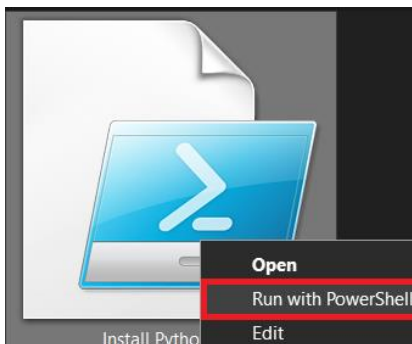


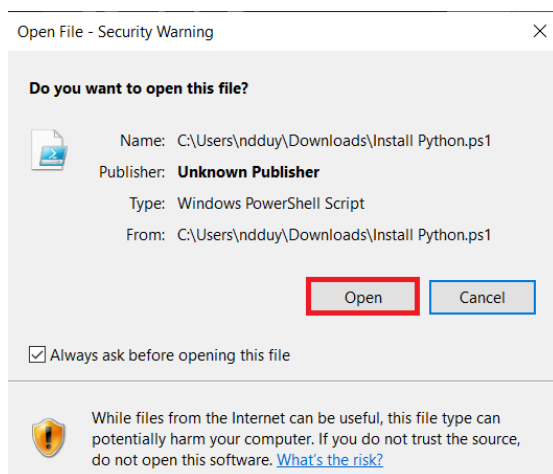
# HƯỚNG DẪN CÀI ĐẶT TOOL VÀ GÁN NHÃN DỮ LIỆU CHO CÁC TÁC PHẨM THƠ

## 1. Hướng dẫn cài đặt tool PPOCRLLabel

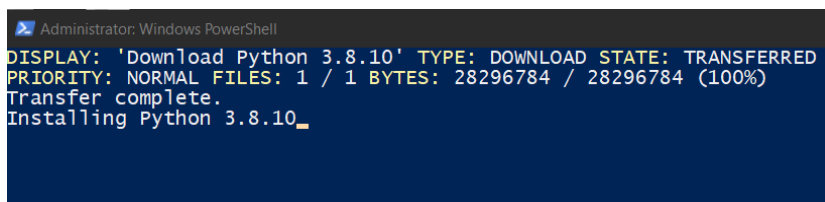
- **Bước 1:** Tải file [InstallPython.ps1](#).
- **Bước 2:** Nhấn chuột phải vào file chọn “Run with PowerShell”.



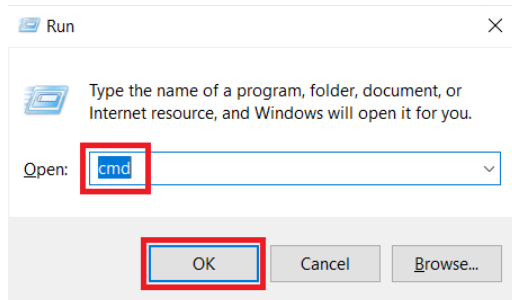
- **Bước 3:** Chọn “Open”.



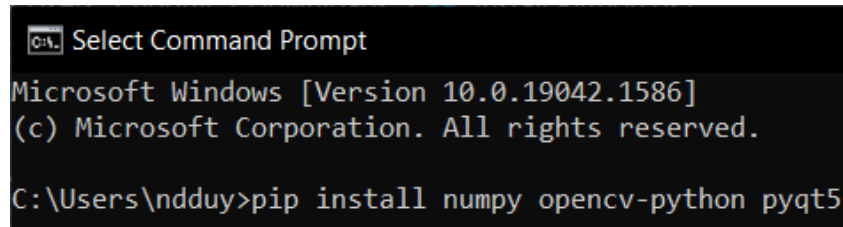
- **Bước 4:** Có cửa sổ thông báo hiện lên chọn “Yes” sau đó đợi đến khi cửa sổ như hình đóng.



- **Bước 5:** Mở “cmd” bằng tổ hợp phím “Window + R” sau đó gõ cmd rồi nhấn “OK”.

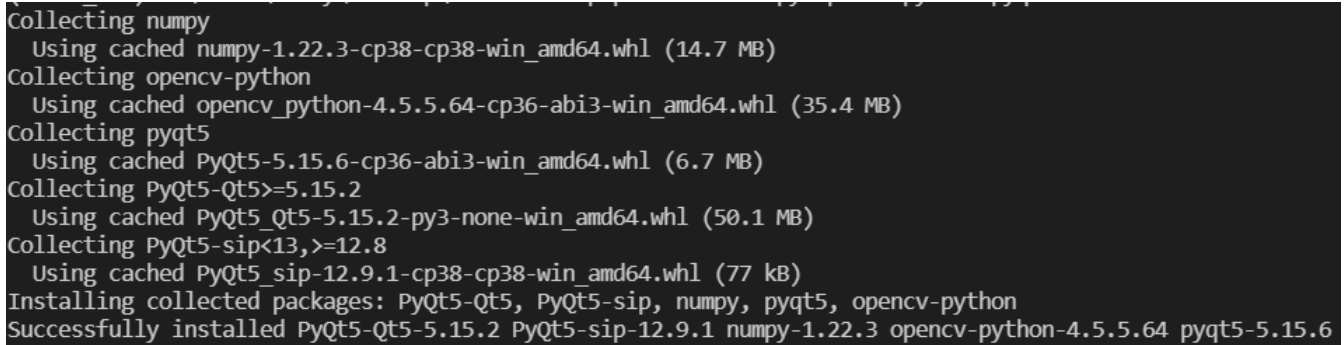


- **Bước 6:** Gõ “pip install numpy opencv-python pyqt5” rồi nhấn Enter.



