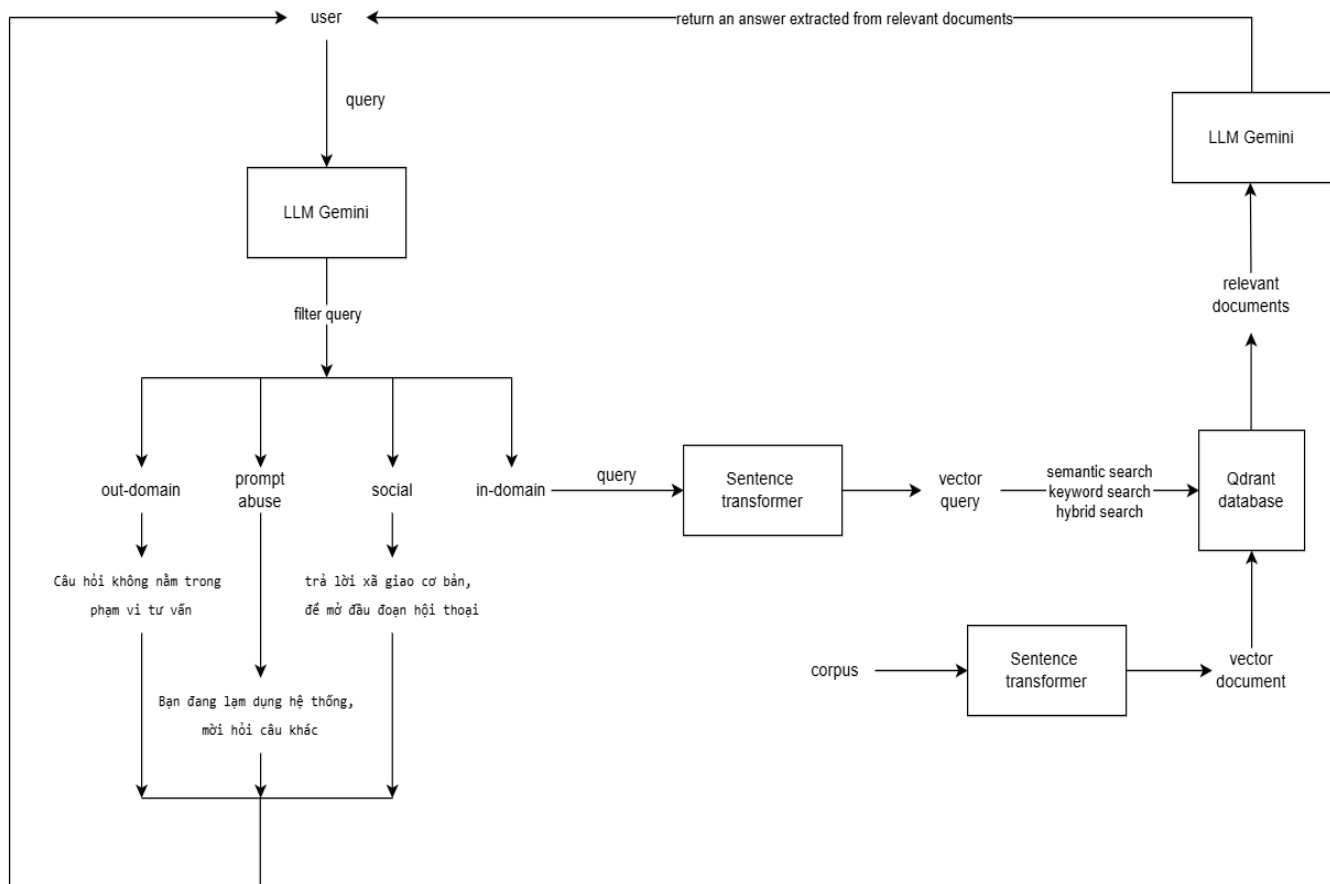
BM25 (Best Matching 25) là một thuật toán truy xuất thông tin tiên tiến, cải tiến từ TF-IDF, bằng cách kết hợp tần suất từ khóa (TF), độ hiếm của từ khóa (IDF), và điều chỉnh độ dài tài liệu thông qua các tham số  $k$  và  $b$ . So với TF-IDF, BM25 xử lý tốt hơn các tài liệu dài và truy vấn phức tạp, mang lại độ chính xác cao hơn. Tuy nhiên, BM25 phức tạp hơn và cần điều chỉnh tham số, trong khi TF-IDF đơn giản, dễ triển khai nhưng kém hiệu quả với tài liệu dài và truy vấn phức tạp. BM25 phù hợp cho hệ thống tìm kiếm hiện đại, còn TF-IDF thích hợp cho các ứng dụng đơn giản.

