ĐẠI HỌC QUỐC GIA TPHCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

Báo cáo Giữa kỳ

Đề tài: Low resource NLP

Môn học: Nhập môn xử lý ngôn ngữ tự nhiên

Sinh viên thực hiện:

Lê Hải Đăng (23122005)

Nguyễn Nhật Minh (23122010)

Nguyễn Văn Khoa (23122016)

Bùi Anh Quân (23122017)

Giáo viên hướng dẫn:

Thầy Đinh Điền

NCS Lâm Hương

Ngày 14 tháng 12 năm 2025

Mục lục

1 Giới thiệu đề tài và mục tiêu của đồ án

1.1 Đề tài

Low language model: Low language model (mô hình ngôn ngữ cho ngôn ngữ ít tài nguyên) là hướng nghiên cứu tập trung vào việc xây dựng và cải thiện các mô hình xử lý ngôn ngữ tự nhiên cho những ngôn ngữ có lượng dữ liệu huấn luyện hạn chế, điển hình như tiếng Việt trong nhiều bài toán OCR và các bài toán liên quan đến mô hình ngôn ngữ lớn (LLM).

Trong thực tế, phần lớn các mô hình ngôn ngữ và hệ thống OCR hiện nay được huấn luyện chủ yếu trên các ngôn ngữ giàu tài nguyên như tiếng Anh, tiếng Trung, dẫn đến hiệu quả thấp khi áp dụng cho ngôn ngữ khác, đặc biệt là tiếng Việt do sự khác biệt về ngữ pháp, dấu thanh, ký tự đặc biệt và cách viết. Vì vậy, việc nghiên cứu low language model là cần thiết nhằm nâng cao độ chính xác nhận dạng, giảm lỗi sai sau OCR và cải thiện chất lượng dữ liệu đầu ra cho các tác vụ xử lý tiếp theo.

Tuy nhiên, việc xây dựng low language model gặp nhiều khó khăn, bao gồm: thiếu dữ liệu huấn luyện chất lượng cao, dữ liệu không đồng nhất về định dạng và nguồn gốc, chi phí lớn cho việc gán nhãn thủ công, cũng như các lỗi phổ biến trong OCR tiếng Việt như nhầm lẫn dấu thanh, ký tự có hình dạng tương tự và lỗi tách/gộp từ. Do đó, cần có một quy trình tiền xử lý dữ liệu chặt chẽ, kết hợp các công cụ OCR hiện có với việc chỉnh sửa và chuẩn hóa nhãn, nhằm tạo ra tập dữ liệu đủ tốt để huấn luyện hoặc tinh chỉnh mô hình ngôn ngữ cho tiếng Việt.

1.2 Mục tiêu của đồ án

1.2.1 Đồ án 1 (giữa kì): OCR một tài liệu tiếng Việt

- Áp dụng kỹ thuật OCR (Google Vision) để thu thập nhãn sơ bộ cho tài liệu tiếng Việt.
- Tinh chỉnh bounding box và label bằng tay nhằm đảm bảo chất lượng dữ liệu.
- Xuất tập dữ liệu chuẩn (Label.txt, rec_gt.txt, crop_img) phục vụ các mô-đun tiếp theo.

1.2.2 Đồ án 2 (cuối kì)

2 Quy trình thực hiện

Trong đồ án này, nhóm thực hiện đã tiến hành xử lý toàn bộ ngữ liệu được giao, bao gồm hai bộ tài liệu chính:

- *Lịch sử Việt Nam tập 14: Từ năm 1975 đến năm 1986* - Trần Đức Cường (2017).
- *Lịch sử Việt Nam tập 03: Từ thế kỷ XV đến thế kỷ XVI* - Tạ Ngọc Liễn (2017).

Tổng khối lượng dữ liệu được xử lý lên đến hơn **1.163 trang** tài liệu. Quy trình xử lý được chia thành các bước cụ thể như sau:

2.1 Chuyển đổi dữ liệu thô

Dữ liệu đầu vào là các tệp PDF scan nguyên bản. Bước đầu tiên là tách (split) toàn bộ các file PDF này thành từng trang ảnh riêng biệt dưới định dạng `.jpg`. Việc chuyển đổi này nhằm mục đích tách biệt dữ liệu để thuận tiện cho việc xử lý cục bộ trên từng trang, đồng thời đảm bảo định dạng đầu vào tương thích với các công cụ OCR.