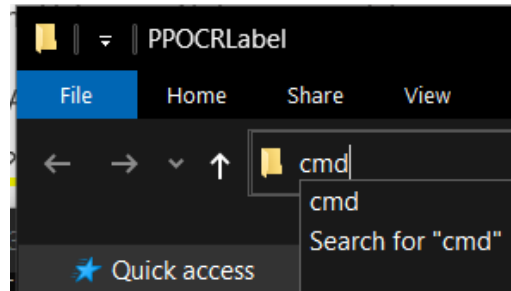
```
Select Command Prompt
Microsoft Windows [Version 10.0.19042.1586]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ndduy>pip install numpy opencv-python pyqt5
```



```
Collecting numpy
  Using cached numpy-1.22.3-cp38-cp38-win_amd64.whl (14.7 MB)
Collecting opencv-python
  Using cached opencv_python-4.5.5.64-cp36-abi3-win_amd64.whl (35.4 MB)
Collecting pyqt5
  Using cached PyQt5-5.15.6-cp36-abi3-win_amd64.whl (6.7 MB)
Collecting PyQt5-Qt5>=5.15.2
  Using cached PyQt5_Qt5-5.15.2-py3-none-win_amd64.whl (50.1 MB)
Collecting PyQt5-sip<13,>=12.8
  Using cached PyQt5_sip-12.9.1-cp38-cp38-win_amd64.whl (77 kB)
Installing collected packages: PyQt5-Qt5, PyQt5-sip, numpy, pyqt5, opencv-python
Successfully installed PyQt5-Qt5-5.15.2 PyQt5-sip-12.9.1 numpy-1.22.3 opencv-python-4.5.5.64 pyqt5-5.15.6
```

- **Bước 7:** Tải tool theo [tài đây](#), giải nén ra, sau đó trên thanh tìm kiếm gõ “cmd” và nhấn Enter.

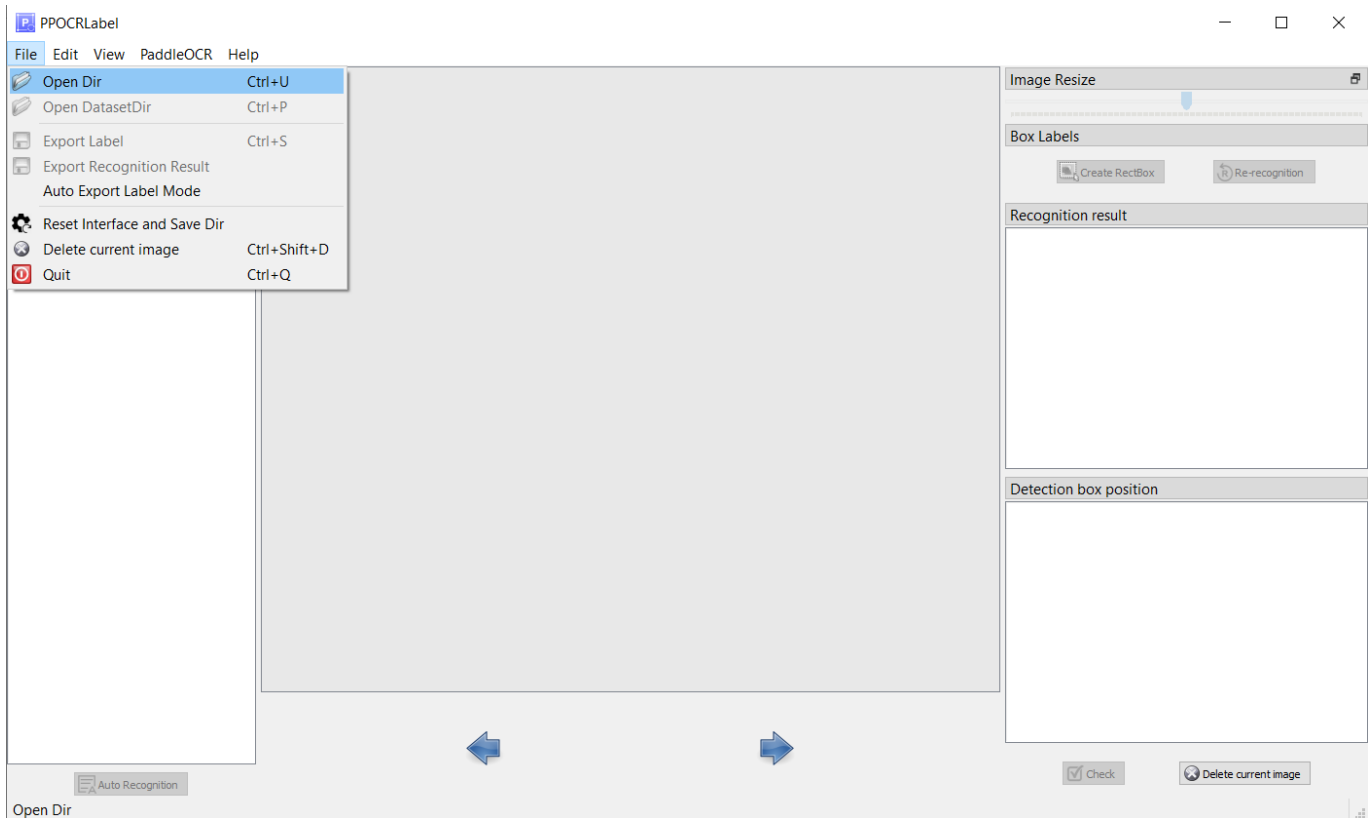


- **Bước 8:** Để khởi động tool gõ “python PPOCRLabel.py” như hình là thành công.



