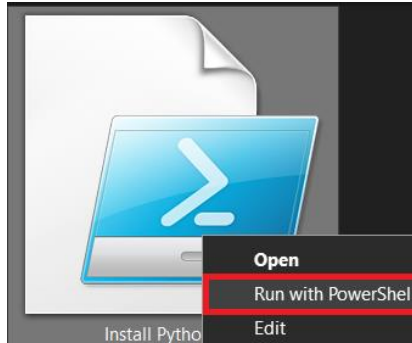


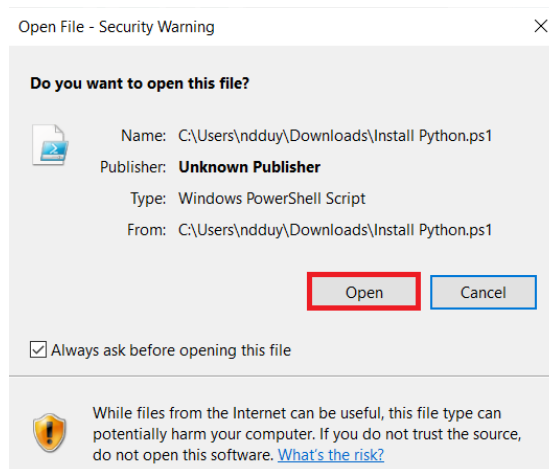
HƯỚNG DẪN CÀI ĐẶT TOOL VÀ GÁN NHÃN DỮ LIỆU CHO CÁC TÁC PHẨM VĂN XUÔI

1. Hướng dẫn cài đặt tool PPOCRLLabel

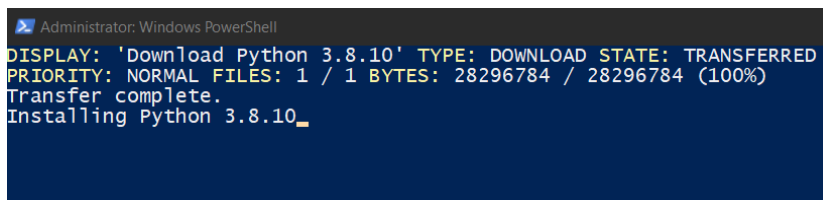
- **Bước 1:** Tải file [InstallPython.ps1](#).
- **Bước 2:** Nhấn chuột phải vào file chọn “Run with PowerShell”.



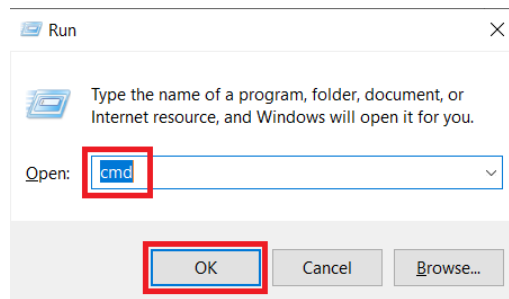
- **Bước 3:** Chọn “Open”.



- **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”.



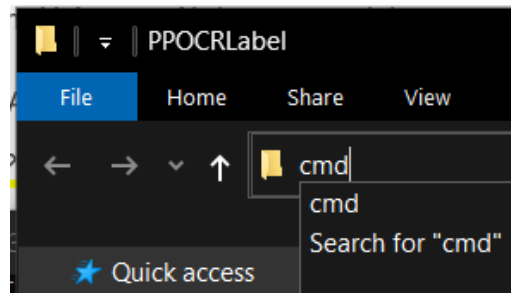
- **Bước 6:** Gõ “pip install numpy opencv-python pyqt5” rồi nhấn Enter.

```
