07/2025

THỰC TẬP DOANH NGHIỆP

ỨNG DỤNG OCR ĐỂ TRÍCH XUẤT THÔNG TIN Y TẾ TỪ GIẤY TỜ KHÁM BỆNH VÀ LƯU VÀO HỆ THỐNG HIS/EMR.

Mentor: MR Kính

Thành viên:

Lê Hoàng Minh Quý (Lead)
Nguyễn Văn Minh Khánh
Trịnh Quốc Dân
Phạm Phước Bảo Tín
Hồ Tăng Nhật Hiếu

Thực tập doanh nghiệp BDATA

ĐẶT VẤN ĐỀ

- Kết quả khám bệnh chủ yếu lưu dưới dạng giấy
- Nhập liệu thủ công tốn thời gian, dễ sai
- Khó quản lý, tìm kiếm và phân tích dữ liệu

LÝ DO CHỌN BÀI TOÁN

- Tự động hóa nhập liệu bằng OCR giúp tiết kiệm thời gian
- Giảm sai sót, nâng cao hiệu quả quản lý
- Phù hợp xu hướng chuyển đổi số trong ngành y tế

MỤC TIÊU

- Có thể tự động trích xuất và giảm tải công việc nhập tay dữ liệu thông tin y tế từ ảnh hoặc PDF.
- Tăng tính chính xác và tốc độ xử lý dữ liệu đầu vào
- Chuẩn hóa dữ liệu lưu vào cơ sở dữ liệu.

ỨNG DỤNG

- Quét và lưu trữ thông tin
- Trích xuất dữ liệu y tế từ tài liệu cũ
- Triển khai tại phòng khám nhỏ hoặc bệnh viện sử dụng hệ thống HIS
- Giúp nhân viên chỉ cần scan hoặc chụp ảnh là có thể lưu thông tin vào hệ thống
- Tiền đề cho việc xây dựng hồ sơ sức khỏe điện tử (EHR)
- Bảo vệ dữ liệu y tế quan trọng: các thông tin nhạy cảm không an toàn ở dạng giấy có thể được số hóa để bảo mật cao hơn.

ĐẦU VÀO

- Giấy tờ khám bệnh (Ảnh hoặc PDF)
- Đơn thuốc hoặc các tài liệu y tế khác

ĐẦU RA

- Văn bản đã được OCR và chuẩn hóa
- Được lưu vào cơ sở dữ liệu (DB)

ROADMAP

Giai đoạn	Nội dung thực hiện	Thời gian
1	Tìm hiểu & lựa chọn công nghệ	Tuần 01
2	Tiền xử lý ảnh và chuẩn hóa dữ liệu	Tuần 02 - 04
3	Nhận dạng văn bản bằng OCR	Tuần 05 - 07
4	Thiết kế giao diện và cơ sở dữ liệu	Tuần 08 - 10
5	Tích hợp toàn bộ hệ thống: Giao diện ↔ OCR ↔ Tiền xử lý ↔ Hậu xử lý ↔ CSDL	Tuần 11 - 13

Phân chia công việc

Giai đoạn	Nội dung thực hiện	Thành viên phụ trách
1	Tìm hiểu & lựa chọn công nghệ	Toàn bộ thành viên
2	Tiền xử lý ảnh và chuẩn hóa dữ liệu	Hiếu, Khánh
3	Nhận dạng văn bản bằng OCR	Dân, Hiếu, Quý
4	Thiết kế giao diện và cơ sở dữ liệu	Tín, Dân
5	Tích hợp toàn bộ hệ thống: Giao diện ↔ OCR ↔ Tiền xử lý ↔ Hậu xử lý ↔ CSDL	Khánh, Quý, Tín

Giai đoạn 1: Tìm hiểu & lựa chọn công nghệ

Mục tiêu: Phân tích yêu cầu bài toán và xác định công nghệ phù hợp

- Tìm hiểu đề tài, xác định đầu vào và đầu ra
- Phân tích yêu cầu hệ thống OCR y tế
- Khảo sát & so sánh các công cụ: Tesseract, EasyOCR, Google Vision,...
- Lựa chọn công nghệ chính thức cho từng bước xử lý
- Thu thập dữ liệu mẫu (ảnh/PDF đơn thuốc, kết quả xét nghiệm)
- Lập kế hoạch, phân chia nhiệm vụ cho các thành viên

Giai đoạn 2: Tiền xử lý ảnh, chuẩn hóa dữ liệu

Mục tiêu: Cải thiện chất lượng đầu vào và đầu ra để tối ưu độ chính xác khi nhận dạng.

- Áp dụng các kỹ thuật tiền xử lý như chuyển ảnh sang trắng đen (binarization), làm thẳng ảnh (deskew), loại bỏ nhiễu (denoise), cắt vùng chứa thông tin quan trọng (ROI).
- Sau OCR, thực hiện hậu xử lý: sửa lỗi chính tả, loại bỏ ký tự dư, chuẩn hóa định dạng (JSON, bảng,...).
- Áp dụng kỹ thuật NER để nhận diện các thực thể như: tên thuốc, ngày tháng, đơn vị đo lường,...

Giai đoạn 3: Nhận dạng văn bản bằng OCR

Mục tiêu: Trích xuất nội dung văn bản từ ảnh

- Cài đặt và tích hợp công cụ OCR đã chọn (Tesseract OCR, EasyOCR)
- Xử lý nhiều loại tài liệu (đơn thuốc, xét nghiệm, chẩn đoán...)
- Đánh giá độ chính xác của kết quả OCR
- Kiểm tra với dữ liệu thật

Giai đoạn 4: Thiết kế giao diện & Cơ sở dữ liệu

Mục tiêu: Xây dựng giao diện nhập liệu thân thiện và hệ thống lưu trữ linh hoạt.

- Thiết kế giao diện web hoặc app đơn giản cho phép người dùng tải lên ảnh/PDF.
- Thiết kế và xây dựng cơ sở dữ liệu (MongoDB, PostgreSQL...) để lưu trữ thông tin đã chuẩn hóa.
- Đảm bảo dữ liệu dễ truy xuất, tìm kiếm và mở rộng trong tương lai.

Giai đoạn 5: Kết nối toàn bộ hệ thống

Mục tiêu: Tích hợp liền mạch các thành phần thành một quy trình tự động hoàn chỉnh.

- Kết nối giao diện với công cụ OCR, mô-đun tiền xử lý, hậu xử lý và cơ sở dữ liệu.
- Tạo một pipeline xử lý tự động: từ khi tải ảnh lên → trích xuất nội dung → xử lý → lưu dữ liệu.
- Kiểm thử hệ thống, đo hiệu suất xử lý và cải tiến dựa trên phản hồi.

CẢM ƠN
ĐÃ LẮNG NGHE !