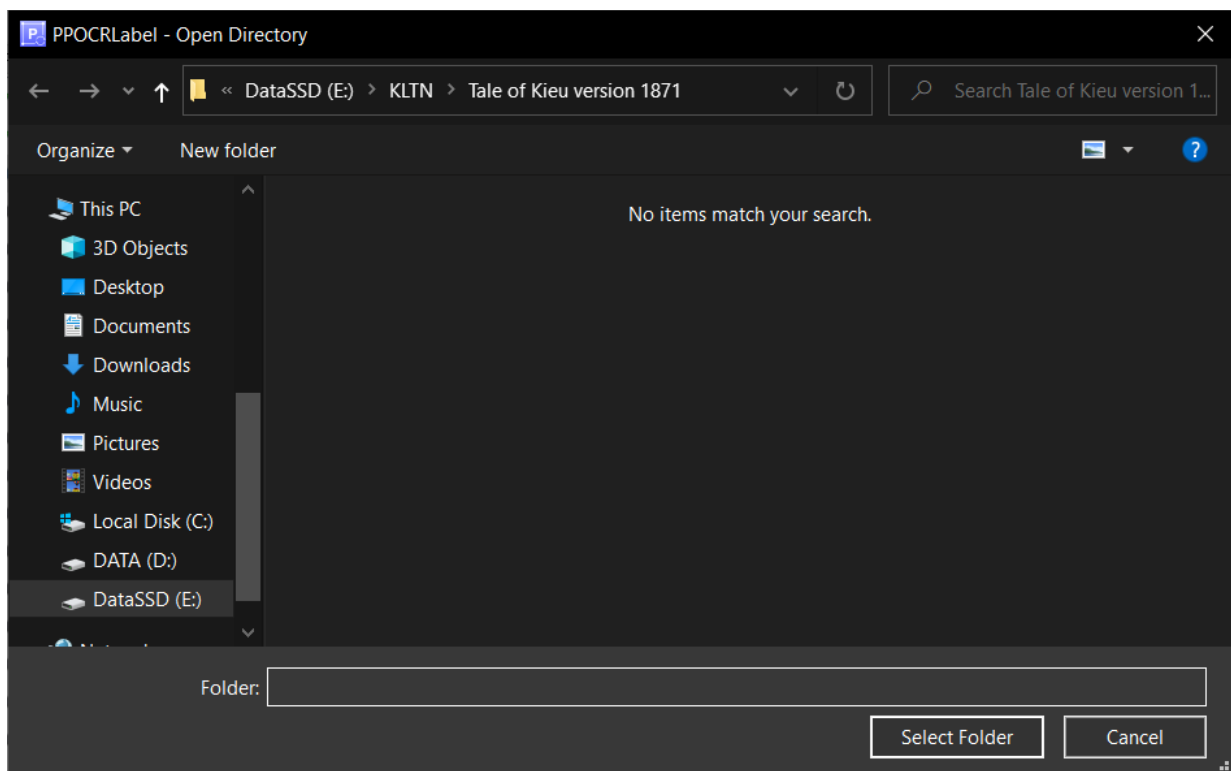
## 2. Hướng dẫn sử dụng tool gán nhãn PPOCRLabel

- **Bước 1:** Mở thư mục chứa dữ liệu, Click File -> Open Dir.