2.2 Tối ưu hóa độ phân giải

Sau khi chuyển đổi sang định dạng ảnh, toàn bộ dữ liệu ảnh được chuẩn hóa lại độ phân giải (DPI). Nhóm thực hiện đã thiết lập thông số DPI trong khoảng **300 - 350 DPI**.
Lý do thực hiện: Đây là ngưỡng độ phân giải tối ưu cho các bài toán OCR tiếng Việt. Nếu DPI quá thấp (dưới 150), các nét chữ sẽ bị mờ, gây khó khăn cho việc nhận dạng dấu. Nếu DPI quá cao (trên 600), dung lượng lưu trữ sẽ tăng đột biến mà không cải thiện đáng kể độ chính xác, gây lãng phí tài nguyên tính toán.

2.3 Tự động hóa trích xuất đặc trưng

Tại bước này, nhóm sử dụng công cụ **Google Vision API** để thực hiện hai nhiệm vụ chính:

1. Tự động phát hiện và vẽ các hộp bao (Bounding Box) quanh các dòng văn bản.

2. Nhận dạng ký tự quang học (OCR) để trích xuất văn bản thô ban đầu.

Kết quả của bước này là dữ liệu `raw`, bao gồm các ảnh crop sơ bộ (`crop_img`) và các tệp thông tin tọa độ, nhãn text tương ứng. Việc sử dụng Google Vision giúp giảm thiểu đáng kể thời gian gán nhãn thủ công ban đầu nhờ độ chính xác tương đối cao của mô hình.
Dữ liệu đầu ra được tổ chức theo hai định dạng chính:

- **rec_gt.txt**: Chứa ánh xạ giữa đường dẫn ảnh crop và văn bản nhận dạng tương ứng.
- **Label.txt**: Chứa tọa độ bounding box dưới dạng 4 điểm (top-left, top-right, bottom-right, bottom-left).

Ví dụ minh họa cấu trúc file rec_gt.txt:

```
fi/p0143_020.jpg    Ban Van co Dai hanh khien va Hanh khien 5 dao dung dau, sau
fi/p0143_021.jpg    cung co lay chuc Boc xa la Hanh khien. Thu den Thuong thu dung
fi/p0143_022.jpg    dau bo. Phan Huy Chu cho rang: "Bay gio moi dat hai bo la Lai bo,
fi/p0143_023.jpg    Le bo; thuoc quan co Lang trung, Vien ngoai lang, Chu su".
```

Ví dụ minh họa cấu trúc file Label.txt:

```
fi/p0449_026.jpg    [[186, 4891], [3842, 4891], [3842, 5058], [186, 5058]]
fi/p0449_027.jpg    [[196, 5089], [1586, 5089], [1586, 5221], [196, 5221]]
fi/p0449_028.jpg    [[407, 5303], [3842, 5303], [3842, 5456], [407, 5456]]
fi/p0449_029.jpg    [[196, 5467], [3837, 5467], [3837, 5630], [196, 5630]]
```

2.4 Quy trình kiểm thử và Căn chỉnh thủ công

Đây là bước quan trọng nhất để đảm bảo chất lượng bộ dữ liệu. Dữ liệu sau khi đi qua Google Vision được kiểm tra thủ công từng trang (human-in-the-loop). Các công việc cụ thể bao gồm:

- **Kiểm tra Bounding Box:** Điều chỉnh lại các box bị lệch, box quá rộng chứa nhiều nền, hoặc box cắt lẹm vào nét chữ.
- **Kiểm tra Text Label:** Đối chiếu text nhận dạng với ảnh gốc để sửa các lỗi chính tả, lỗi nhận diện dấu câu hoặc các ký tự đặc biệt.
- **Loại bỏ nhiễu:** Xóa bỏ các box nhận diện sai (nhiễu nền, vết bẩn trên giấy scan, số trang, header/footer không cần thiết).

