ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Hỏi đáp dịch vụ hành chính công

PHẠM DUY ANH
anh.pd183481@sis.hust.edu.vn

Ngành: Khoa học máy tính

Giảng viên hướng dẫn: PGS.TS. Lê Thanh Hương _____

Chữ kí GVHD

Khoa: Khoa học máy tính

Trường: Công nghệ thông tin và Truyền thông

HÀ NỘI, 08/2023

LỜI CAM KẾT

Họ và tên sinh viên: ..... Phạm Duy Anh .....
Điện thoại liên lạc: ..... 0857163189 .....
Email: ..... anh.pd183481@sis.hust.edu.vn .....
Lớp: ..... Khoa học máy tính 03 - K63 .....
Hệ đào tạo: ..... Kỹ sư chính quy .....

Tôi – *Phạm Duy Anh* – cam kết Đồ án Tốt nghiệp (ĐATN) là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của *PGS.TS. Lê Thanh Hương*. Các kết quả nêu trong ĐATN là trung thực, là thành quả của riêng tôi, không sao chép theo bất kỳ công trình nào khác. Tất cả những tham khảo trong ĐATN – bao gồm hình ảnh, bảng biểu, số liệu, và các câu từ trích dẫn – đều được ghi rõ ràng và đầy đủ nguồn gốc trong danh mục tài liệu tham khảo. Tôi xin hoàn toàn chịu trách nhiệm với dù chỉ một sao chép vi phạm quy chế của nhà trường.

Hà Nội, ngày tháng năm

Tác giả ĐATN

Phạm Duy Anh

LỜI CẢM ƠN

Trước tiên, tôi xin bày tỏ lòng biết ơn sâu sắc đến Trường Công nghệ thông tin và truyền thông đã cung cấp cho tôi một môi trường học tập chất lượng và các nguồn tài nguyên quý báu để thực hiện Đồ án tốt nghiệp. Sự hỗ trợ và sự chu đáo từ Trường đã đóng góp rất lớn vào thành công của tôi.

Tiếp theo, tôi muốn gửi lời cảm ơn chân thành tới PGS.TS. Lê Thanh Hương - người hướng dẫn tận tâm của tôi. Sự kiên nhẫn, kiến thức sâu rộng và sự đồng hành suốt quá trình nghiên cứu đã giúp tôi phát triển và hoàn thiện Đồ án tốt nghiệp của mình. Tôi rất biết ơn vì những chỉ dẫn quý giá và sự động viên không ngừng từ PGS.TS. Lê Thanh Hương.

Tôi cũng muốn cảm ơn những giảng viên tại Trường Công nghệ thông tin và Truyền thông - Đại học Bách Khoa Hà Nội, đã truyền tải những kiến thức một cách đầy đủ, dễ hiểu nhất suốt những năm tháng được học trên giảng đường. Và cũng đặc biệt gửi lời cảm ơn tới anh Hoàng Thành Đạt cùng những người bạn luôn đồng hành cùng tôi trong chuyến hành trình vừa qua, Đặng Thái Sơn, Nguyễn Hữu Kiệt, Đặng Duy Anh, Lê Thị Vân, Nguyễn Quang Huy, Nguyễn Đình Thắng, Nguyễn Nam Hán, Nguyễn Hải Anh.

Và cuối cùng, tôi muốn gửi lời cảm ơn đến gia đình, bạn bè, những người luôn sẵn sàng chia sẻ và giúp đỡ tôi trong học tập và cuộc sống. Mong rằng, chúng ta sẽ mãi mãi gắn bó với nhau.

Xin chân thành cảm ơn.

TÓM TẮT NỘI DUNG ĐỒ ÁN

Đề tài đồ án của tôi tập trung vào nghiên cứu và giải quyết bài toán hỏi đáp trong lĩnh vực hành chính công. Hiện nay, việc cung cấp thông tin và hỗ trợ dịch vụ cho công dân trong hành chính công gặp phải nhiều khó khăn, mất thời gian và không hiệu quả. Việc tìm kiếm thông tin và thực hiện các thủ tục hành chính cần được cải thiện để nâng cao trải nghiệm người dùng.

Để giải quyết vấn đề này, tôi đã lựa chọn hướng tiếp cận sử dụng các mô hình truy xuất thông tin dựa trên nền tảng xử lý ngôn ngữ tự nhiên. Hướng tiếp cận này cho phép xây dựng các mô hình thông minh có khả năng tự động tìm kiếm và cung cấp thông tin chính xác từ các tài liệu văn bản pháp luật.

Trong đề tài, tôi đã tiến hành xây dựng hệ thống hỏi đáp dịch vụ công bằng cách kết hợp nhiều mô hình truy xuất thông tin, trong đó có các mô hình dựa trên thuật ngữ và mô hình ngữ nghĩa dựa trên BERT - một mô hình ngôn ngữ tiên tiến với khả năng biểu diễn ngôn ngữ tự nhiên hiệu quả. Cụ thể ở đây là sử dụng SimCSE cùng với các mô hình xếp hạng kết hợp với mô hình BM25+ truyền thống. Trong đồ án này, tôi cũng xây dựng bộ dữ liệu phục vụ riêng cho quá trình hoàn thiện đồ án.