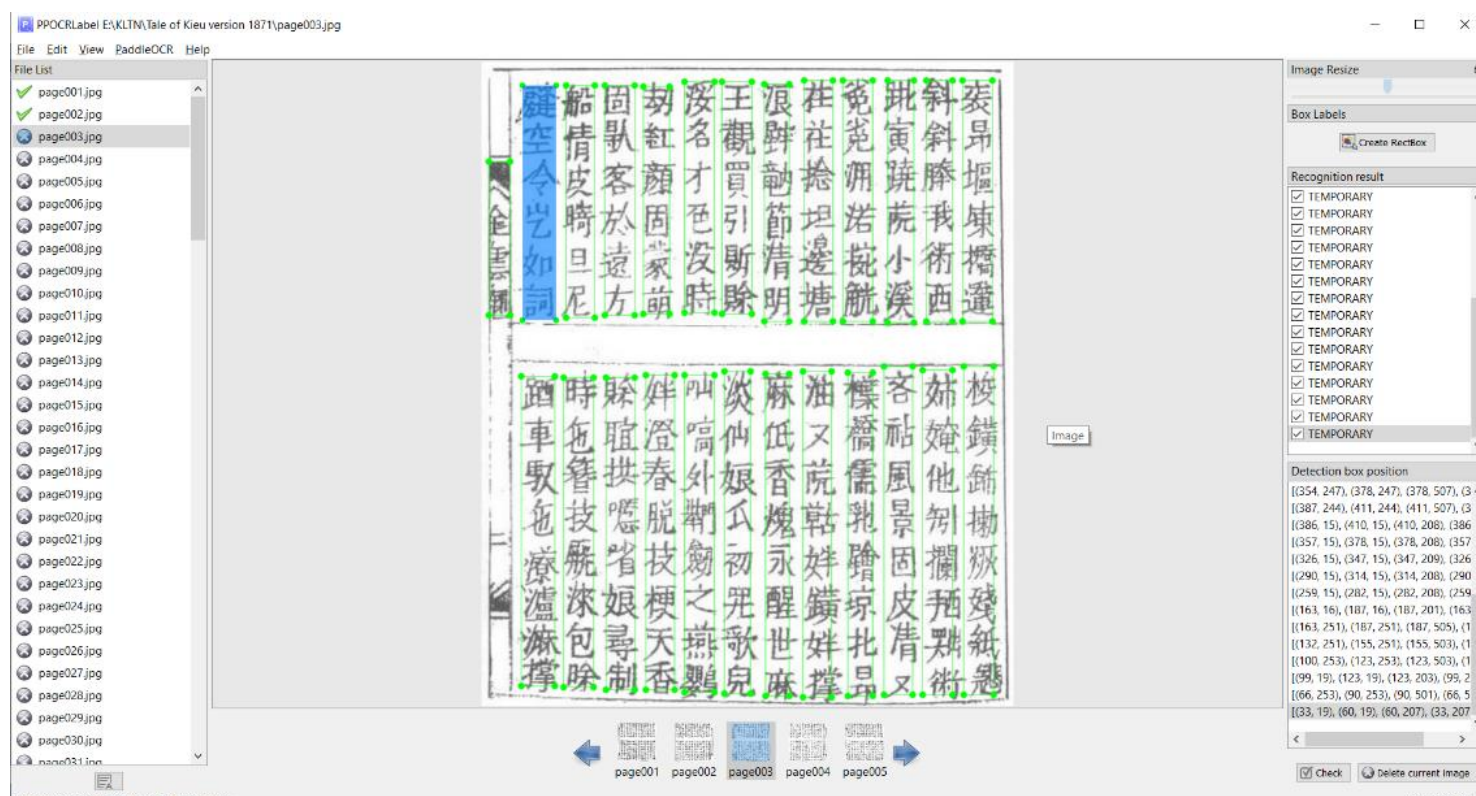


- **Bước 2:** Chọn đến thư mục chứa dữ liệu, nhấn “Select Folder”

- Trong trường hợp này thư mục chứa dữ liệu là “Tale of Kieu version 1871”



- Sau khi chọn đúng thư mục chứa dữ liệu ta sẽ được như hình:



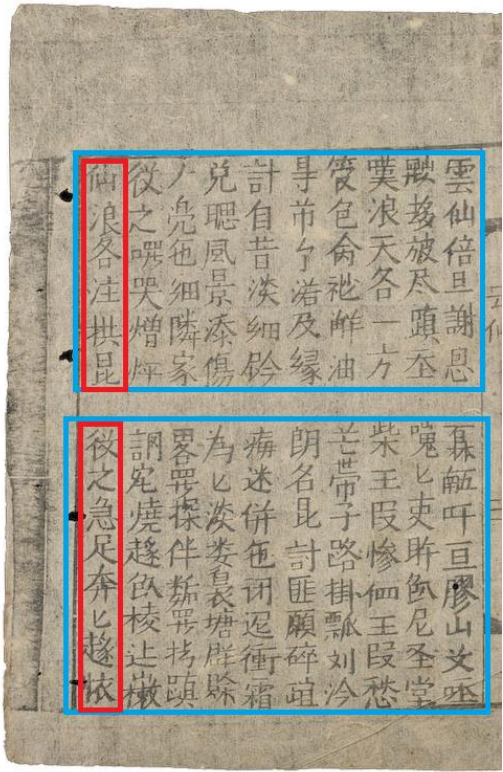
## - Các phím tắt:

- Bouding Box: các hình chữ nhật màu xanh trong hình trên được gọi là Bouding Box.

Chức năng	Phím tắt	Cách sử dụng
Kéo Bouding Box	W	Nhấn phím W. Sau khi thực hiện chức năng này không hủy được, phải kéo đại Bouding Box rồi dùng chức năng xóa để xóa Bouding Box đó.
Xóa Bouding Box	Backspace	Chọn Bouding Box (nhấn vào giữa Bouding Box) cần xóa, sau đó nhấn Backspace.
Chỉnh sửa Bouding Box		Kéo các ô tròn ở góc các Bouding Box
Check	Ctrl + V	
Quay lại thao tác trước	Ctrl + Z	
Zoom in	Ctrl + lăn chuột lên	
Zoom out	Ctrl + lăn chuột xuống	

3. Tiến hành gán nhãn dữ liệu

- Lưu ý:
  - Các Bouding Box sau khi chỉnh sửa không được đè vào nhau.
  - Để dễ gán hơn nên Zoom in vào cho dễ xem.
  - Thứ tự đọc của các tác phẩm được sử dụng là trên xuống dưới, phải qua trái.
  - Các câu trong 2 tác phẩm Lục Vân Tiên và Truyện Kiều đều có 2 phần: 6 từ/kí tự cho phần trên và 8 từ/kí tự cho phần dưới.



