

Cách gán nhãn sắc phong bằng PP-OCR

Mục lục

I.	Dữ liệu hiện có.....	1
1.	Dữ liệu ảnh.....	1
2.	Dữ liệu văn bản.....	1
3.	Dữ liệu văn bản đã ghép các ký tự lại và xóa dấu câu.....	2
II.	Cách gán nhãn.....	3
1.	Cài đặt PPOCR	3
2.	Chọn thư mục cần gán nhãn.....	3
3.	Gán nhãn tự động cho phát hiện và nhận diện văn bản.....	4
4.	Tải mô hình của chính mình lên như thế nào?	5
5.	Gán nhãn cho phát hiện văn bản	5
6.	Gán nhãn cho nhận dạng văn bản	6
III.	Tham khảo	8

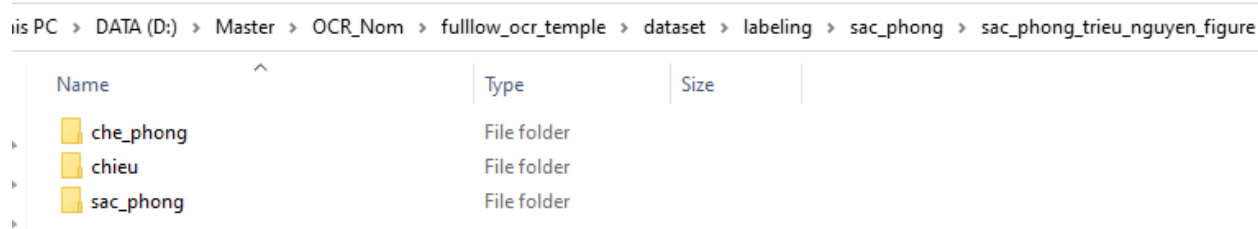
I. Dữ liệu hiện có

1. Dữ liệu ảnh

Bao gồm 3 thư mục con là:

- Chiều: 19 ảnh.
- Chế phong: 100 ảnh.
- Sắc phong: 180 ảnh.

Link drive: <https://drive.google.com/file/d/10xocehRNXFkuqAubWkvfNXqLO-8OojQe/view?usp=sharing>



PC > DATA (D:) > Master > OCR_Nom > fulllow_ocr_temple > dataset > labeling > sac_phong > sac_phong_trieu_nguyen_figure		
Name	Type	Size
che_phong	File folder	
chieu	File folder	
sac_phong	File folder	

Hình 1: Minh họa dữ liệu hình ảnh

2. Dữ liệu văn bản

Dữ liệu văn bản chứa các văn bản đã được số hóa từ các hình ảnh phân theo từng chiều, chế phong và sắc phong.

Link drive:

https://drive.google.com/file/d/1_uEXPjgzD9ZW8G4uvbhudaCwNeJI6Za/view?usp=sharing

his PC > DATA (D:) > Master > OCR_Nom > fulllow_ocr_temple > dataset > labeling > sac_phong > sac_phong_digitilization_KHOA

Name	Date modified	Type	Size
CHẾ PHONG.xlsx	11/03/2022 11:32 PM	Microsoft Excel W...	75 KB
CHIẾU.xlsx	20/10/2023 11:30 PM	Microsoft Excel W...	17 KB
SẮC PHONG.xlsx	11/03/2022 11:32 PM	Microsoft Excel W...	65 KB

Hình 2: Minh họa dữ liệu văn bản

3. Dữ liệu văn bản đã ghép các ký tự lại và xóa dấu câu
Dữ liệu chứa các văn bản được đánh số theo thứ tự của từng chế phong, sắc phong và chiếu.