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 [tại đây](#), giải nén ra, sau đó trên thanh tìm kiếm gõ “cmd” và nhấn Enter.

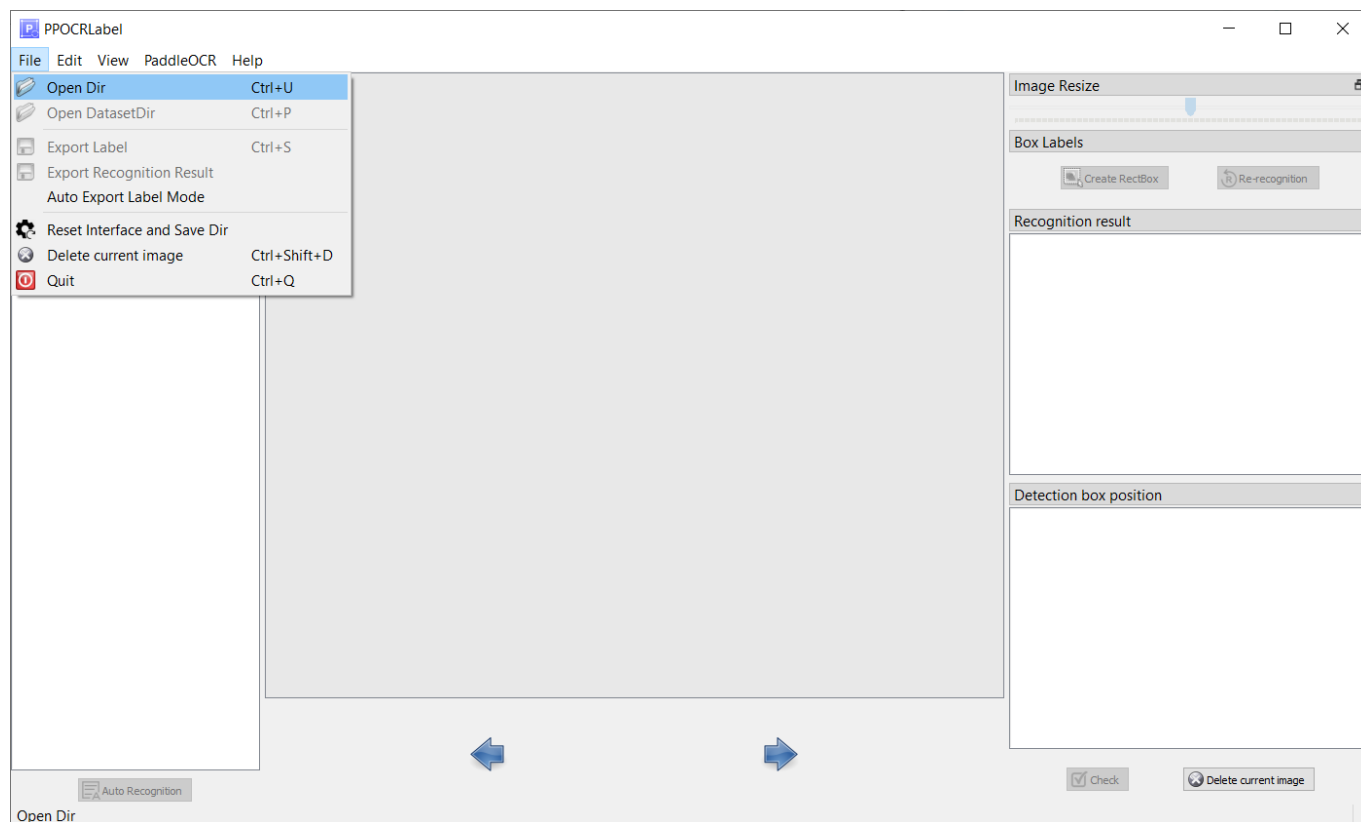


- **Bước 8:** Để khởi động tool gõ “python PPOCRLabel.py” như hình là thành công.



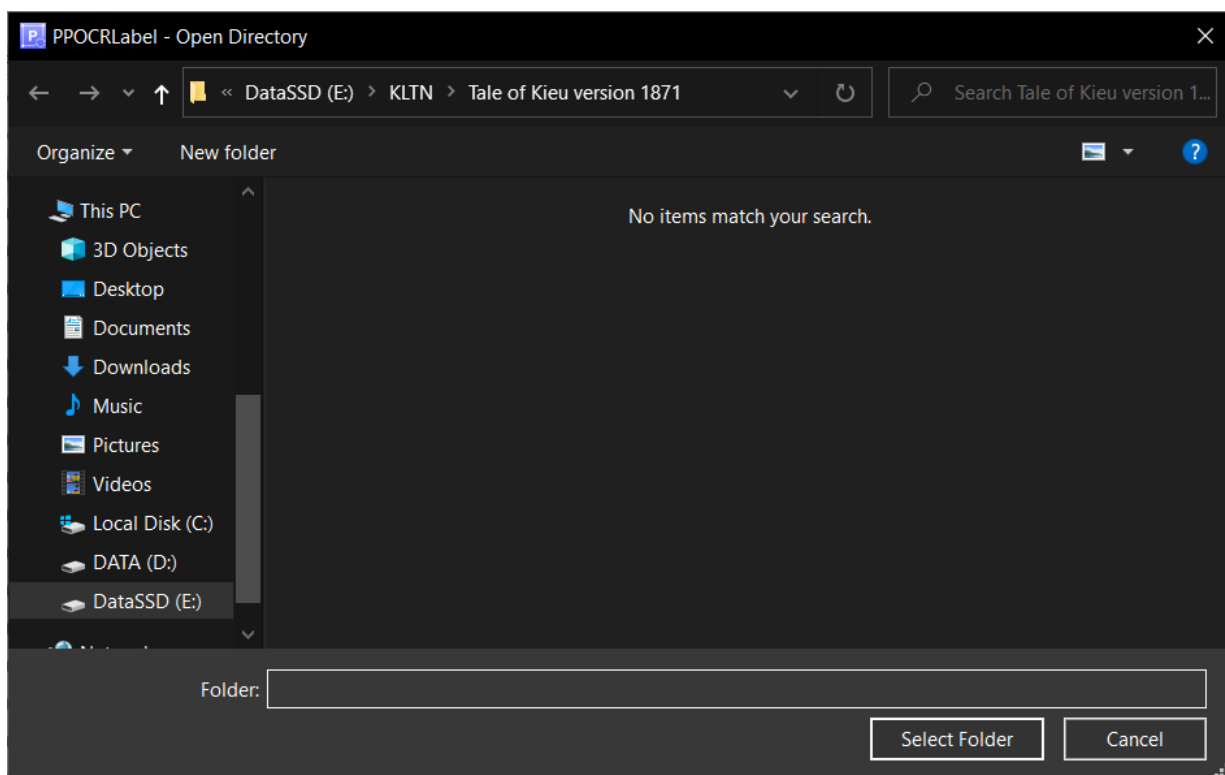
2. Hướng dẫn sử dụng tool gắn nhãn PPOCRLabel

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

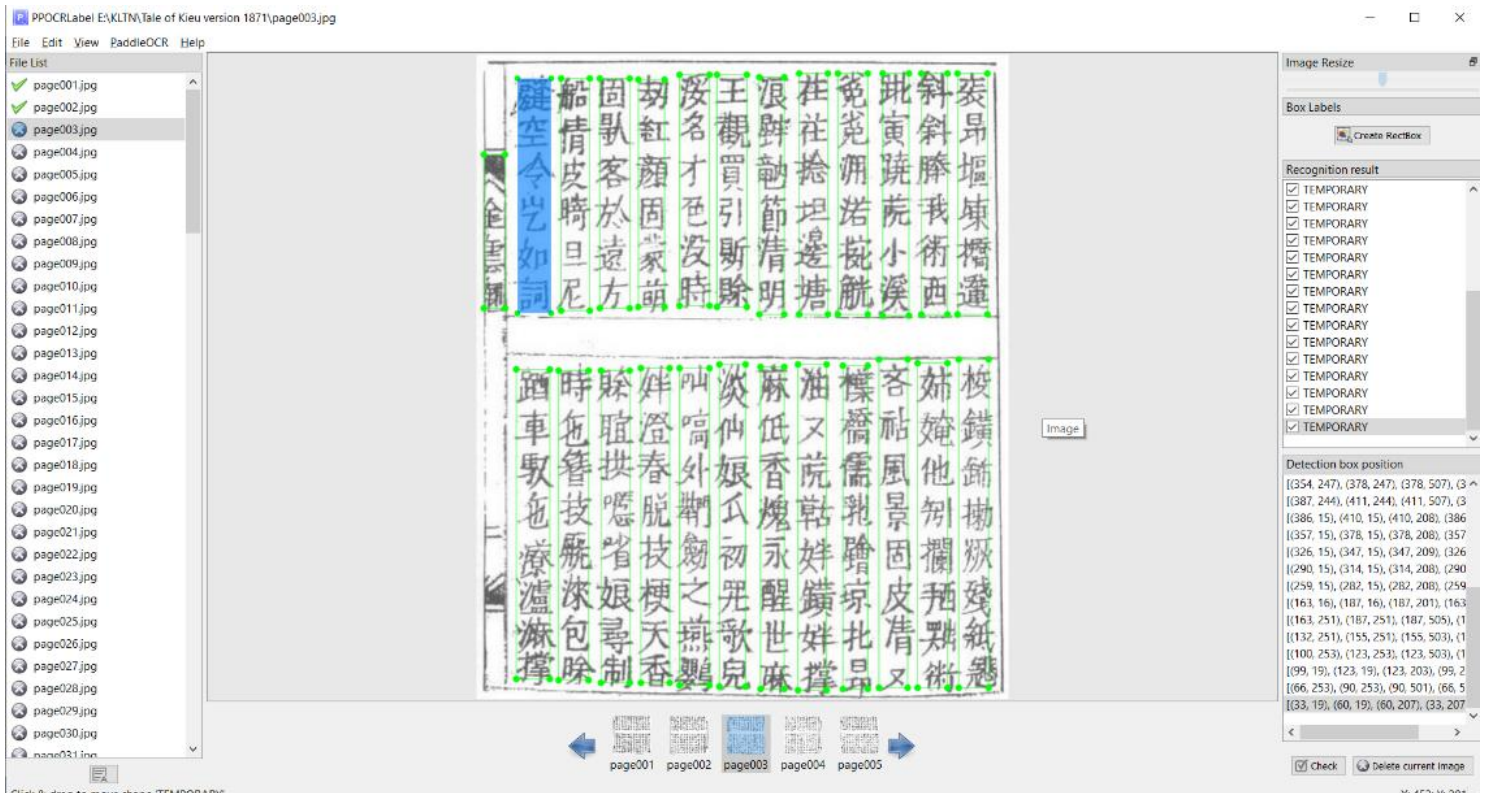


- **Bước 2:** Chọn đến thư mục chứa dữ liệu, nhấn “Select Folder”

- 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:

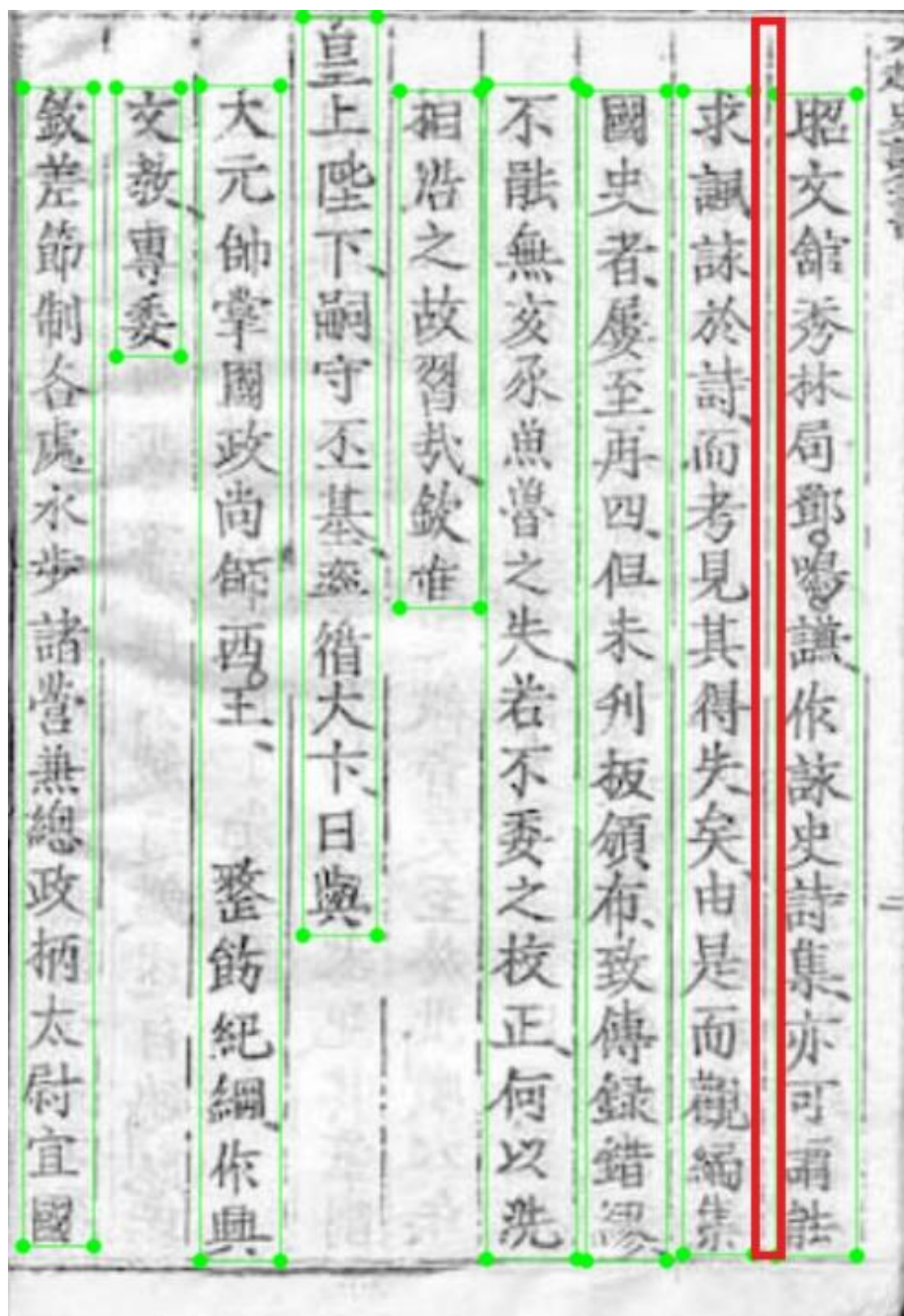
- Bounding Box: các hình chữ nhật màu xanh trong hình trên được gọi là Bounding Box.