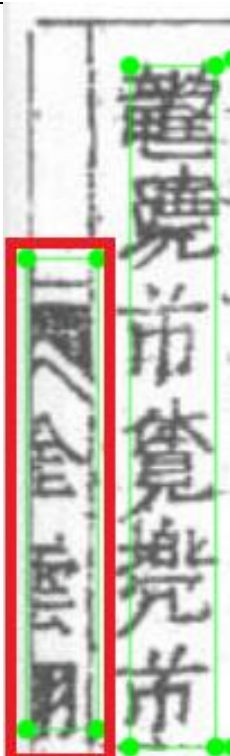
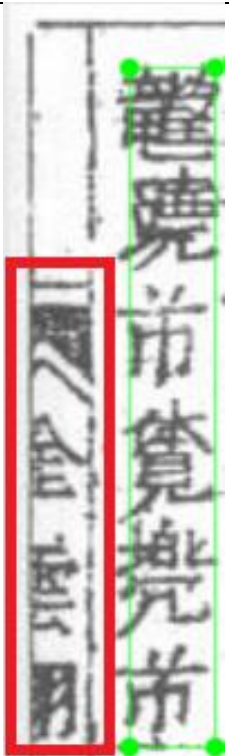
- Bước 1: Kiểm tra từng Bouding Box.
  - Trường hợp 1: thiếu Bouding Box.

Vấn đề	Xử lý	Kết quả
	Nhấn W rồi kéo khung sao cho vừa đủ chữ.	

- Trường hợp 2: Bouding Box chưa đủ phủ hết chữ, hoặc phủ dư quá nhiều.

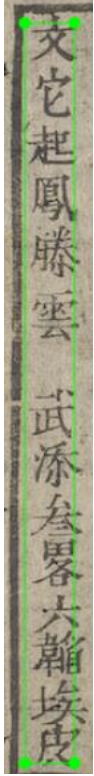

Vấn đề	Xử lý	Kết quả
 The image shows a vertical calligraphy strip with two columns of text. A green bounding box is drawn around the text, but it is too narrow on the left side, failing to cover the first column of characters.	<p>Ta có thể thấy bên phải không đủ phủ hết chữ.</p> <p>Ta chỉ cần nắm kéo ô tròn ở dưới/trên và bên phải, sang bên phải.</p>	 The image shows the same calligraphy strip after adjustment. The green bounding box has been shifted to the right, now correctly encompassing both columns of text.

- Trường hợp 3: các Bouding Box nằm ở ngoài văn bản chính.



Vấn đề	Xử lý	Kết quả
 The image shows a vertical calligraphy strip with two columns of text. A green bounding box is drawn around the text. A red rectangle highlights a separate bounding box on the left side, which is positioned outside the main text area.	Thực hiện chức năng xóa.	 The image shows the same calligraphy strip after adjustment. The red rectangle, which was previously outside the text, has been removed, leaving only the green bounding box around the text.




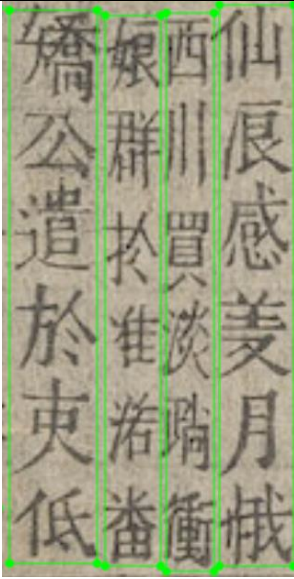
- Trường hợp 4: Bouding Box bọc cả 2 câu.