This PC > DATA (D:) > Master > OCR_Nom > fulllow_ocr_temple > dataset > labeling > sac_phong > sac_phong_digitilization_ghep_ky_tu

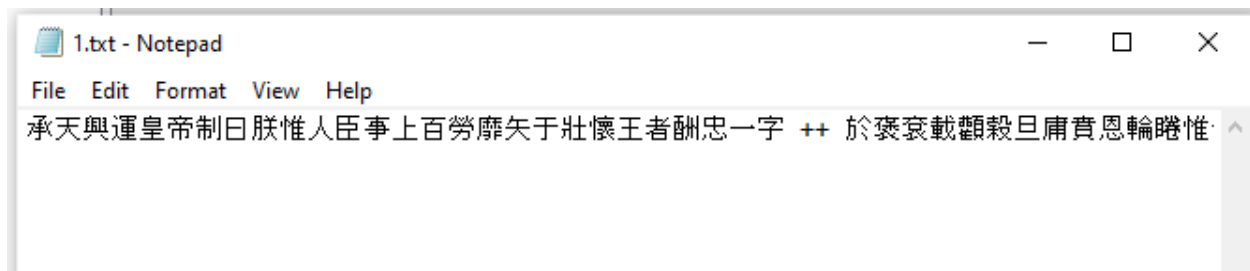
Name	Date modified	Type	Size
che_phong	27/10/2023 3:12 PM	File folder	
chieu	27/10/2023 3:25 PM	File folder	
sac_phong	27/10/2023 3:25 PM	File folder	
CHẾ PHONG.xlsx	11/03/2022 11:32 PM	Microsoft Excel W...	75 KB
CHIẾU.xlsx	20/10/2023 11:30 PM	Microsoft Excel W...	17 KB
SẮC PHONG.xlsx	11/03/2022 11:32 PM	Microsoft Excel W...	65 KB

Hình 3: Dữ liệu văn bản đã ghép các ký tự và xóa dấu câu

his PC > DATA (D:) > Master > OCR_Nom > fulllow_ocr_temple > dataset > labeling > sac_phong > sac_phong_digitilization_ghep_ky_tu > che_phong

Name	Date modified	Type	Size
1.txt	27/10/2023 3:12 PM	Text Document	1 KB
2.txt	27/10/2023 3:12 PM	Text Document	1 KB
3.txt	27/10/2023 3:12 PM	Text Document	1 KB
4.txt	27/10/2023 3:12 PM	Text Document	1 KB
5.txt	27/10/2023 3:12 PM	Text Document	1 KB
6.txt	27/10/2023 3:12 PM	Text Document	1 KB
7.txt	27/10/2023 3:12 PM	Text Document	1 KB
8.txt	27/10/2023 3:12 PM	Text Document	1 KB
9.txt	27/10/2023 3:12 PM	Text Document	1 KB
10.txt	27/10/2023 3:12 PM	Text Document	1 KB
11.txt	27/10/2023 3:12 PM	Text Document	1 KB
12.txt	27/10/2023 3:12 PM	Text Document	1 KB
13.txt	27/10/2023 3:12 PM	Text Document	1 KB

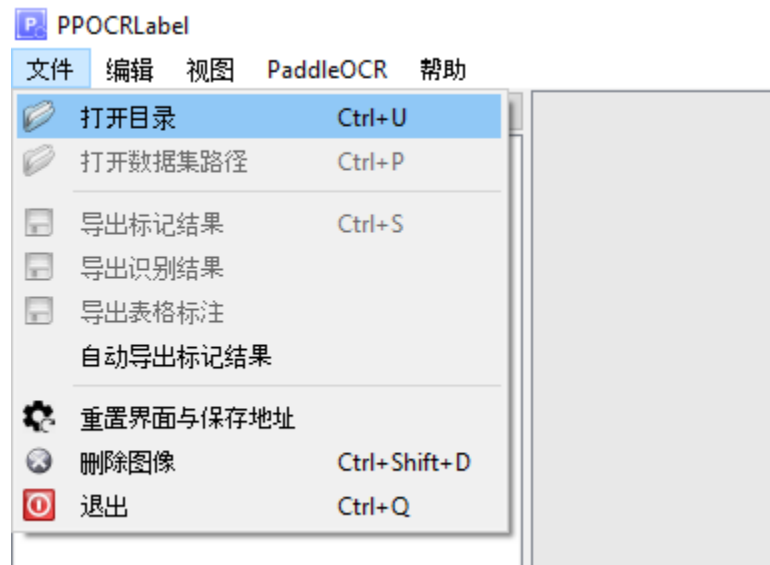
Hình 4: Trong thư mục “Chế phong” bao gồm các tập tin .txt chứa từng văn bản





