**HƯỚNG DẪN THỰC HIỆN BÀI TẬP 2**

File toàn hình, xử lý liên tục

Sau bài tập 1, chúng ta đã có kết quả các box và chữ Hán Nôm được OCR, sắp xếp và dóng hàng. Bài tập 2 chúng ta sẽ thực hiện trên 2 dạng file:

* Dạng 1: File nhiều trang. Các file dữ liệu thực tế các bạn sẽ phải làm là những file dài có nhiều trang. Các bạn cần viết code để có thể chạy tự động những file này.
* Dạng 2: File toàn hình. Có nhiều file dữ liệu được scan/chụp từ những cuốn sách có sẵn dẫn đến toàn bộ các các chữ Hán Nôm và chữ Quốc ngữ cũng là hình ảnh, chúng ta cần OCR luôn chữ Quốc ngữ (và kiểm lỗi chính tả sau khi OCR). Gợi ý (thầy chỉ gợi ý, các bạn có thể làm cách khác cũng được):
  + OCR chữ QN: dùng Google Cloud Vision API:

<https://www.youtube.com/watch?v=q8QRd4CUuvs>

* + Kiểm lỗi chính tả:
    - Dùng ChatGPT để kiểm lỗi
    - hoặc dùng transformers (tham khảo code ví dụ đính kèm)

**Nội dung bài tập:**

Bước 1: Viết code “extract\_continuous.py” dựa trên code “extract.py” của Bài tập 1 để trích xuất các hình ảnh và các văn bản (Hán Nôm và QN) tương ứng từ file nhiều trang “Bondaytrecon.pdf” (file dạng 1).

Bước 2: Viết code “extract\_ocr.py” dựa trên code “extract.py” của Bài tập 1, bổ sung thêm phần OCR chữ Quốc ngữ (dùng Google Cloud Vision API) để trích xuất hình ảnh và văn bản (Hán Nôm, QN) tương ứng (chủ yếu quan tâm OCR chữ QN vì dùng nó làm chuẩn) từ file “Kieu1866page045.pdf” (file dạng 2).

Bước 3: Sử dụng PPOCRLabel để OCR hình ảnh trích xuất được.

Bước 4: Chỉnh file “convert\_output.py” để trích nội dung file “Label.txt” cho phù hợp với cả 2 dạng văn bản trên.

**Yêu cầu nộp bài:**

* Tạo thư mục đặt tên là MSSV (thay MSSV bằng thông tin của sinh viên)
* Chép các file “extract\_continuous.py”, “ extract\_ocr.py”, “convert\_output.py”, “output.xlsx” vào thư mục MSSV trên
* Nén thư mục MSSV thành “MSSV.zip” vào nộp lên moodle
* Hạn nộp: 23h59 ngày 13/11/2024 (Thứ Tư).