Vấn đề	Xử lý	Kết quả
	Từ 1 Box, ta kéo thành 2 Bouding Box như kết quả	

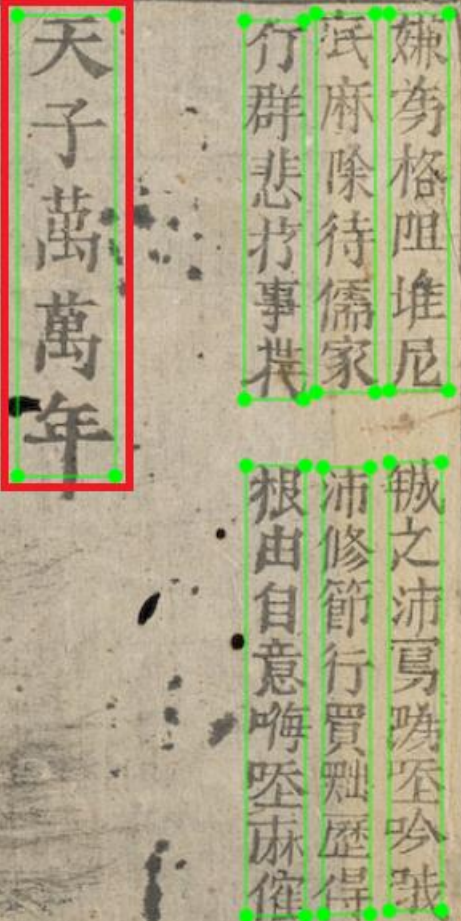
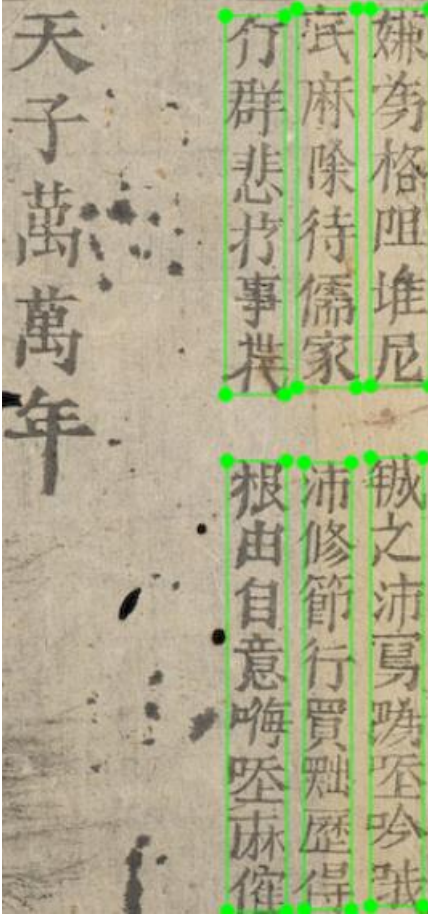
- Trường hợp 5: trong Bouding Box có vết mực hoặc rách.

Vấn đề	Xử lý	Kết quả
	Trong Box có vết mực, ta chỉ cần kéo cho vừa đủ chữ không cần kéo hết vết mực.	

- Trường hợp 6: các chữ nhỏ hơn các chữ khác trong ảnh.

Vấn đề	Xử lý	Kết quả
	Các chữ nhỏ hơn so với các chữ khác và khoảng cách giữa các câu thì kéo vừa đủ, một số nét kéo dài không cần phải kéo hết để tránh trường hợp các Bouding Box bị đè vào nhau.	

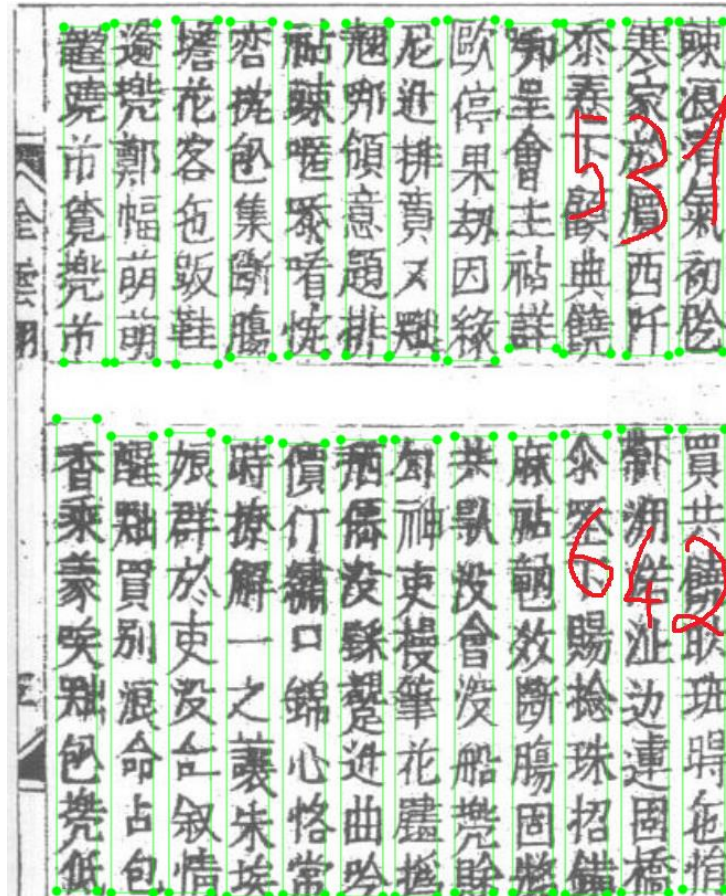
- Trường hợp 7: các chữ to hơn so với các chữ khác và số lượng không đúng theo định dạng là 6 từ/ kí tự ở đoạn trên và 8 từ/kí tự ở đoạn dưới.

Vấn đề	Xử lý	Kết quả
	Thực hiện chức năng xóa box.	