Hình 5: Nội dung tập tin “1.txt” trong thư mục “ché phong”

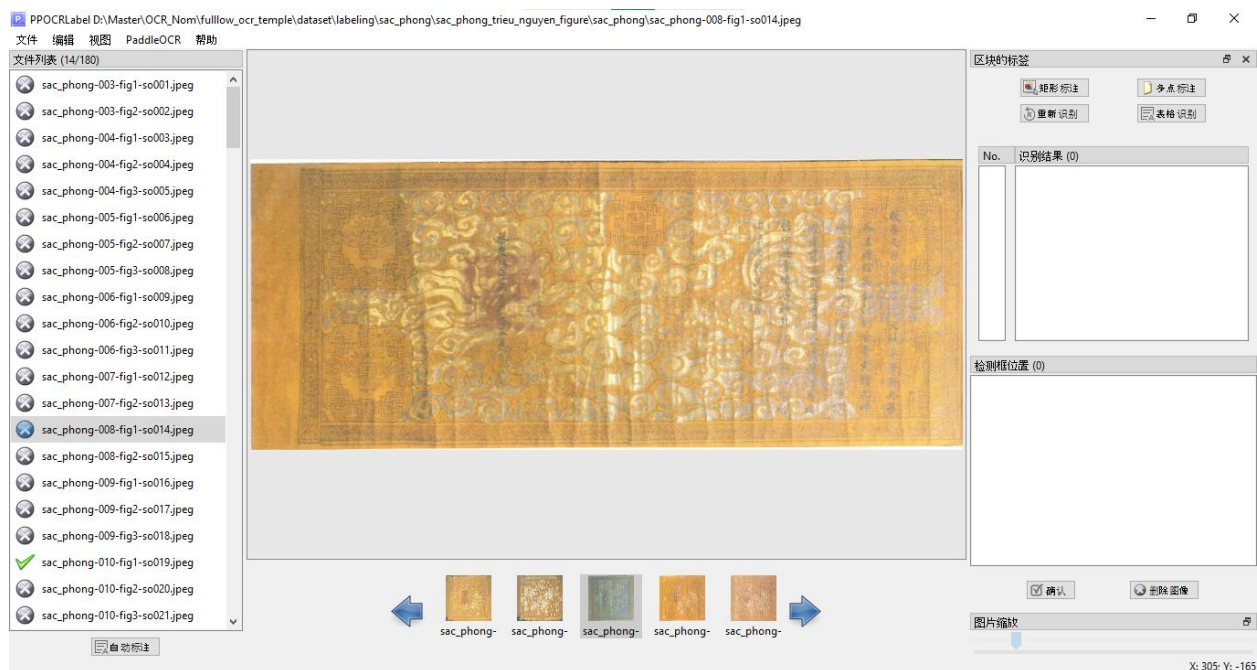
II. Cách gán nhãn

1. Cài đặt PPOCR
 - Để cài đặt thư viện PPOCRLabel ta dùng Pip:
`pip install PPOCRLabel`
 - Để kích hoạt PPOCRLabel, ta sử dụng lệnh:
`PPOCRLabel -lang ch`
2. Chọn thư mục cần gán nhãn