2.5 Đóng gói và Xuất dữ liệu

Sau khi hoàn tất quá trình kiểm thử, dữ liệu sạch được xuất ra theo cấu trúc yêu cầu của đề bài.

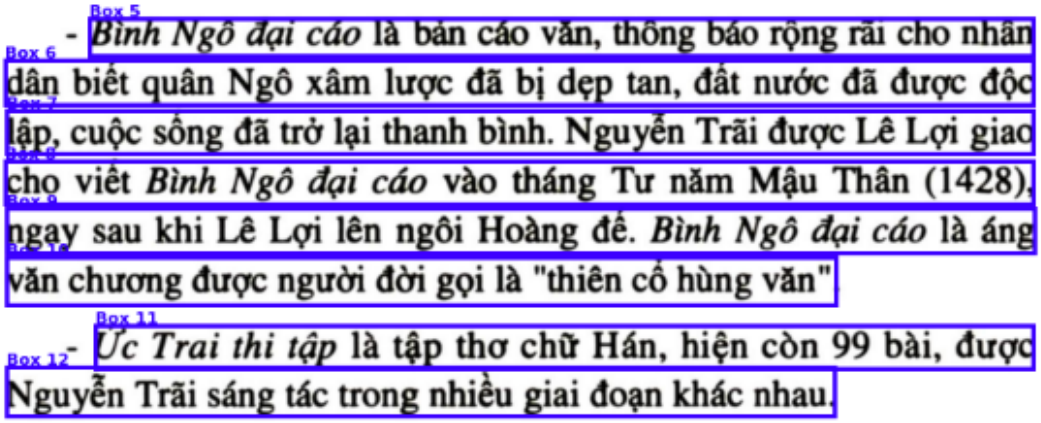
- **Hình ảnh:** Thư mục `final` chứa các ảnh cắt dòng đã được căn chỉnh chính xác.
- **Nhãn:** Các tệp `Label.txt`, `rec_gt.txt` chứa thông tin tọa độ và nội dung text đã được chuẩn hóa.

Bộ dữ liệu cuối cùng đảm bảo tính nhất quán giữa ảnh và nhãn, sẵn sàng cho việc huấn luyện và đánh giá mô hình.

2.6 Các lỗi phổ biến về Định dạng và Nhận diện

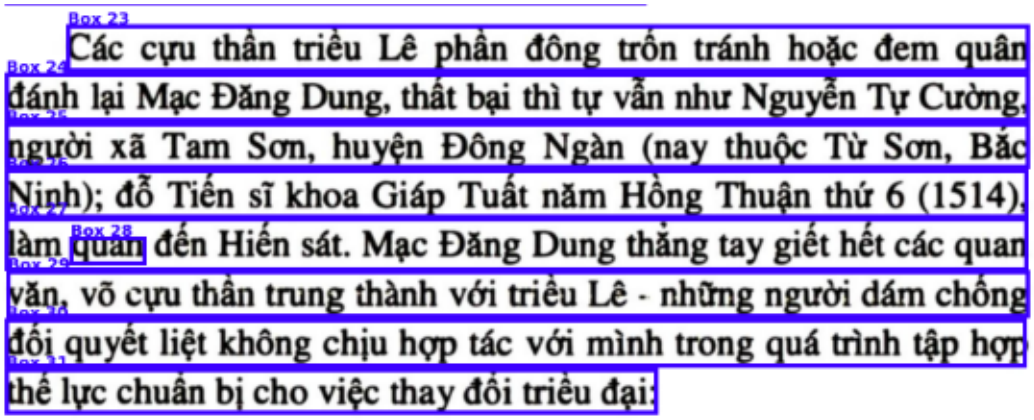
Đây là các lỗi xuất hiện thường xuyên do hạn chế của thuật toán OCR đối với văn bản tiếng Việt và bố cục trang in:

- **Lỗi khoảng trắng:** Mô hình thường xuyên thêm hoặc bớt khoảng trắng không chính xác, đặc biệt là ở các vị trí tiếp giáp giữa các từ hoặc dấu câu.
- **Bỏ qua ký tự liệt kê:** Các ký tự đánh dấu đầu dòng (gạch đầu dòng, chấm tròn) thường bị OCR bỏ qua hoặc nhận diện nhầm thành các ký tự nhiễu (noise), gây mất cấu trúc của các đoạn liệt kê.



