**HƯỚNG DẪN THỰC HIỆN BÀI TẬP 1**

Sắp xếp bbox, dóng hàng, đối sánh

Sau khi nghiên cứu cách sử dung công cụ PPOCRLabel, chúng ta đã có kết quả các bounding box (bbox) và các chữ Hán Nôm được OCR ra. Qua thực tập, các bạn cũng thấy các bbox có thể bị nhận diện sai/thiếu/thừa hay không đúng thứ tự (từ trên xuống dưới và từ trái sang phải theo cách viết chữ Hán xưa) và các chữ Hán Nôm có thể bị OCR ra sai/thiếu/thừa (tô màu đỏ trong ví dụ đính kèm) hay OCR ra đúng nhưng khác mã unicode (tô màu xanh). Yêu cầu bài tập 1 này là phát hiện và khắc phục các lỗi sai trên. Cụ thể như sau:

1. Sắp xếp các box theo đúng thứ tự (phải qua trái, trên xuống): Dựa vào tọa độ (lưu ý đến độ lệch)
2. Dóng hàng các box (box alignment) theo các câu dịch của chữ Quốc ngữ: Dựa vào chiều dài tương ứng của từng box với chiều dài câu QN.
3. Dóng hàng từng chữ Hán Nôm được OCR ra với từng chữ QN tương ứng trong 1 box (char alignment): Dựa vào giải thuật M.E.D Levenstein. Trong đó, việc so sánh 2 chữ cần dóng hàng được thực hiện bằng cách lấy phần giao giữa 2 tập S1 và S2, trong đó:

S1 = {tập những chữ Hán Nôm có hình dáng tương đồng (dùng tự điển SinoNom\_Similar\_Dic)} và

S2 = {tập những chữ Hán Nôm khả dĩ được dịch từ 1 chữ QN (dùng tự điển QuocNgu\_SinoNom\_Dic)}.

Việc lấy phần giao này có thể xảy ra 1 trong 3 trường hợp sau:  
a. Kết quả giao = 1: Đó chính là chữ OCR đúng và trùng với chữ QN tương ứng.

b. Kết quả giao > 1: Lấy chữ OCR có độ tương đồng cao nhất (từ trái sang phải) và chữ đó cũng được xem là OCR đúng.

c. Kết quả giao = 0: Không xác định. Đánh dấu, . Xem đó là chữ OCR sai.

Ví dụ: cặp chữ đầu tiên trong box đầu tiên, ta có:

chữ “𤾓”có S1 = {'𤾓','庹', '度', '税', '脣', '賡', '榥', '板', '𥝽', '釈', '椩', '𣈓', '棅', '𣌋', '梘', '磨', '稂', '𥟌', '庚', '佒', '算']}

chữ “trăm” có S2 = {啉, 林, 百,𤾓,𬃴, 󰭖, 󱦏}

Kết quả giao S1&S2 = {𤾓} => đây là chữ OCR đúng và trùng với chữ QN “trăm”

**Nội dung bài tập:**

Bước 1: Viết code “extract.py” để trích xuất hình ảnh và văn bản (Hán Nôm và QN) từ file “LCPv\_001.pdf” (dựa theo file code ví dụ đính kèm).

Bước 2: Sử dụng PPOCRLabel để OCR hình ảnh trích xuất được và lưu file "Label.txt".

Bước 3: Tự viết 1 file python script "convert\_output.py" để trích nội dung file "Label.txt" ra file "output.xlsx" với các yêu cầu sau:

* Sắp xếp các box theo đúng thứ tự (phải qua trái, trên xuống)
* Dóng hàng các box (box alignment) theo các câu dịch của chữ QN: Dựa vào chiều dài tương ứng của từng box với chiều dài câu QN được trích ở Bước 1. **Đánh dấu** các box không có câu QN tương ứng.
* Dóng hàng từng chữ Hán Nôm được OCR ra với từng chữ QN tương ứng trong 1 box (char alignment) dựa vào giải thuật M.E.D Levenstein. **Đánh dấu** các chữ có kết quả giao khác 1.
* Xuất file “output.xlsx” theo mẫu “LCPv.001.xlsx” với các yêu cầu sau:
  + Cột “ID”: lấy tên file ảnh làm ID và có thêm thông tin:
    - ppp: trang của hình ảnh Hán Nôm trong file ngữ liệu thô đó
    - ss: thứ tự của box
  + Cột “Image Box”: thông tin tọa độ của box (sau khi sắp xếp), box nào dóng hàng mà không có câu QN tương ứng thì tô màu xanh lá (#00FF00) (tham khảo code ví dụ đính kèm để biết cách tô màu trong excel)
  + Cột “SinoNom OCR”: kết quả OCR tương ứng với box đó sau khi đã được dóng hàng chữ:
    - Chữ nào có kết quả giao = 1 thì để màu mặc định
    - Chữ nào có kết quả giao > 1 thì tô màu xanh dương (#0000FF)
    - Chữ nào có kết quả giao = 0 thì tô màu đỏ (#FF0000)
  + Cột “SinoNom Char”: văn bản Hán Nôm trích từ file PDF
  + Cột “Chữ Quốc ngữ”: văn bản QN trích từ file PDF

**Yêu cầu nộp bài:**

* Tạo thư mục đặt tên là MSSV (thay MSSV bằng thông tin của sinh viên)
* Chép các file "extract.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 tối Thứ Hai tuần sau (28/10/2024).