Chức năng	Phím tắt	Cách sử dụng
Kéo Bounding Box	W	Nhấn phím W. Sau khi thực hiện chức năng này không hủy được, phải kéo đại Bounding Box rồi dùng chức năng xóa để xóa Bounding Box đó.
Xóa Bounding Box	Backspace	Chọn Bounding Box (nhấn vào giữa Bounding Box) cần xóa, sau đó nhấn Backspace.
Chỉnh sửa Bounding Box		Kéo các ô tròn ở góc các Bounding 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

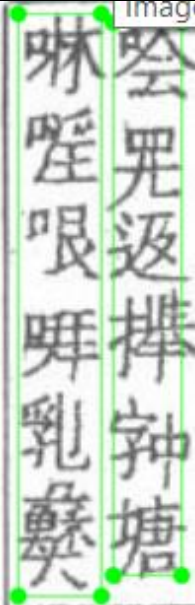
- Lưu ý:

- Các Bounding 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.
- Nên tránh các đường phân cách giữa các cột nếu có thể.
- Ưu tiên kéo Bounding box trọn vẹn câu trong hình:
 - Nếu kéo trọn vẹn các box bị đè lên nhau thì chấp nhận mất một số nét kéo dài của chữ.
 - Nếu kéo trọn vẹn mà bị đè lên đường phân cách giữa các cột vẫn chấp nhận

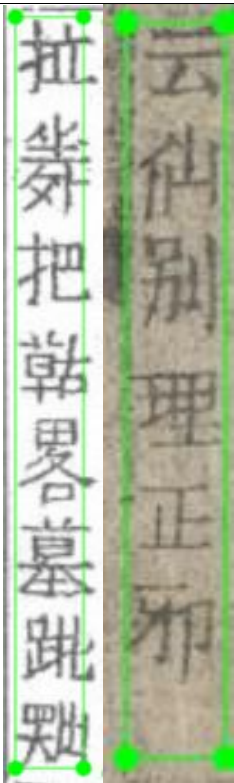



- **Bước 1:** Kiểm tra từng Bounding Box.


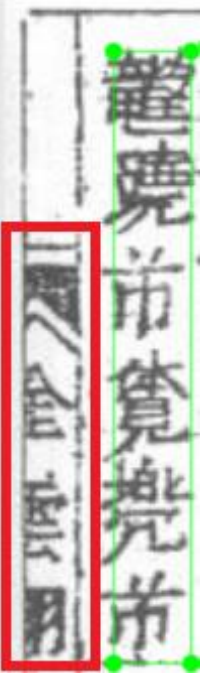
- Trường hợp 1: thiếu Bounding Box.

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

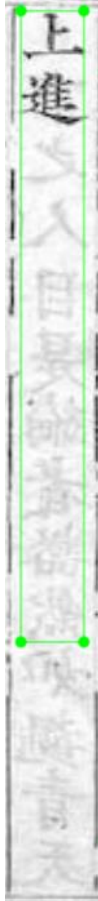

- Trường hợp 2: Bounding Box chưa đủ phủ hết chữ, hoặc phủ dư quá nhiều. Lưu ý: Bounding Box phải phủ hết câu không được thiếu nét của chữ, trong trường hợp Bounding Box bị đè lên nhau chấp nhận kéo thiếu các nét kéo dài của chữ.