Kết quả nghiên cứu của tôi cho thấy hệ thống đạt được hiệu suất tốt trong việc trả lời câu hỏi và truy xuất thông tin hành chính công. Đây là đóng góp chính của đề tài, cải thiện trải nghiệm người dùng và giúp nâng cao hiệu quả hoạt động của hệ thống hành chính công.

Với thành công đạt được trong việc áp dụng mô hình truy xuất thông tin trong lĩnh vực hành chính công, đề tài hy vọng góp phần thúc đẩy sự phát triển và ứng dụng của trí tuệ nhân tạo trong lĩnh vực này, đồng thời cung cấp một giải pháp hiệu quả và tiện lợi cho việc truy cập và sử dụng dịch vụ hành chính công.

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 1

1.1 Đặt vấn đề 1

1.2 Các giải pháp hiện tại và hạn chế 2

1.3 Mục tiêu và định hướng giải pháp 6

1.4 Đóng góp của đồ án 7

1.5 Bố cục đồ án 7

CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT 8

2.1 Tổng quan về bài toán truy xuất thông tin 8

2.2 Các phương pháp truy xuất dựa trên thuật ngữ 9

2.2.1 TF-IDF (Term Frequency - Inverse Document Frequency) 9

2.2.2 BM25 (Best Matching 25) 10

2.3 Các phương pháp truy xuất dựa trên ngữ nghĩa 11

2.3.1 Sử dụng phương pháp biểu diễn từ 11

2.3.2 Sử dụng các mô hình chủ đề 14

2.4 Các phương pháp truy xuất dựa trên mạng neural 14

2.4.1 Mô hình BERT 14

2.4.2 Sentence Transformer 17

2.5 Các phương pháp xếp hạng tài liệu 19

2.6 Các phương pháp đánh giá hiệu năng hệ thống truy xuất thông tin 20

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT 24

3.1 Xây dựng bộ dữ liệu 24

3.1.1 Mục tiêu xây dựng bộ dữ liệu 24

3.1.2 Phương pháp thu thập dữ liệu 25

3.1.3 Tiền xử lý dữ liệu 26

3.2 Tổng quan hệ thống hỏi đáp dịch vụ công 26
3.2.1 Đầu vào hệ thống hỏi đáp dịch vụ công 27
3.2.2 Đầu ra hệ thống hỏi đáp dịch vụ công 28
3.2.3 Kiến trúc hệ thống hỏi đáp dịch vụ công 28
3.3 Mô hình truy xuất tài liệu 29
3.3.1 Truy xuất mức từ vựng 29
3.3.2 Truy xuất mức ngữ nghĩa 29
3.4 Mô hình xếp hạng tài liệu 32
3.5 Mô hình kết hợp (Ensemble Model) 33
CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM 35
4.1 Mô tả bộ dữ liệu 35
4.2 Các phương pháp đánh giá 36
4.3 Môi trường cài đặt 37
4.4 Phương pháp thực nghiệm 38
4.5 So sánh các mô hình Truy xuất thông tin 39
4.6 So sánh các mô hình Xếp hạng tài liệu 40
4.7 Đánh giá ảnh hưởng của chiến lược tăng cường dữ liệu 41
4.8 Đánh giá hiệu năng của hệ thống hỏi đáp dịch vụ công 42
CHƯƠNG 5. KẾT LUẬN 45
5.1 Kết luận 45
5.2 Hướng phát triển trong tương lai 45
TÀI LIỆU THAM KHẢO 48
PHỤ LỤC 50
A. MỘT SỐ KẾT QUẢ CỦA HỆ THỐNG HỎI ĐÁP DỊCH VỤ CÔNG .. 50

DANH MỤC HÌNH VẼ

Hình 2.1 Tổng quan hệ thống truy xuất thông tin 8
Hình 2.2 Biểu diễn đầu vào của mô hình BERT [5] 15
Hình 2.3 Kiến trúc mô hình BERT [5] 16
Hình 2.4 Kiến trúc sentence transformer 18
Hình 2.5 Minh họa cách tính P(k) 22

Hình 3.1 Kiến trúc hệ thông hỏi đáp dịch vụ công 28
Hình 3.2 (a) Unsupervised SimCSE (b) Supervised SimCSE [24] 30
Hình 3.3 Mô hình Bi-Encoder và Cross-Encoder 32

Hình 4.1 So sánh chiều dài tài liệu trong kho dữ liệu 36
Hình 4.2 So sánh chiều dài câu hỏi trong bộ dữ liệu 37