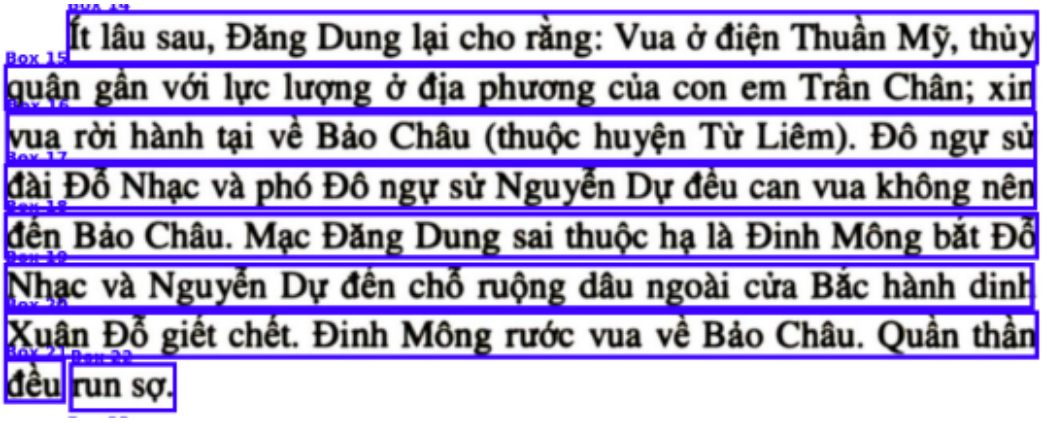
Hình 1: Ví dụ lỗi bỏ qua ký tự liệt kê

- **Nhận diện không đồng bộ:** Xuất hiện hiện tượng trên cùng một dòng văn bản thực tế lại tồn tại hai dòng OCR chồng lấp lên nhau với nội dung khác biệt. Đây là lỗi nghiêm trọng cần phải xóa bỏ thủ công một trong hai để tránh trùng lặp dữ liệu.



Hình 2: Ví dụ lỗi nhận diện không đồng bộ

- **Lỗi phân mảnh dòng:** Hiện tượng một dòng văn bản vật lý bị thuật toán OCR tách thành hai hoặc nhiều bounding box riêng biệt, thường xảy ra ở cuối dòng hoặc với các từ ngắn. Do sự biến thiên không đồng đều của khoảng cách giữa các từ (inter-word spacing) trong văn bản in cũ, khiến mô hình nhận diện sai ngưỡng phân cách (thresholding gap), dẫn đến việc ngắt mạch liên kết của dòng (text line connectivity). Việc này tạo ra các mẫu dữ liệu vụn vặt, làm mất ngữ cảnh nếu không được gộp (merge) lại thủ công.