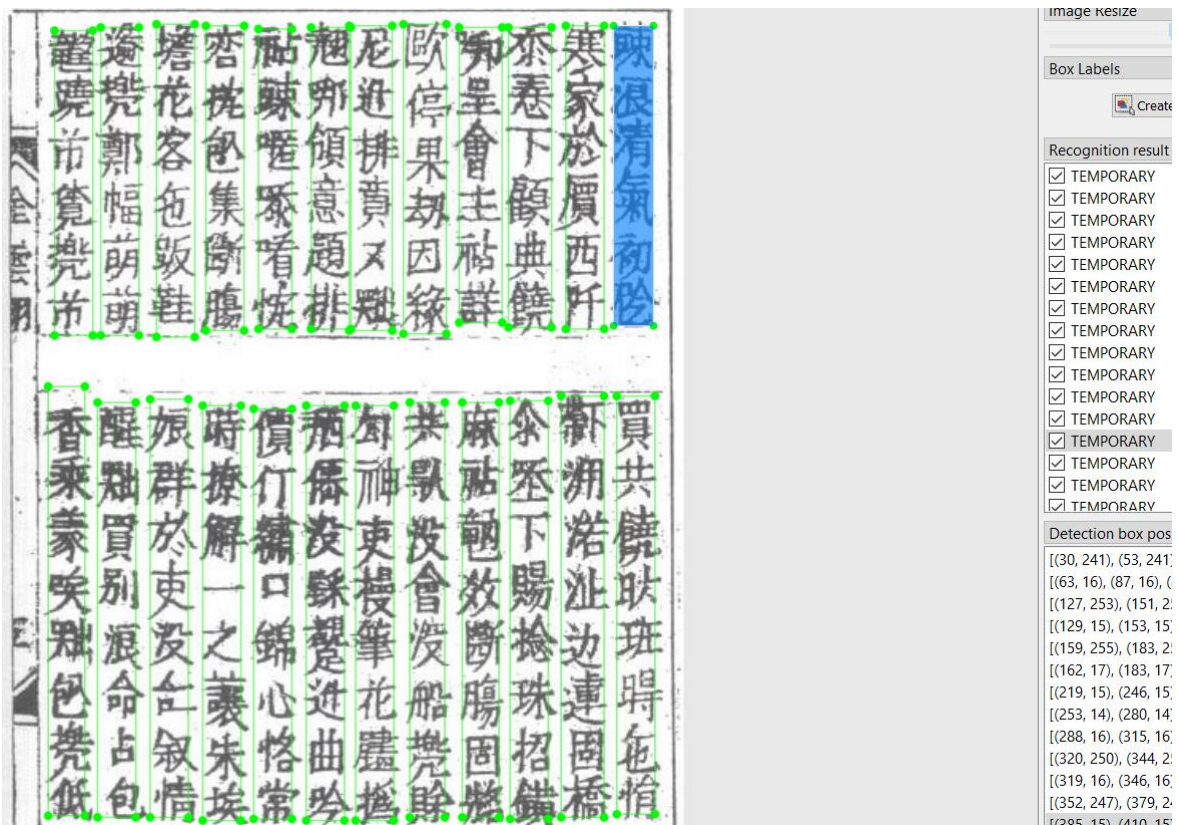


- **Bước 2:** Gán nhãn cho từng Bouding Box

- Thứ tự gán nhãn: Gán nhãn các Bouding Box từ trên xuống dưới và từ phải sang trái như hình.



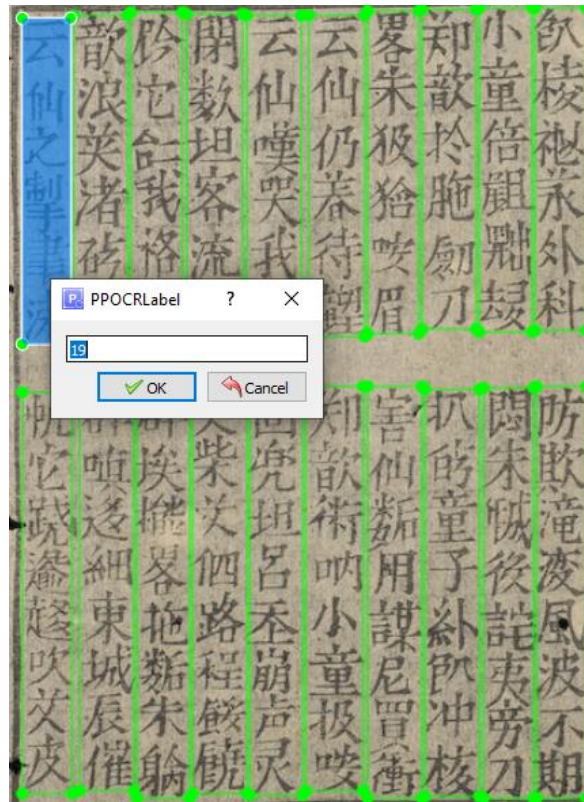
- Đầu tiên chọn vào Bouding Box cần gán nhãn.