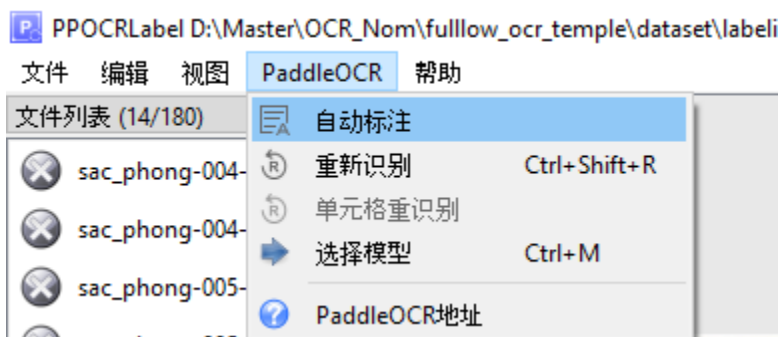


Hình 6: Chọn thư mục cần gán nhãn

Sau khi chọn thư mục “Sắc phong”, giao diện sẽ hiển thị các ảnh trong thư mục. Các ảnh chưa được gán nhãn sẽ được ký hiệu là  và ảnh đã được gán có ý hiệu là .

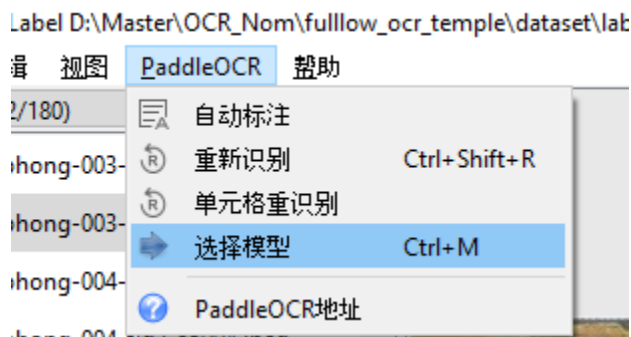


3. Gán nhãn tự động cho phát hiện và nhận diện văn bản.
Để gán nhãn tự động văn bản, chúng ta chọn tab PaddleOCR và chọn mục đầu tiên.



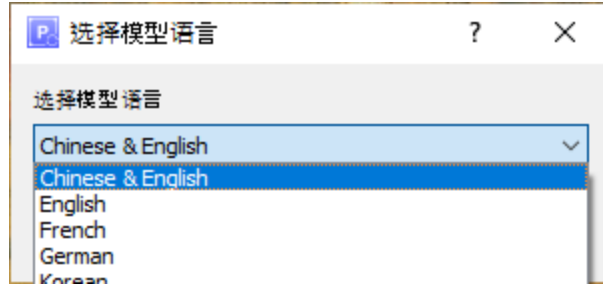
Hình 7: Nhận diện tự động

Chọn các mô hình nhận diện ký tự khác bằng cách:



Hình 8: Chọn mô hình nhận diện tự động

Các mô hình có thể chọn bao gồm: Chinese & English, English, Frace, German, Korean, Japanese.