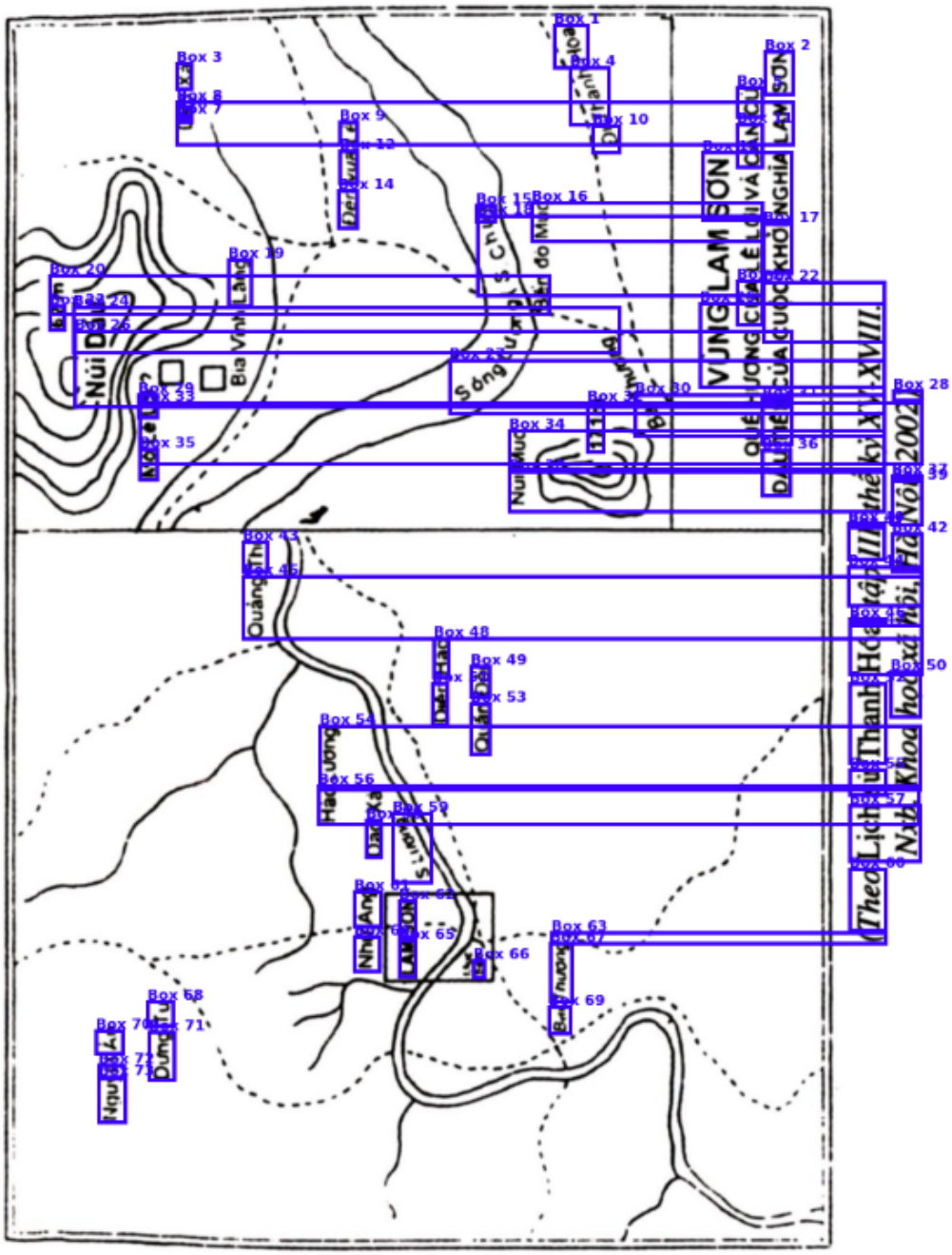
Hình 3: Ví dụ lỗi phân mảnh dòng

- **Lỗi mã hóa nội tại:** Một số trường hợp xuất hiện các token lạ như `PROTECTED_DECIMAL`, đây là lỗi do cơ chế xử lý số học hoặc tokenization của hệ thống OCR khi gặp các định dạng số đặc biệt.