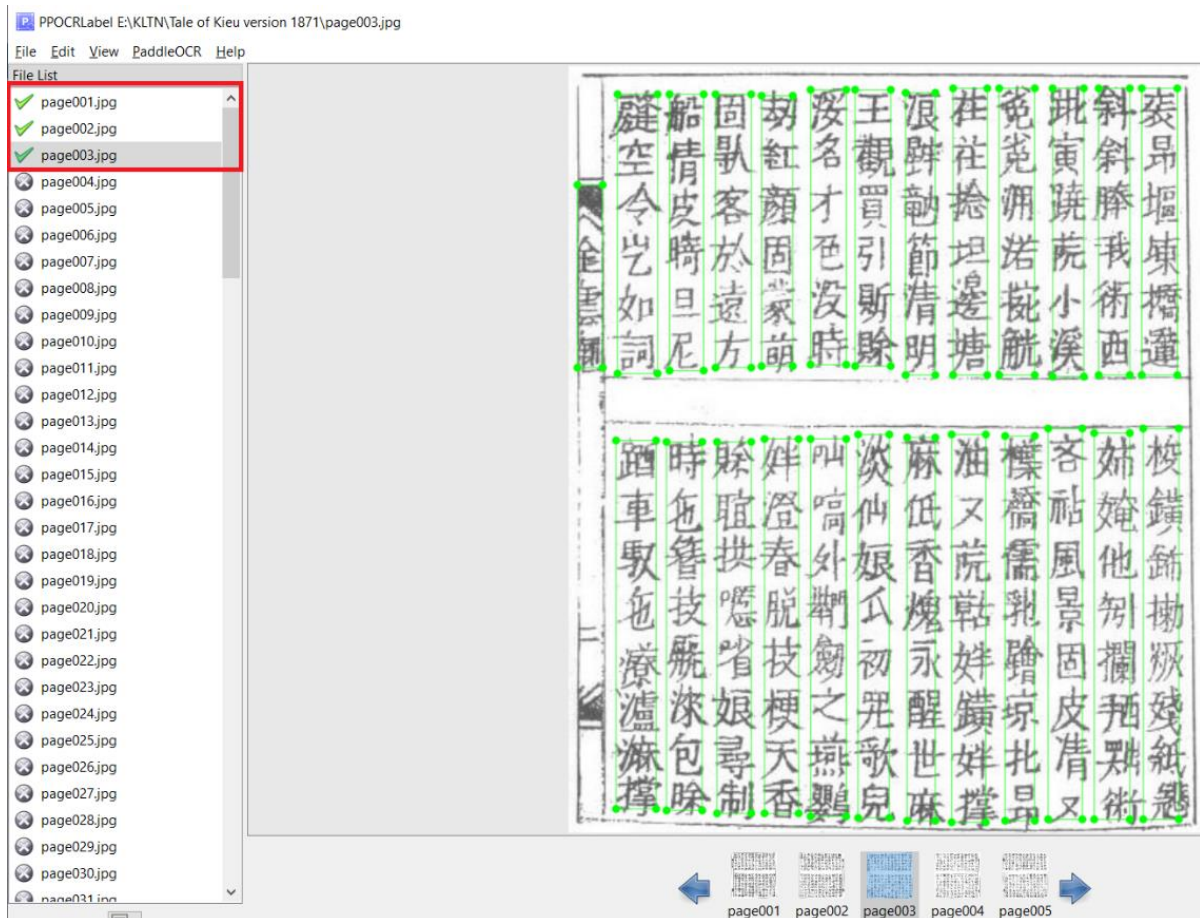


- Sau đó nhận nhãn là số thứ tự của câu rồi click “Ok” hoặc nhấn enter.



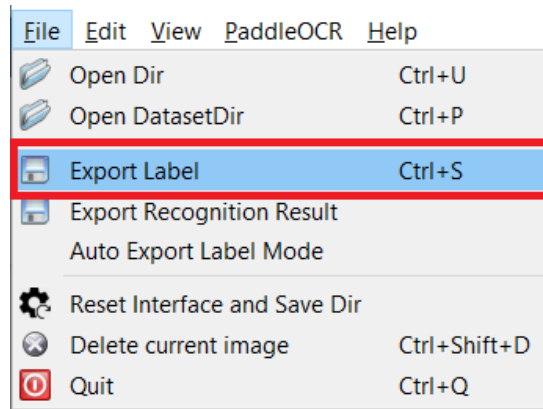
- **Bước 3:** Sau khi gán nhãn hết các Bouding Box ở trên ta thực hiện chức năng “check”.

- Sau khi check xong thì sẽ có dấu tích màu xanh như hình.





- **Bước 4:** Sau khi check hết tất cả các hình trong bộ dữ liệu ta thực hiện xuất file label.



- **Các lưu ý khi sử dụng tool:**

- Đối với những bạn sử dụng Telex thì phải tắt chế độ viết tiếng Việt.
- Sau khi chỉnh sửa và gán nhãn các Bouding Box thì phải “Check” cho mỗi hình. Hoàn thành hình hiện tại sau đó mới được chuyển hình khác.
- Mọi thắc mắc về tool cũng như các trường hợp nằm ngoài các trường hợp trên xin liên hệ theo thông tin bên dưới:

Facebook: [Quân Đặng](#)

Gmail: [18520339@gm.uit.edu.vn](mailto:18520339@gm.uit.edu.vn)

Facebook: [Nguyễn Đức Duy Anh](#)

Gmail: [18520455@gm.uit.edu.vn](mailto:18520455@gm.uit.edu.vn)

## 4. Tham khảo

- **Mã nguồn gốc PPOCRLabel:**

- Từ PaddleOCR: <https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.4/PPOCRLabel>
- Từ Repo riêng: <https://github.com/Evezerest/PPOCRLabel>

- **Nguồn lấy dữ liệu:**

- Truyện Kiều: <http://www.nomfoundation.org/nom-project/Tale-of-Kieu>
- Lục Vân Tiên: <https://www.nomfoundation.org/nom-project/Luc-Van-Tien>
- Đại Việt sử kí: <https://www.nomfoundation.org/nom-project/History-of-Greater-Vietnam>