Vấn đề	Xử lý	Kết quả
	Ta có thể thấy bên phải không đủ phủ hết chữ. Ta chỉ cần nắm kéo ô tròn ở dưới/trên và bên phải, sang bên phải.	

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

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



- Trường hợp 4: không gán các chữ mờ.

Vấn đề	Xử lý	Kết quả
	Ta kéo cho vừa đủ 2 chữ. Khi đếm số từ để gán nhãn cũng chỉ đếm 2 từ.	



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

Vấn đề	Xử lý	Kết quả
	<p>Trong Box có vết mực, ta chỉ cần kéo cho vừa đủ chữ không cần kéo hết vết mực.</p> <p>Chữ bị vết mực vẫn đếm trong trường hợp này (6) nhưng khi gán nhãn bỏ dấu ✓</p>	



- Trường hợp 6: cột bị dính hoặc chứa dấu mộc.

Vấn đề	Xử lý	Kết quả
	<p>Chỉ kéo box kết thúc các chữ, không kéo dấu mộc và cũng không bỏ các chữ bị dính dấu mộc.</p>	

- Trường hợp 7: có 2 câu nhỏ (2 từ) trong một cột.

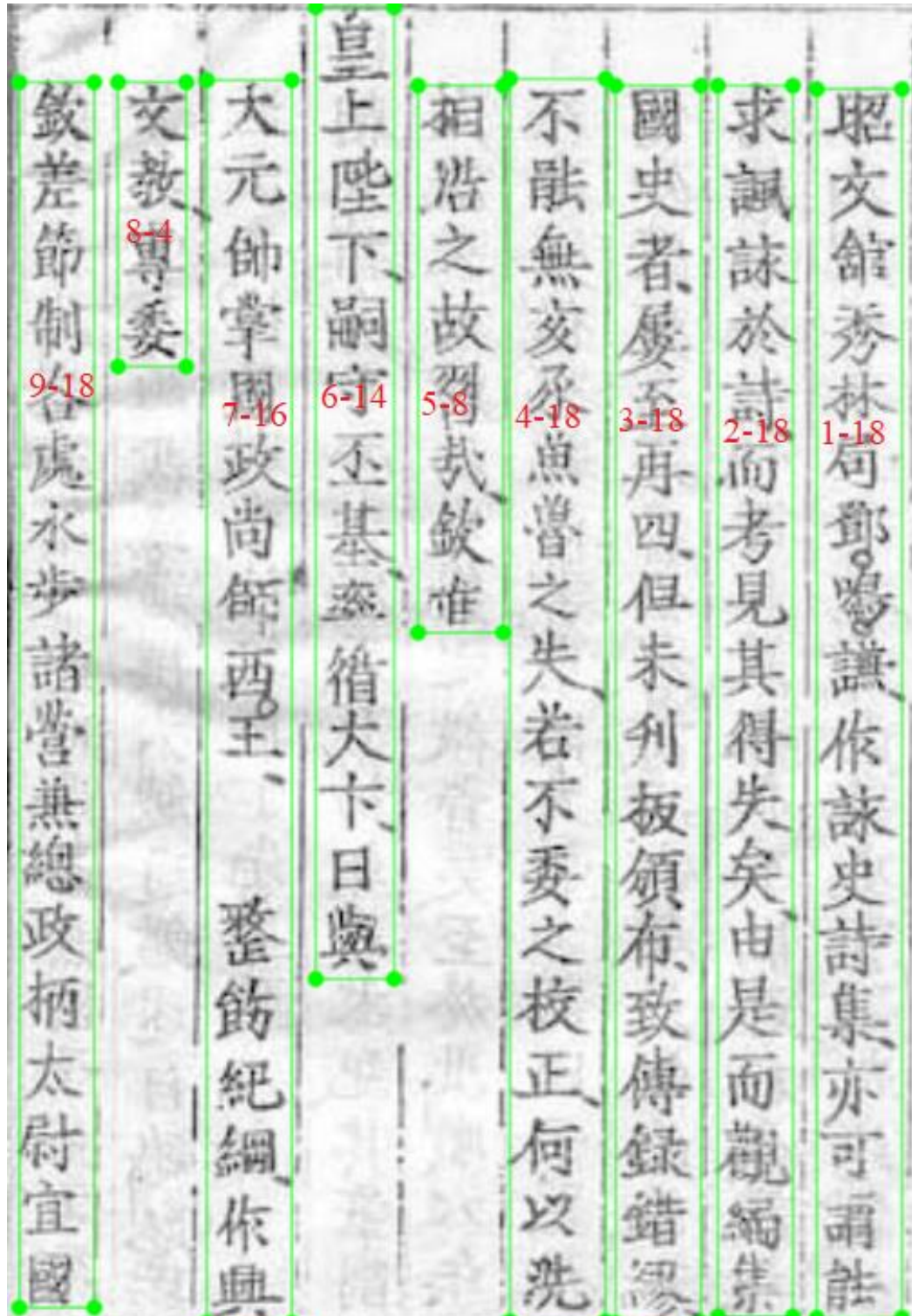
Vấn đề	Xử lý	Kết quả
	<p>Khi 2 câu (2 từ) trong một cột ta gán từ trên xuống dưới và phải qua trái.</p> <p>Thứ tự câu từ trên xuống và phải qua trái như hình.</p>	

- Trường hợp 8: mất chữ ở giữa câu.

