CHƯƠNG 5: ĐÓNG GÓP ĐỒ ÁN

Trong suốt quá trình thực hiện đồ án. Các đóng góp được phân chia thành từng mục, mỗi mục đều có phần dẫn dắt, giải pháp đề xuất, và kết quả đạt được. Đây là cơ sở để khẳng định tính sáng tạo và giá trị thực tiễn của đề tài.

0.1 Xây dựng quy trình xử lý trọn vẹn từ ảnh đến nhãn tài liệu

Giới thiệu vấn đề

Trong nhiều hệ thống OCR hiện tại, quá trình xử lý thường chỉ dừng lại ở việc trích xuất văn bản từ ảnh, chưa có bước phân loại để gán nhãn loại tài liệu. Điều này gây khó khăn khi quản lý tập tài liệu lớn và đa dạng.

Giải pháp

Đồ án đề xuất một quy trình hoàn chỉnh gồm: (i) tiền xử lý ảnh bằng OpenCV, (ii) trích xuất văn bản bằng PaddleOCR, (iii) biểu diễn ngữ nghĩa bằng PhoBERT, và (iv) phân loại bằng Logistic Regression. Các bước này được tích hợp trong một pipeline thống nhất.

Kết quả

Pipeline hoạt động ổn định, có khả năng xử lý nhiều loại giấy tờ khác nhau. Hệ thống nhận dạng và phân loại tài liệu phổ biến như CCCD, hóa đơn, hợp đồng với độ chính xác cao, giảm thiểu sai sót so với nhập liệu thủ công.

0.2 Ứng dụng PhoBERT để cải thiện độ chính xác phân loại

Giới thiệu vấn đề

Các phương pháp cũ chủ yếu sử dụng đặc trưng thủ công hoặc embedding đa ngôn ngữ (mBERT), dẫn đến hạn chế khi áp dụng cho tiếng Việt – một ngôn ngữ giàu ngữ cảnh, nhiều từ ghép và dấu thanh.

Giải pháp

Đồ án áp dụng PhoBERT – mô hình ngôn ngữ tiền huấn luyện cho tiếng Việt – để tạo vector đặc trưng. Nhờ đó, hệ thống không chỉ nhận diện ký tự mà còn hiểu được ngữ cảnh.

Kết quả

Thử nghiệm cho thấy độ chính xác phân loại cải thiện rõ rệt so với cách chỉ dựa vào từ khóa hoặc embedding cơ bản. PhoBERT giúp mô hình Logistic Regression đạt hiệu suất tốt với tập dữ liệu giới hạn.

0.3 Kết hợp PaddleOCR và Tesseract để tăng độ tin cậy

Giới thiệu vấn đề

Một công cụ OCR đơn lẻ khó đảm bảo độ chính xác tuyệt đối trong mọi trường hợp. Ví dụ, Tesseract dễ bị lỗi với tiếng Việt, trong khi PaddleOCR có lúc nhận dạng sai ký tự ở ảnh mờ.

Giải pháp

Hệ thống triển khai cả PaddleOCR và Tesseract. Văn bản từ hai công cụ được ghép lại và chuẩn hóa trước khi phân loại, giúp giảm thiểu mất thông tin.

Kết quả

Độ đầy đủ của văn bản OCR tăng lên, đặc biệt trong trường hợp ảnh chất lượng trung bình. Hệ thống hoạt động linh hoạt và có thể so sánh, chọn kết quả OCR tốt hơn.

0.4 Xây dựng ứng dụng web trực quan bằng Streamlit

Giới thiệu vấn đề

Nhiều hệ thống nhận dạng hiện tại chỉ có dạng dòng lệnh (CLI), khó tiếp cận với người dùng phổ thông, gây hạn chế trong thử nghiệm và trình diễn kết quả.

Giải pháp

Đồ án phát triển ứng dụng web dựa trên Streamlit, cho phép người dùng tải ảnh, xem ảnh gốc và ảnh tiền xử lý, hiển thị văn bản OCR, và kết quả phân loại theo thời gian thực.

Kết quả

Ứng dụng thân thiện, dễ sử dụng, hỗ trợ trực quan hóa pipeline xử lý. Giao diện giúp cả người không có nền tảng kỹ thuật cũng có thể trải nghiệm hệ thống.

0.5 Khả năng mở rộng và huấn luyện lại mô hình

Giới thiệu vấn đề

Trong thực tế, nhu cầu phân loại giấy tờ rất đa dạng và có thể thay đổi theo thời gian. Một hệ thống cứng nhắc sẽ nhanh chóng lỗi thời khi xuất hiện loại giấy tờ mới.

Giải pháp

Đồ án thiết kế quy trình huấn luyện lại mô hình Logistic Regression bằng tập dữ liệu mới. Nhãn tài liệu được lưu trong file JSON, dễ bổ sung và cập nhật.

Kết quả

Hệ thống có khả năng mở rộng linh hoạt: khi thêm loại giấy tờ mới, chỉ cần bổ sung dữ liệu và chạy lại script huấn luyện. Điều này giúp ứng dụng duy trì hiệu quả trong môi trường thực tế.

Kết chương

Chương này đã tổng hợp các đóng góp chính của đồ án: xây dựng pipeline xử lý ảnh – văn bản – phân loại, áp dụng PhoBERT cho tiếng Việt, kết hợp nhiều công cụ OCR để tăng độ tin cậy, phát triển ứng dụng Streamlit trực quan, và thiết kế khả năng mở rộng hệ thống. Những đóng góp này thể hiện sự sáng tạo, tính thực tiễn và giá trị ứng dụng của đề tài trong quá trình số hóa và quản lý tài liệu.