CHƯƠNG 3: CÔNG NGHỆ SỬ DỤNG

Chương này trình bày các công nghệ, nền tảng và thuật toán được sử dụng trong đồ án "Nhận dạng và phân loại giấy tờ". Mỗi công nghệ đều được phân tích mục đích sử dụng, vai trò trong hệ thống, và lý do lựa chọn so với các giải pháp thay thế. Qua đó, nội dung chương này tạo cầu nối giữa yêu cầu đã phân tích ở Chương 2 và phần thiết kế – hiện thực ở các chương tiếp theo.

Ứng dụng được xây dựng trên Python 3.10 với nhiều thư viện hỗ trợ, mỗi thư viện đảm nhận một vai trò riêng trong toàn bộ quy trình xử lý ảnh và phân loại giấy tờ:

- Streamlit: Cung cấp môi trường nhanh gọn để phát triển giao diện web. Người dùng có thể tải ảnh, xem kết quả OCR, loại giấy tờ và tải dữ liệu xuất ra mà không cần lập trình front-end phức tạp.
- OpenCV: Thư viện xử lý ảnh mạnh mẽ, đảm nhiệm các bước làm sạch ảnh như chuyển sang thang xám, khử nhiễu, tăng tương phản và phân ngưỡng, giúp cải thiện chất lượng đầu vào cho OCR.
- Pillow: Công cụ xử lý ảnh cơ bản, hỗ trợ mở, chuyển đổi và chuẩn bị dữ liệu hình ảnh trước khi đưa qua OpenCV hoặc OCR.
- PyTesseract: Giao diện Python cho Tesseract OCR Engine. Cho phép trích xuất văn bản số hóa từ ảnh tài liệu, đặc biệt hữu ích với văn bản in rõ nét.
- PaddleOCR & PaddlePaddle: Hệ thống OCR tối ưu cho nhiều ngôn ngữ, trong đó có tiếng Việt. PaddleOCR giúp nhận dạng chính xác hơn với ảnh phức tạp, chữ in mờ hoặc bố cục nhiều cột.
- Torch: Nền tảng tính toán tensor và deep learning. Đóng vai trò backend cho PhoBERT, hỗ trợ tính toán embedding văn bản trên CPU/GPU.
- Transformers: Thư viện của HuggingFace. Trong dự án, PhoBERT được sử dụng để biến đổi văn bản OCR thành vector đặc trưng, phục vụ phân loại.
- Scikit-learn: Cung cấp các thuật toán machine learning. Logistic Regression được sử dụng để dự đoán loại tài liệu dựa trên embedding từ PhoBERT.
- Joblib: Dùng để lưu và tải lại mô hình Logistic Regression, giúp triển khai nhanh mà không cần huấn luyện lại.
- Pandas & XlsxWriter: Hỗ trợ xử lý kết quả OCR, quản lý bảng dữ liệu và xuất file CSV/Excel để lưu trữ hoặc chia sẻ.
- Regex (re): Thư viện chuẩn của Python, được dùng để lọc và chuẩn hóa văn

bản (ví dụ: loại bỏ ký tự đặc biệt, định dạng ngày tháng).

Kết chương

Chương này đã giới thiệu và phân tích chi tiết các công nghệ sử dụng trong đồ án: PaddleOCR, Tesseract, PhoBERT, Logistic Regression, PyTorch/Transformers, Streamlit, OpenCV và các công cụ bổ trợ. Với mỗi công nghệ, đã chỉ ra vai trò, lý do lựa chọn và so sánh với các giải pháp thay thế. Đây chính là cơ sở để bước sang giai đoạn thiết kế và hiện thực hóa ở các chương tiếp theo.