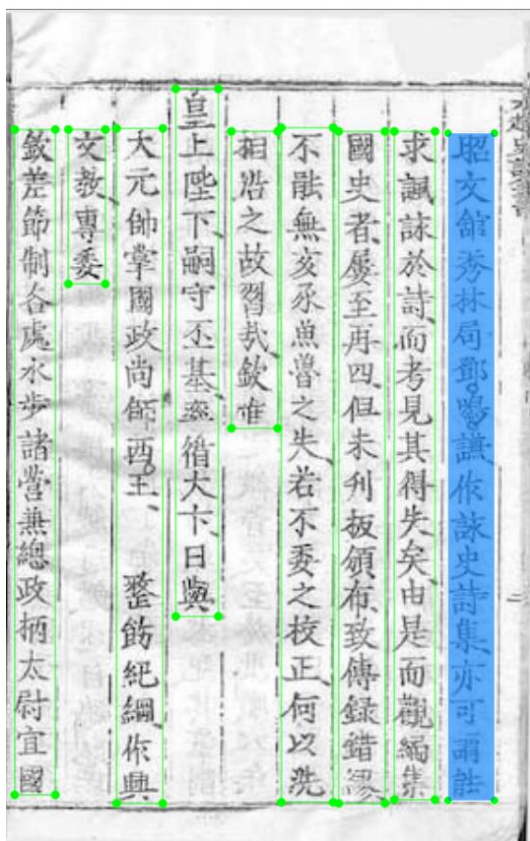
Vấn đề	Xử lý	Kết quả
	<p>Khi mất chữ ở giữa câu cũng kéo trong một box, chỉ đếm các từ không bị mất, trong trường hợp này là 16 từ.</p>	

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

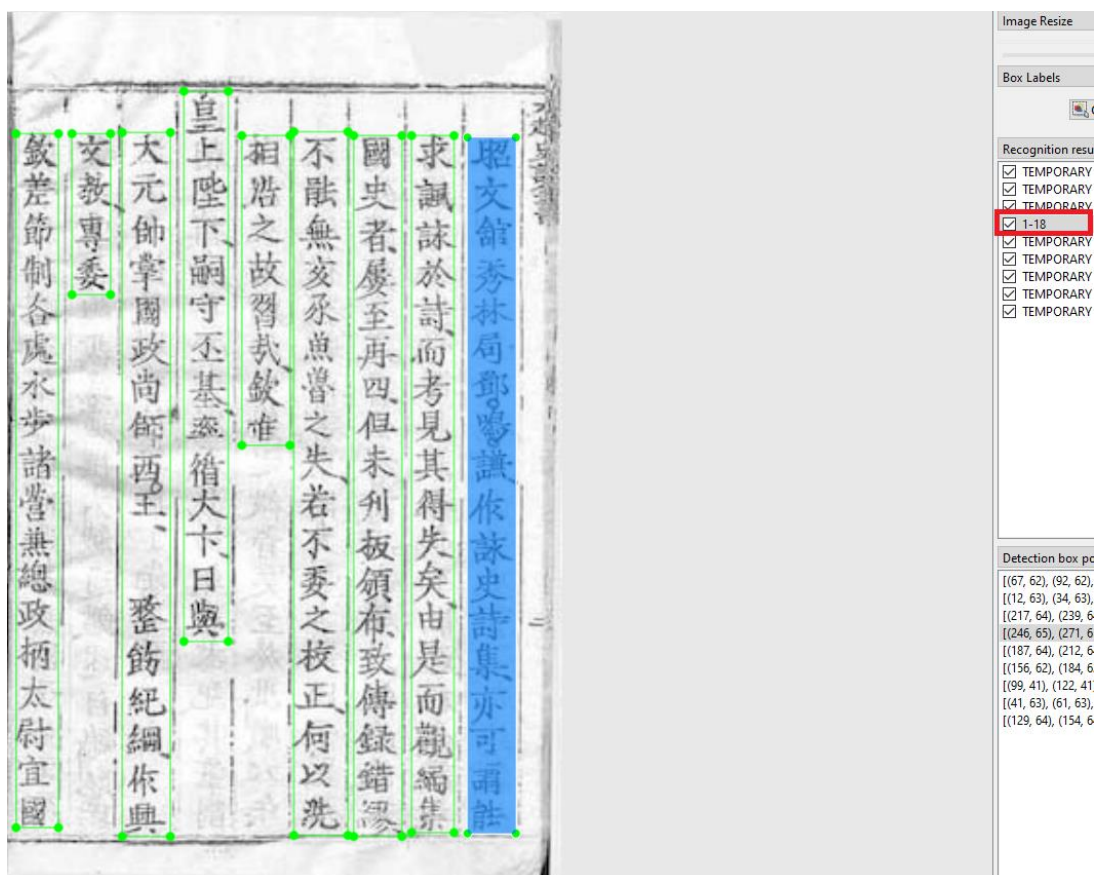
- Thứ tự gán nhãn: Gán nhãn các Bounding Box từ trên xuống dưới và từ phải sang trái như hình.
 - Nhãn gồm 2 phần:
 - Phần 1: Số thứ tự của Bounding Box được tính từ trên xuống dưới và từ phải sang.