2.7 Thách thức với dữ liệu Bảng biểu và Bản đồ

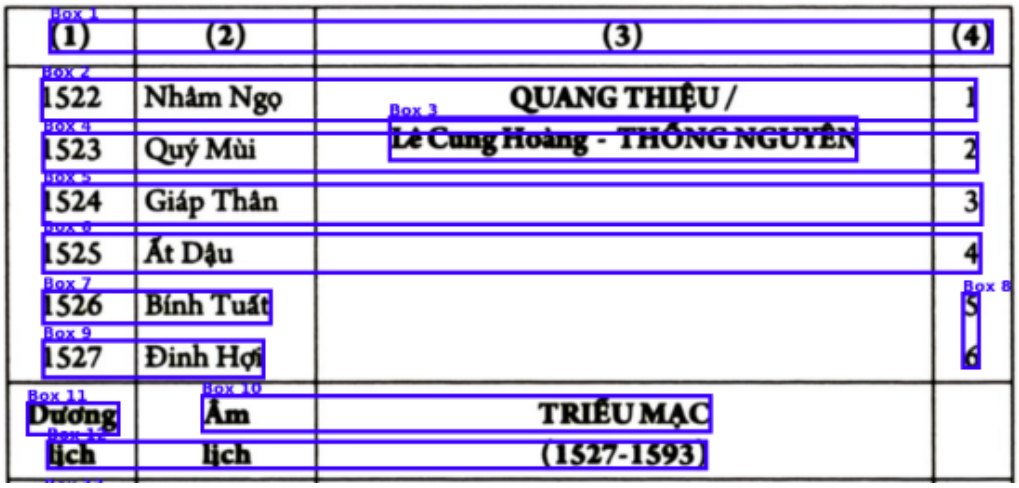
Dữ liệu lịch sử chứa nhiều bảng thống kê và bản đồ minh họa, gây khó khăn lớn cho các mô hình OCR tiêu chuẩn:

- **Đối với Bản đồ:** OCR hoạt động kém hiệu quả do chữ viết nằm rải rác, xoay nhiều hướng và lẫn vào nền đồ họa phức tạp.



Hình 4: Ví dụ OCR bản đồ

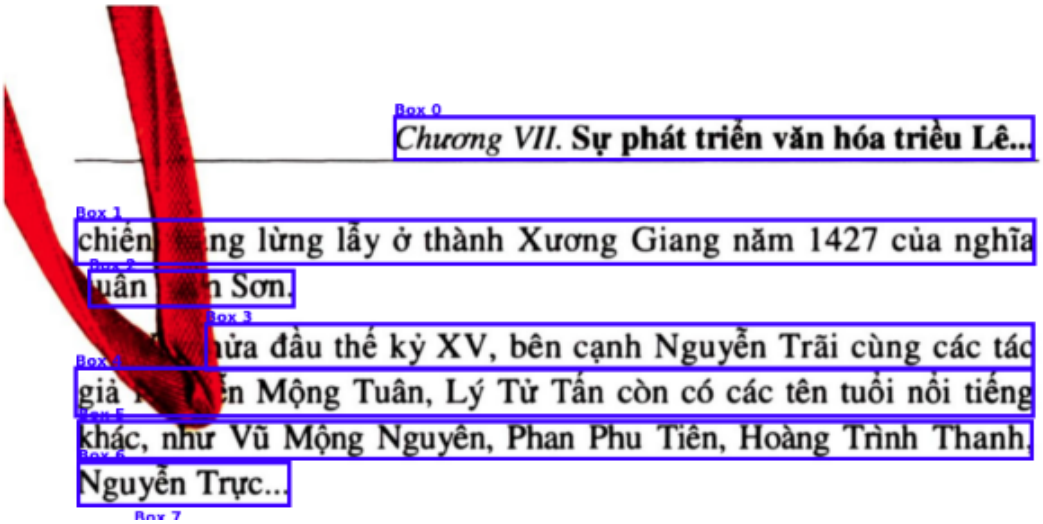
- **Đối với Bảng biểu:** Cấu trúc dòng/cột bị phá vỡ. Việc gán nhãn đòi hỏi nỗ lực thủ công lớn để định nghĩa lại các bounding box sao cho khớp với từng ô (cell) dữ liệu, đảm bảo tính toàn vẹn của thông tin thống kê.



Hình 5: Ví dụ OCR dạng bảng

2.8 Nhiễu vật lý

Chất lượng bản scan gốc ảnh hưởng trực tiếp đến kết quả đầu ra. Ghi nhận tại trang 347 xuất hiện dị vật (sợi dây) trong quá trình scan, tạo thành các đường kẻ giả (fake lines) cắt ngang văn bản, làm sai lệch việc phân đoạn dòng (line segmentation).



Hình 6: Ví dụ về dị vật

2.9 Giới hạn vật lý của tín hiệu

Một vấn đề đặc thù được ghi nhận là việc OCR không thể nhận diện các ký tự chú thích (footnote markers, ví dụ: [1],[2]) nằm sát dấu ngoặc kép hoặc cuối câu.

