QUY ĐỊNH ĐỒ ÁN GIỮA KỲ

I. YÊU CẦU CHUNG VỀ PHƯƠNG PHÁP VÀ NGỮ LIỆU

1. Mục tiêu:

Sinh viên xây dựng một bộ dữ liệu OCR hoàn chỉnh gồm:

- Ảnh gốc đã xử lý (nếu cần).
- Ảnh crop theo từng dòng/khu vực văn bản.
- Bounding box chuẩn cho từng dòng
- Gióng hàng để text khớp chính xác với ảnh

2. Công cụ:

- **Không ràng buộc** các phải sử dụng các công cụ, giải thuật, phương pháp hay mô hình cung cấp
- Nếu các em tự tìm công cụ/mô hình khác, cần **trao đổi trước** để tránh xử lý/đánh giá sai.
- Một số công cụ tốt để OCR gợi ý là: Paddle OCR, Google Vision,...
- Có thể sử dụng PPOCRLabel (Paddle Label)**:** kiểm tra và chỉnh sửa bounding box, export tự động. https://github.com/PFCCLab/PPOCRLabel

3. Yêu cầu :

Sinh viên **bắt buộc** đảm bảo:

- Bounding box thẳng, không nghiêng, không thừa nền.
- Không crop thiếu chữ.
- Text khớp hoàn toàn với ảnh.
- Dòng phải được phân tách chính xác.
- Ảnh xử lý phải rõ ràng, không làm mất nét.

4. Khối lượng: Mỗi sinh viên phải hoàn thành tối thiểu 3.000 dòng (= 3.000 bbox + 3.000 ảnh crop + 3.000 text tương ứng).

5. Qui trình:

🎬 **Chuẩn bị dữ liệu**: Chuyển đổi pdf (dữ liệu thô đã được cung cấp) sang ảnh , text

🎬 **Xử lý ảnh (nếu cần)**: crop, deskew, tăng tương phản, giảm nhiễu.

🎬 **Tạo Bounding Box**: dùng OCR/tool, kiểm tra thủ công, bbox sát chữ, không bỏ sót dòng.

🎬 **Dóng hàng text**: nhập text khớp 100% với ảnh, dựa trên TXT, không sửa lỗi gốc.

🎬 **Crop ảnh**: mỗi bounding box → 1 ảnh crop

🎬 **Xuất dữ liệu**:

- raw/: Label.txt, fileState.txt, rec_gt.txt
- final/: Label.txt, fileState.txt, rec_gt.txt, final_labels.txt/json
- crop_img/: ảnh crop dùng chung hoặc tách raw/final

II. HẠN NỘP VÀ BÁO CÁO

Hạn nộp: Thầy sẽ thông báo sau (dự kiến 4-5 tuần tính từ 17/11/2025)

Các dữ liệu cần nộp:

Mục	Bắt buộc	Mô tả
Ảnh gốc / ảnh xử lý	✔	Nếu có xử lý
Raw OCR	✔	Để đối chiếu độ chính xác
Final labels (bbox + text)	✔	File quan trọng nhất
Ảnh crop theo dòng	✔	1 bbox = 1 ảnh
Báo cáo	✔	Ghi rõ số dòng ≥ 3.000, khoảng 7-12 trang, ghi tất cả các việc đã làm đồ án. bao gồm cả việc xử lý data, cách tạo bbox,..
Source code (nếu dùng)	Optional	Nộp nếu có dùng code, cần ghi rõ hướng dẫn chi tiết chạy source code ở README.md

Định dạng file cần nộp:

Folder / File	Ý nghĩa
raw/	Kết quả OCR tự động
├─ Label.txt	Bounding box detection đã lưu tự động
├─ fileState.txt	Trạng thái các ảnh đã được xác nhận
└─ rec_gt.txt	Text OCR nhận dạng tự động
final/	Kết quả sau khi dóng hàng
├─ Label.txt	Bounding box detection đã chỉnh sửa
├─ fileState.txt	Trạng thái ảnh đã kiểm tra
├─ rec_gt.txt	Text OCR đã chỉnh sửa/dóng hàng
crop_img/	Ảnh crop dùng chung cho raw và final (hoặc có thể tách theo subfolder raw_crop/ và final_crop/)
img/	Ảnh đã chuyển từ pdf
txt/	Text đã chuyển từ pdf