 - Phần 2: Số lượng chữ có trong Bounding Box.
- ⇒ VD: 1-18 (nghĩa là Bounding box đầu tiên và có 18 chữ).



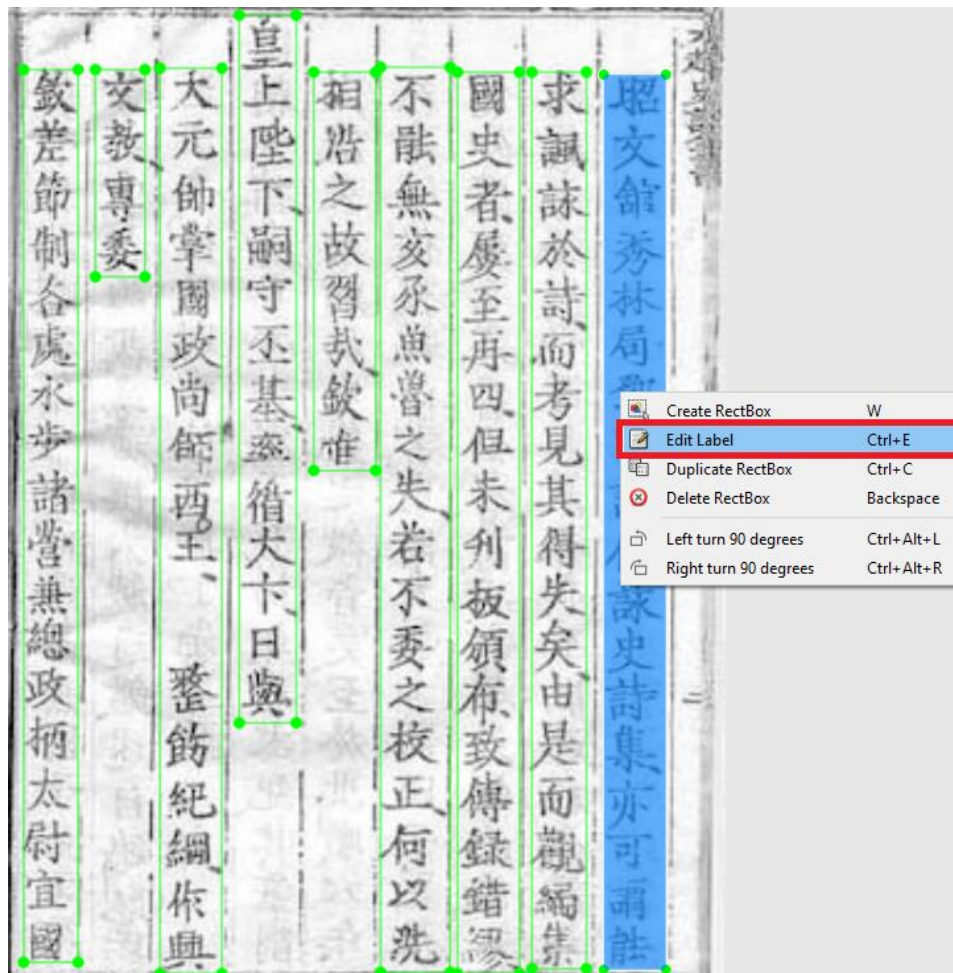
- Đầu tiên click chọn vào Bounding Box cần gán nhãn.



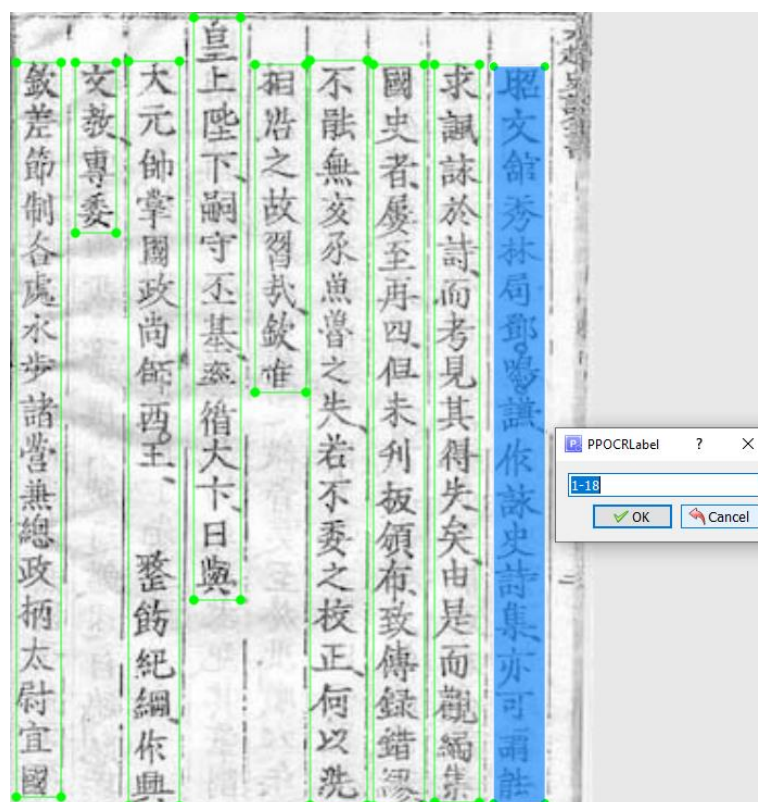
- Cách 1: Bên khung bên phải (được tô màu đỏ), click double chuột vào đánh số thứ tự của box và số chữ trong box, trong trường hợp này nhãn là 1-18.