Phân tích nguyên nhân: Đây không được coi là lỗi thuật toán, mà là *giới hạn vật lý của tín hiệu ảnh.*

1. **Kích thước quá nhỏ:** Các ký tự footnote thường có kích thước chỉ bằng 1/3 hoặc 1/4 ký tự thường. Với độ phân giải scan tiêu chuẩn, số lượng pixel cấu thành ký tự này không đủ để tạo nên các đặc trưng hình học rõ ràng.
2. **Độ tương phản thấp:** Nét mực mờ và mảnh khiến chúng chìm vào nhiễu nền.
3. **Thiếu ngữ cảnh:** OCR hoạt động dựa trên nhận dạng hình ảnh chứ không có khả năng suy diễn ngữ pháp hay ngữ cảnh để "đoán" sự tồn tại của chú thích.

Giải pháp: Nhóm thực hiện xác định đây là các "unrecoverable glyphs" và chấp nhận loại bỏ hoặc xử lý bằng quy tắc hậu kiểm thay vì cố gắng gán nhãn cưỡng bức.

2.10 Kết quả thực hiện

Tính đến thời điểm báo cáo, nhóm đã hoàn tất việc gán nhãn và kiểm thử cho toàn bộ ngữ liệu, bao gồm cả các phần dữ liệu khó như bảng biểu và các trang bị nhiễu vật lý. Bộ dữ liệu sau khi làm sạch đảm bảo độ tin cậy cao cho việc huấn luyện mô hình.

3 Thống kê kết quả và kết luận

3.1 Thống kê

Tài liệu	Thành viên	Số trang	Số dòng
LỊCH SỬ VIỆT NAM - TẬP 3	Bùi Anh Quân	680	21340
LỊCH SỬ VIỆT NAM - TẬP 14	Nguyễn Nhật Minh	160	3870
LỊCH SỬ VIỆT NAM - TẬP 14	Nguyễn Văn Khoa	160	5002
LỊCH SỬ VIỆT NAM - TẬP 14	Lê Hải Đăng	163	5864

3.2 Kết luận

Sau quá trình OCR bằng Google Vision và tinh chỉnh thủ công, ta thu được bộ nhãn sơ bộ có chất lượng đủ để xây dựng tập train cho bước tiếp theo. Việc đối chiếu và chỉnh sửa thủ công giúp giảm

các loại lỗi phổ biến (nhầm lẫn ký tự, tách/gộp bounding box, encoding ký tự đặc biệt). Kết quả đầu ra (Label.txt và rec_gt.txt) có thể dùng để huấn luyện/finetune mô hình nhận dạng ký tự và mô hình ngôn ngữ cho tiếng Việt.

4 Phân công công việc

Thành viên	MSSV	Công việc
Lê Hải Đăng	23122005	Tinh chỉnh kết quả OCR (100%)
Nguyễn Nhật Minh	23122010	Tinh chỉnh kết quả OCR; Tổng hợp báo cáo (100%)
Nguyễn Văn Khoa	23122016	Tinh chỉnh kết quả OCR (100%)
Bùi Anh Quân	23122017	OCR bằng Google Vision; Tinh chỉnh kết quả OCR; Tổng hợp báo cáo (100%)

Tài liệu tham khảo

- **Lịch sử Việt Nam - Tập 3: Từ thế kỷ XV đến thế kỷ XVI** — Tạ Ngọc Liễn (Chủ biên). NXB Khoa học Xã hội (2017).
- **Lịch sử Việt Nam - Tập 14: Từ năm 1975 đến năm 1986** — Trần Đức Cường (Chủ biên). NXB Khoa học Xã hội (2017).
- **Google Cloud Vision API Documentation** — Google Cloud. cloud.google.com/vision/docs/ocr
- **PP-OCR: A Practical Ultra Lightweight OCR System** — Yuning Du, Chenxia Li, Ruoyu Guo, et al. (Baidu Inc.). *arXiv preprint arXiv:2009.09941*, 2020. (Cơ sở lý thuyết của công cụ PPOCRLabel). arxiv.org/abs/2009.09941