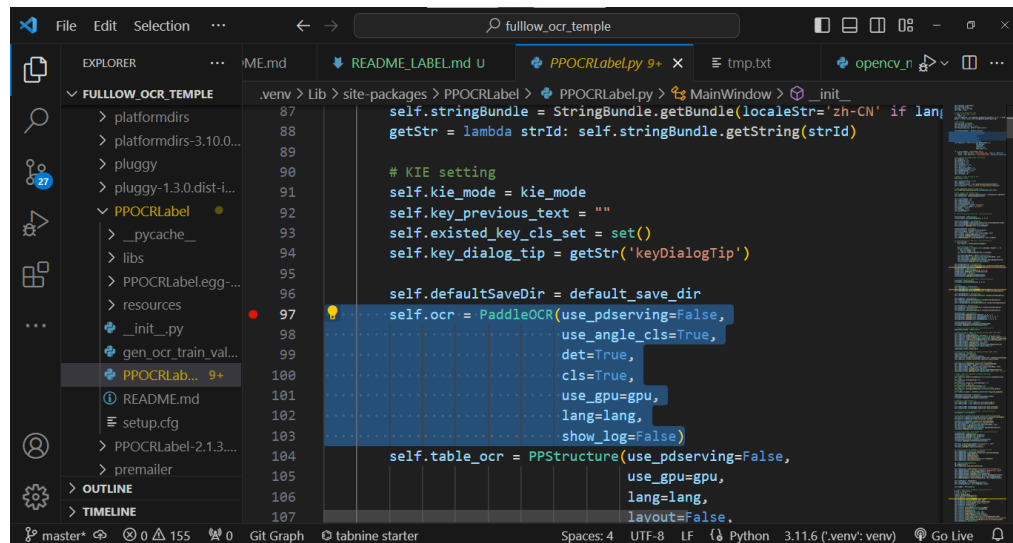


Hình 9: Các mô hình có thể chọn để nhận diện ký tự

4. Tải mô hình của chính mình lên như thế nào?

Do **PPOCRLabel** đã được đóng gói thành thư viện. Nên chúng ta có thể tìm code trong môi trường ảo để thêm đường dẫn vào mô hình.

Tại dòng 97 của `../PPOCRLabel/PPOCRLabel.py` chứa code để tải mô hình nhận diện và phát hiện văn bản.



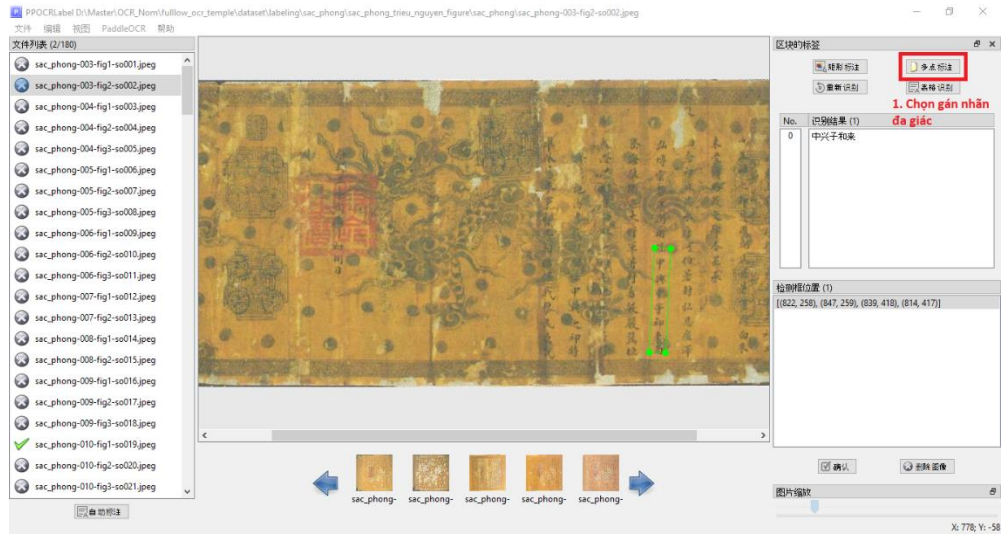
Hình 10: Mã nguồn gọi mô hình phát hiện và nhận dạng ký tự

Để thêm mô hình mới vào, chúng ta sẽ thêm các tham số vào khởi tạo `PaddleOCR()`. Cụ thể cần thêm những tham số sau:

- `det_model_dir`: Đường dẫn đến thư mục chứa checkpoint của mô hình phát hiện văn bản.
- `rec_model_dir`: Đường dẫn đến thư mục chứa checkpoint của mô hình nhận diện văn bản.
- `rec_char_dict_path`: Đường dẫn đến file chứa các ký tự cần nhận diện.
- `cls_model_dir`: Đường dẫn đến mô hình phân lớp hướng của của bounding box (0° , 180°)