- Cách 2: Có thể click chuột phải chọn “Edit Label”.

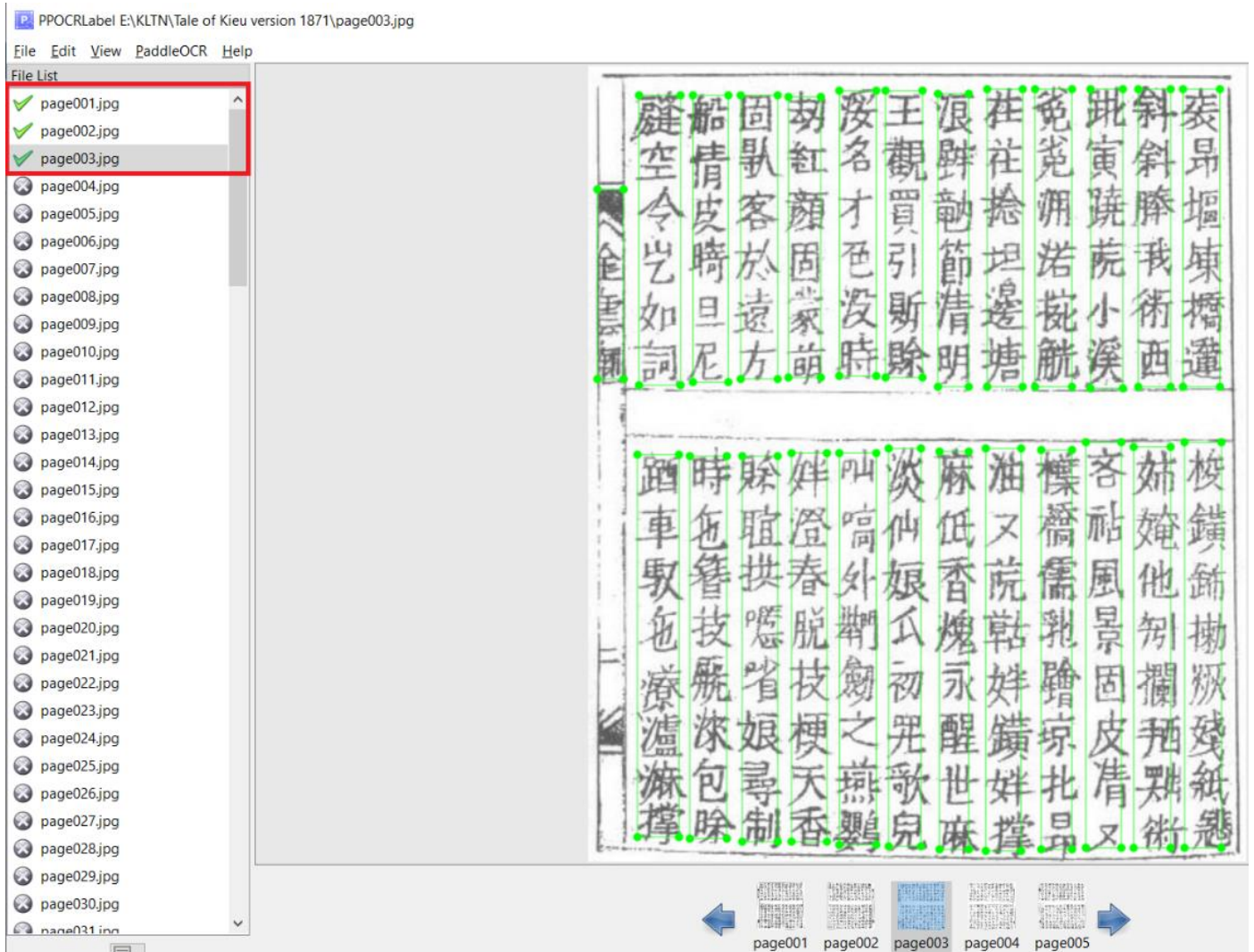


- Sau đó nhận nhãn là số thứ tự của box và số chữ trong box, rồi click “Ok” hoặc nhấn enter.

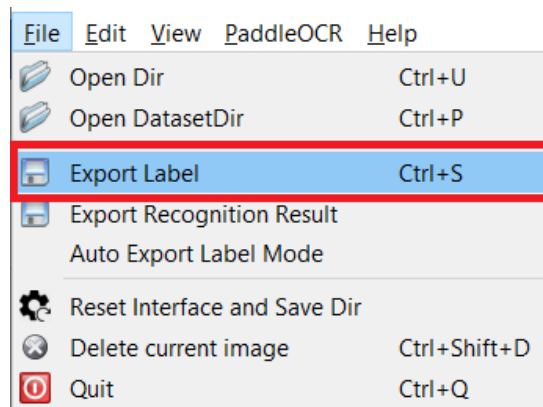


- **Bước 3:** Sau khi gán nhãn hết các Bounding Box ở trên ta thực hiện chức năng “check”.

- 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 Bounding 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 Đăng](#)

Gmail: 18520339@gm.uit.edu.vn

 Facebook: [Nguyễn Đức Duy Anh](#)

Gmail: 18520455@gm.uit.edu.vn

4. Tham khảo

- Mã nguồn gốc PPOCRLLabel:

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

- 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>
- ĐVSKTT: <https://www.nomfoundation.org/nom-project/History-of-Greater-Vietnam>