DANH MỤC BẢNG BIỂU

Bảng 1.1 Ví dụ mẫu câu trả lời cho câu hỏi dịch vụ công 6

Bảng 3.1 Ví dụ về dữ liệu văn bản pháp luật thu thập 24
Bảng 3.2 Ví dụ về dữ liệu hỏi đáp dịch vụ công thu thập 25
Bảng 3.3 Ví dụ về dữ liệu câu hỏi sau khi tiền xử lý 27

Bảng 4.1 Phân bố chiều dài tài liệu trong corpus 35
Bảng 4.2 Số lượng các văn bản liên quan đối với từng câu truy vấn . . . 35
Bảng 4.3 So sánh các mô hình Truy xuất thông tin 39
Bảng 4.4 So sánh các mô hình xếp hạng sau khi tìm kiếm bằng SimCSE với k=100 40
Bảng 4.5 So sánh các mô hình xếp hạng sau khi tìm kiếm bằng BM25+ với k=100 40
Bảng 4.6 Đánh giá ảnh hưởng của chiến lược tăng cường dữ liệu với k=100 41
Bảng 4.7 Đánh giá ảnh hưởng của chiến lược tăng cường dữ liệu với k=5 42
Bảng 4.8 Đánh giá hiệu năng của hệ thống hỏi đáp dịch vụ công 42
Bảng 4.9 Đánh giá hiệu năng truy xuất top 5 tài liệu của hệ thống 43

Bảng A.1 50
Bảng A.2 51
Bảng A.3 52
Bảng A.4 53
Bảng A.5 54
Bảng A.6 55

DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	**Ý nghĩa**
AI	Trí tuệ nhân tạo
BE	Mô hình mã hóa kép
CE	Mô hình mã hóa chéo
DVC	Dịch vụ công
IR	Truy xuất thông tin
LSI	Phân tích ngữ nghĩa ẩn
MAP	Mean Avarage Precision
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
NDCG	Mô hình ngôn ngữ
NLP	Xử lý ngôn ngữ tự nhiên
QA	Hỏi đáp
TF-IDF	Tần suất thuật ngữ - Tần suất nghịch đảo tài liệu

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

Việt Nam đang chứng kiến sự phát triển như vũ bão của công nghệ thông tin trong những năm gần đây. Điều này đã tạo ra một cơ hội lớn để nâng cao chất lượng và hiệu quả của các dịch vụ hành chính công thông qua quá trình số hóa. Sự tiến bộ trong lĩnh vực trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên cũng đã mở ra cánh cửa cho việc tạo ra các mô hình ngôn ngữ mạnh mẽ có khả năng giao tiếp gần như hoàn hảo với con người.

Việc số hóa các dịch vụ hành chính công tại Việt Nam đã nhận được sự quan tâm và ủng hộ từ các cơ quan chính phủ và tổ chức từ trung ương tới địa phương. Mục tiêu của quá trình số hóa này là tạo ra một môi trường điện tử trong đó người dân và doanh nghiệp có thể truy cập và sử dụng dịch vụ hành chính công một cách thuận tiện và hiệu quả hơn. Điều này không chỉ giúp tiết kiệm thời gian và công sức, mà còn tạo ra sự minh bạch và đáng tin cậy trong việc cung cấp thông tin và hỗ trợ từ phía chính quyền.

Với sự phát triển trong tác vụ xử lý ngôn ngữ tự nhiên (NLP), việc nghiên cứu và áp dụng các mô hình ngôn ngữ cho bài toán hỏi đáp dịch vụ hành chính công sẽ mang lại nhiều lợi ích đáng kể. Việc tận dụng các tiến bộ trong lĩnh vực trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên để xây dựng các hệ thống thông minh và tự động có khả năng tương tác với người dùng sẽ đóng góp quan trọng vào việc cải thiện trải nghiệm người dùng và tăng cường hiệu quả của dịch vụ hành chính công trong việc số hóa quy trình và cung cấp thông tin chính xác và kịp thời.

Bài toán **Hỏi đáp dịch vụ hành chính công** không chỉ đơn thuần là giải quyết vấn đề truy cập thông tin và thực hiện các thủ tục hành chính. Nó còn có ý nghĩa quan trọng trong việc thay thế vai trò của con người trong việc đưa ra câu trả lời nhanh chóng và chính xác, đặc biệt trong bối cảnh nhà nước đang tăng cường quá trình số hóa hồ sơ và các thủ tục hành chính.

Sự phát triển các mô hình hỏi đáp dịch vụ hành chính công sẽ tạo ra những lợi ích đáng kể. Thay vì phải tìm kiếm thông tin trên nhiều nguồn khác nhau hoặc phải đối mặt với các quy trình phức tạp và rườm rà, người dùng có thể tương tác với hệ thống thông qua việc đặt câu hỏi và nhận được câu trả lời tức thì. Điều này giúp tiết kiệm thời gian, tăng tính tiện lợi và giảm bớt công đoạn rườm rà và phiền toái.

Đồng thời, việc sử dụng các mô hình hỏi đáp trong dịch vụ hành chính công đáng chú ý làm giảm sự phụ thuộc vào con người và đảm bảo tính chính xác và