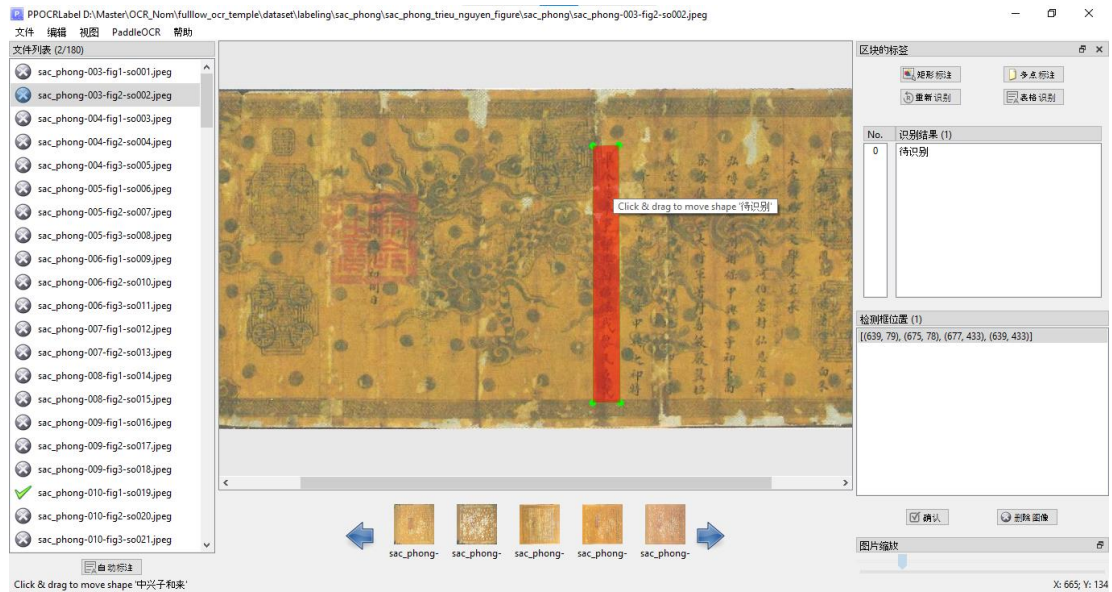
5. Gán nhãn cho phát hiện văn bản

Chọn nút “Gán nhãn đa giác”.



Hình 11: Chọn gán nhãn đa giác

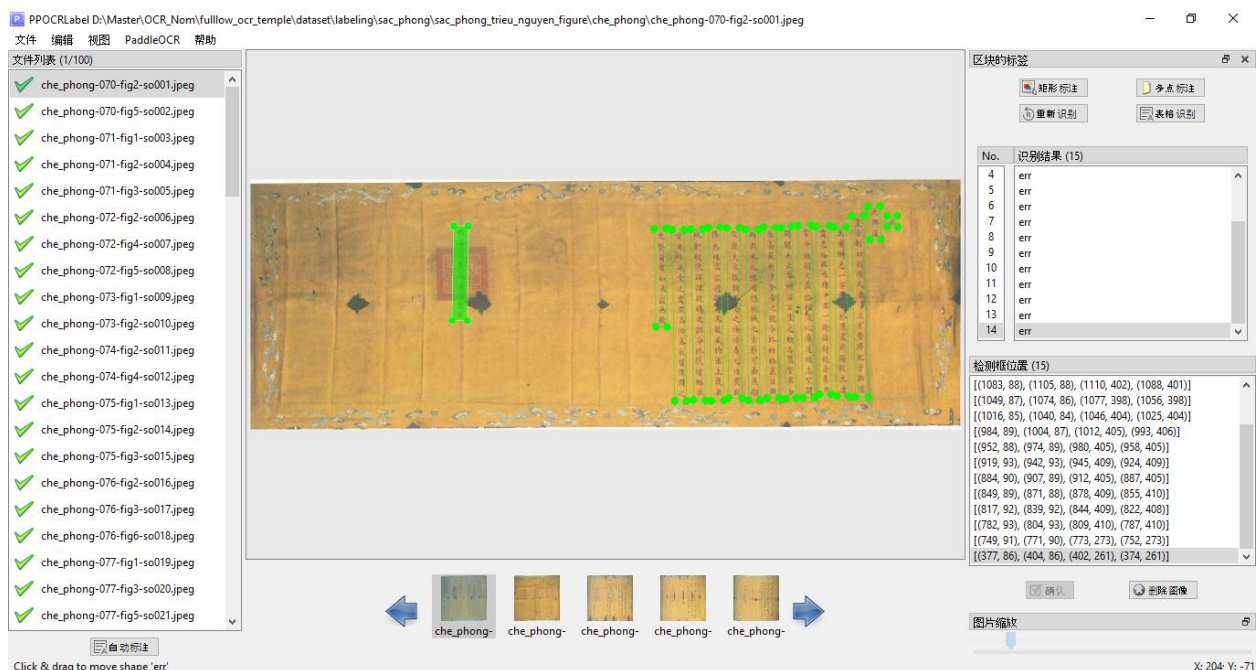
Sau đó vẽ 4 điểm để tạo thành một đa giác bao quanh vùng cần gán nhãn.