## Hybrid search

Hybrid search là kỹ thuật kết hợp semantic search và keyword search với trọng số alpha tùy chỉnh từ 0 tới 1:

- Nếu  $\alpha = 0$ , chỉ sử dụng bm25 (key word search)
- Nếu  $\alpha = 1$ , chỉ sử dụng vector search (semantic search)

## 4. Kiến trúc hệ thống



## 5. Đánh giá truy xuất thông tin

| Mô hình                           | Phương pháp truy xuất | Hit_rate@10   | MRR@10        | ndcg@10       |
|-----------------------------------|-----------------------|---------------|---------------|---------------|
| Vietnamese-bi-encoder (pre-train) | Keyword search        | 0.7284        | 0.4990        | 0.5540        |
|                                   | Semantic search       | 0.7469        | 0.4935        | 0.5538        |
|                                   | Hybrid 0.5            | 0.7284        | 0.4990        | 0.5540        |
| Vietnamese-bi-encoder (trained)   | Keyword search        | 0.9395        | 0.7059        | 0.7629        |
|                                   | Semantic search       | 0.9620        | 0.7793        | 0.8245        |
|                                   | Hybrid 0.5            | 0.9395        | 0.7059        | 0.7629        |
| Halong_embedding (pre-train)      | Keyword search        | 0.8739        | 0.6367        | 0.6941        |
|                                   | Semantic search       | 0.8995        | 0.6949        | 0.7449        |
|                                   | Hybrid 0.5            | 0.8739        | 0.6367        | 0.6941        |
| Halong_embedding (trained)        | Keyword search        | 0.9620        | 0.7468        | 0.7995        |
|                                   | Semantic search       | <b>0.9733</b> | <b>0.8263</b> | <b>0.8626</b> |
|                                   | Hybrid 0.5            | 0.9620        | 0.7468        | 0.7995        |


Sau khi được huấn luyện, có thể thấy hiệu quả cải thiện rõ rệt. Bên cạnh đó, Halong\_embedding có context length 512 tokens, lớn gấp đôi so với Vietnamese-bi-encoder cũng cho phép mô hình được nhận nhiều thông tin hơn từ văn bản đầu vào, từ đó cho kết quả truy xuất tốt hơn.


## 6. DEMO


Một số hình ảnh demo


### Hệ thống tư vấn quy định đào tạo UIT


 xin chào


 Chào bạn! Tôi là trợ lý ảo, hỗ trợ giải đáp các thắc mắc về nội quy của Trường Đại học Công nghệ Thông tin (UIT). Bạn cần giúp đỡ gì thêm không?


 điều kiện vào lớp tài năng

 Dựa vào thông tin của văn bản **QUY ĐỊNH ĐÀO TẠO CHƯƠNG TRÌNH TÀI NĂNG** Điều 8. *Phương thức xét tuyển đầu vào*, câu trả lời của tôi là: Điều kiện vào lớp tài năng gồm 2 phương án: Tuyển từ năm 1 (dựa trên kết quả đầu vào tuyển sinh, có thể có phỏng vấn) và tuyển sinh viên năm trên (ĐTBTL và ĐTBHK từ 7.5 trở lên, không chậm tiến độ, kết quả môn học cốt lõi tốt, ưu tiên sinh viên đạt giải thưởng "Sinh viên tài năng chạm tới đỉnh cao", có thể có kiểm tra, phỏng vấn thêm). Chi tiết từng điều kiện xem văn bản gốc.

 khi nào thì bị đuổi học

 Dựa vào thông tin của văn bản **QUY CHẾ ĐÀO TẠO TỪ XA TRÌNH ĐỘ ĐẠI HỌC CỦA TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**, câu trả lời của tôi là: Sinh viên bị buộc thôi học nếu đã bị cảnh báo học vụ 2 lần liên tiếp nhưng tiếp tục vi phạm mức cảnh báo học vụ; bị đình chỉ học vụ lần thứ hai; đã hết thời gian tối đa hoàn thành khóa học; không tham gia học tập 1 năm trở lên; vi phạm kỷ luật có tính chất và mức độ nghiêm trọng, gây ảnh hưởng xấu đến Trường và xã hội; vi phạm pháp luật bị xử phạt tù giam. Dựa vào thông tin của văn bản **QUY CHẾ ĐÀO TẠO THEO HỌC CHẾ TÍN CHỈ CHO HỆ ĐẠI HỌC CHÍNH QUY CỦA TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**, câu trả lời của tôi là: Sinh viên bị buộc thôi học nếu ĐTBHK hai học kỳ liên tiếp gần nhất bằng 0; đã hết thời hạn được cho phép bảo lưu nhưng không làm thủ tục nhập học lại; đã hết thời gian tối đa hoàn thành khóa học quy định nhưng vẫn chưa hoàn thành chương trình đào tạo; sau khi hết thời gian được cho thôi học theo mục b khoản này; bị cảnh báo học vụ lần thứ hai liên tiếp và vẫn tái phạm; bị đình chỉ học tập lần thứ hai; vi phạm kỷ luật có tính chất và mức độ nghiêm trọng, gây ảnh hưởng xấu đến Trường và xã hội; hoặc vi phạm pháp luật bị xử phạt tù giam.

 Với mức điểm TOEIC nào đối với CTTT sinh viên sẽ học anh văn 5?

 Dựa vào thông tin của văn bản **QUY ĐỊNH ĐÀO TẠO NGOẠI NGỮ ĐỐI VỚI HỆ ĐẠI HỌC CHÍNH QUY CỦA TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN** Điều 4. *Kiểm tra xếp lớp đầu khóa cho sinh viên thuộc CTC, CTTN, CTTT*, câu trả lời của tôi là: Từ 500 điểm trở lên.