Lưu ý:

* Đối với Windows OS, bạn nên dùng convert\_data\_to \_labelsPaddle\_v2.py.
* Đối với Linux OS, bạn nên dùng convert\_data\_to \_labelsPaddle\_linux.py.

B0. Đây là ví dụ mẫu của file Excel output.xlsx của bạn. Cột “ID” có định dạng:  
 {A}.{XXX}.{YYY}   
Trong đó: +A: Tên sách (file) PDF mà bạn đang làm

+XXX: Thứ tự trang

+YYY: Thứ tự bounding box.

A screen shot of a computer

Description automatically generated

Đây là ví dụ file ảnh:

A screenshot of a computer

Description automatically generated

Các bạn cần đảm bảo rằng, ảnh của mỗi trang sách ở đúng định dạng

{Tên sách}\_page{XXX}.png

Ví dụ: Cột “ID” trong file Excel: Bondaytrecon.001.002

🡪Tên ảnh là: Bondaytrecon\_page001.png

Lưu ý: Nếu các bạn đã có cột “ID” ở đúng định dạng như đã nêu, nhưng tên file ảnh .png khác với định dạng trên thì có thể tự chỉnh sửa hàm ***convert\_ID\_To\_png()*** trong file convert\_data\_to \_labelsPaddleđể chuyển đổi “ID” về đúng tên ảnh.

B1. Thay file result.xlsx trong thư mục **check\_label** thành file excel ngữ liệu GK của bạn

Lưu ý:

* Cột ‘Image Box’ cần dùng dấu ngoặc vuông như hình, có dạng: [[x1,y1],[x2,y2],[x3,y3],[x4,y4]] hoặc [(x1,y1),(x2,y2),(x3,y3),(x4,y4)]

B2. Copy thư mục hình ảnh cần check tọa độ vào thư mục **check\_label**, và đặt đúng tên **images\_label**

B3. Chạy file convert\_data\_to \_labelsPaddle

(Các bạn có thể chỉnh sửa code phần “#YOU CAN CHANGE HERE” để chỉnh lại tên cột phù hợp với file excel của mình)

B4: Mở công cụ PPOCRLabel và trỏ tới thư mục chứa ảnh cần kiểm tra. Như hình bên dưới, các box và text từ file result.xlsx đã được hiển thị trên file ảnh. (Lưu ý: Cần đảm bảo sau khi chạy file convert\_data\_to \_labelsPaddle sẽ tạo ra được 2 file fileState.txt và Label.txt trong folder chứa ảnh của bạn)

A screenshot of a computer

Description automatically generated

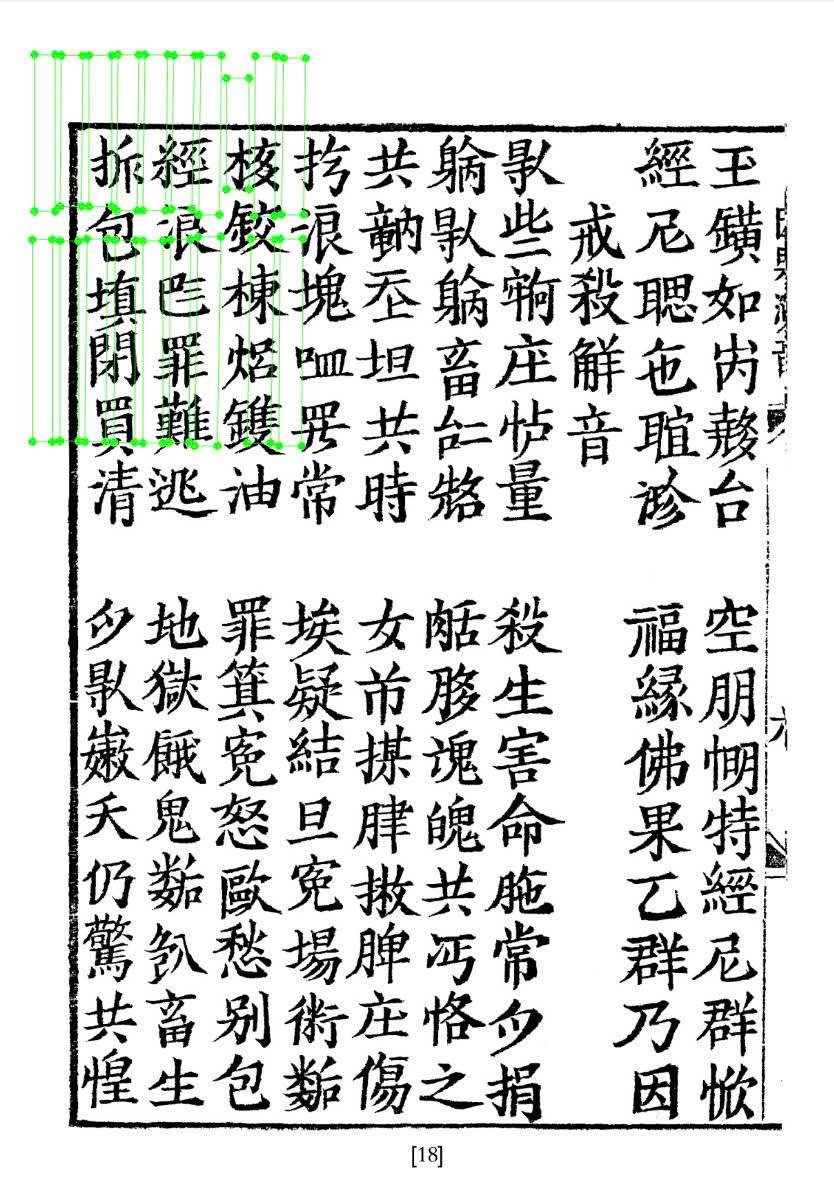
B5. Vào File → Export Recognition Result → OK để xuất label cho giai đoạn nhận dạng văn bản. Kết quả như hình dưới.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**Lưu ý nếu gặp lỗi :** Khi dùng PPOCRLabel kiểm tra box mà bị lỗi như hình dưới



Nguyên nhân là do khi sử dụng API CLC để lấy tọa độ, các hình có kích thước (dài + rộng) >1200 thì API CLC sẽ resize về 1200 trước khi gọi OCR. Do đó, Các hình gốc mà bạn có cần resize về 1200 → Chỉ cần chạy file **thunhoanh.py** (có kèm theo trong thư mục Tools) cho tất cả các ảnh của bạn.