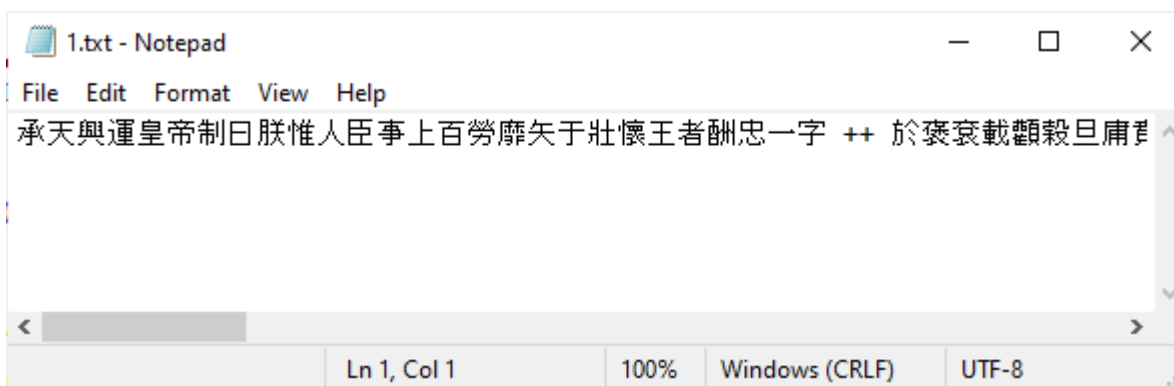
Hình 12: Chọn 4 điểm bao quanh vùng cần gán nhãn để tạo thành một bounding box

Để phóng to hình ảnh, nhấn dè Ctrl + lăn con chuột lên xuống.

6. Gán nhãn cho nhận dạng văn bản
Hình ảnh sắc phong đã gán nhãn phần phát hiện văn bản.

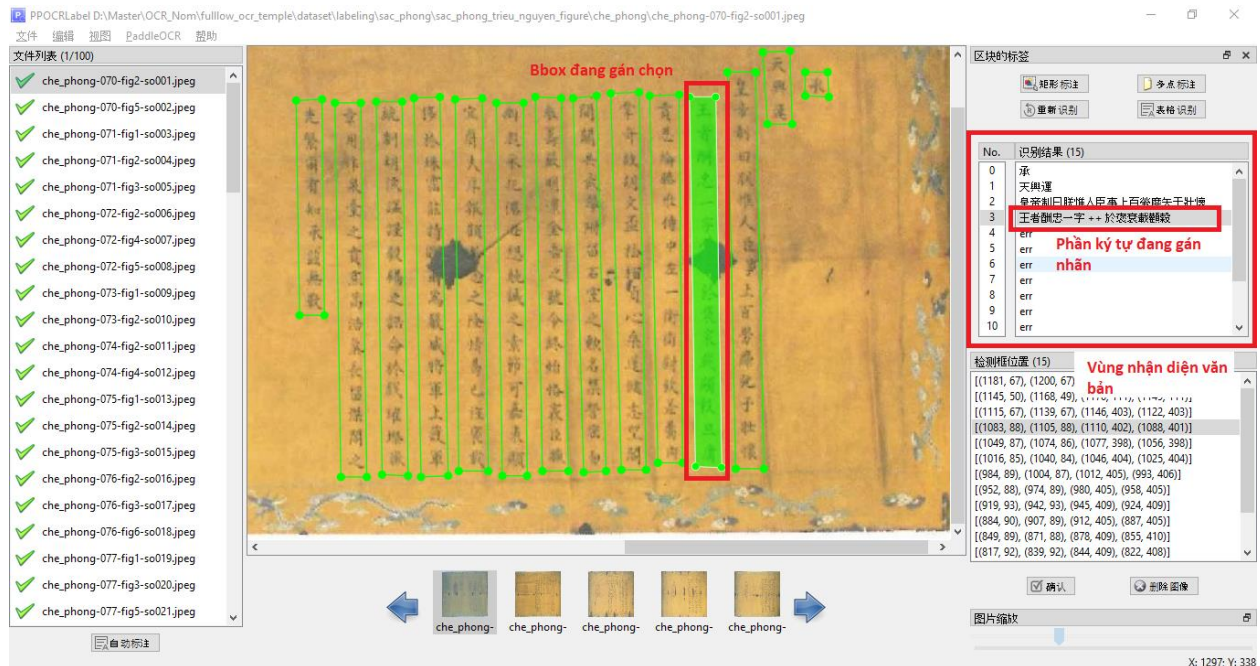


Hình 13: Sắc phong đã gán nhãn phân phát hiện văn bản



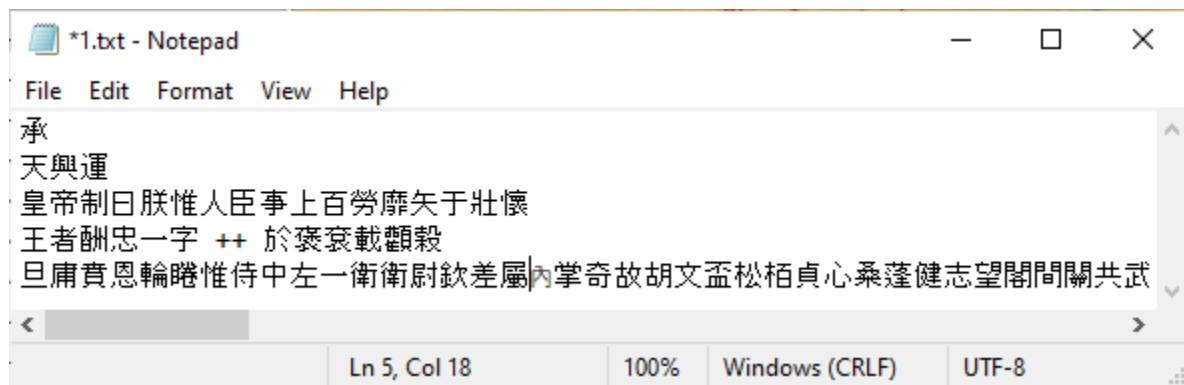
Hình 14: Phần văn bản cần bỏ vào

Cách làm hiện tại tôi sẽ đến số lượng ký tự trong mỗi bounding box và lấy số chữ tương ứng để bỏ vào phần nhận diện văn bản.



Mẹo khi gán nhãn nhận dạng văn bản.

Chúng ta có thể dựa vào vị trí con trỏ trong notepad để biết được trước đó có bao nhiêu ký tự. Ví dụ trong hình bên dưới, vị trí sẽ là “Ln 5, Col 18” sẽ cho biết dòng nằm trước đó đang có 17 ký tự (vì con trỏ đang ở vị trí 18). Để phân biệt các dòng chúng ta có thể xuống hàng để dễ quan sát hơn.



III. Tham khảo

<https://github.com/Evezerest/PPOCRLabel>