Cách gán nhãn sắc phong bằng PP-OCR

Mục lục

[I. Dữ liệu hiện có 1](#_Toc149320295)

[1. Dữ liệu ảnh 1](#_Toc149320296)

[2. Dữ liệu văn bản 1](#_Toc149320297)

[3. Dữ liệu văn bản đã ghép các ký tự lại và xóa dấu câu 2](#_Toc149320298)

[II. Cách gán nhãn 3](#_Toc149320299)

[1. Cài đặt PPOCR 3](#_Toc149320300)

[2. Chọn thư mục cần gán nhãn 3](#_Toc149320301)

[3. Gán nhãn tự động cho phát hiện và nhận diện văn bản. 4](#_Toc149320302)

[4. Tải mô hình của chính mình lên như thế nào? 5](#_Toc149320303)

[5. Gán nhãn cho phát hiện văn bản 5](#_Toc149320304)

[6. Gán nhãn cho nhận dạng văn bản 6](#_Toc149320305)

[III. Tham khảo 8](#_Toc149320306)

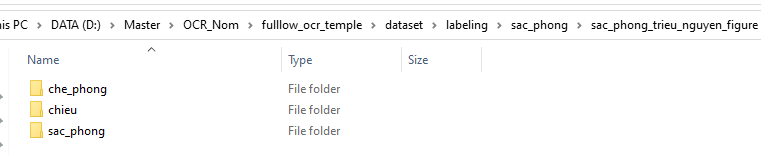
# Dữ liệu hiện có

## Dữ liệu ảnh

Bao gồm 3 thư mục con là:

* Chiếu: 19 ảnh.
* Chế phong: 100 ảnh.
* Sắc phong: 180 ảnh.

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



Hình :Minh họa dữ liệu hình ảnh

## Dữ liệu văn bản

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

Link drive: <https://drive.google.com/file/d/1_uEXPjgzD9ZW8G4uvbhhudaCwNeJI6Za/view?usp=sharing>

A screenshot of a computer

Description automatically generated

Hình : Minh họa dữ liệu văn bản

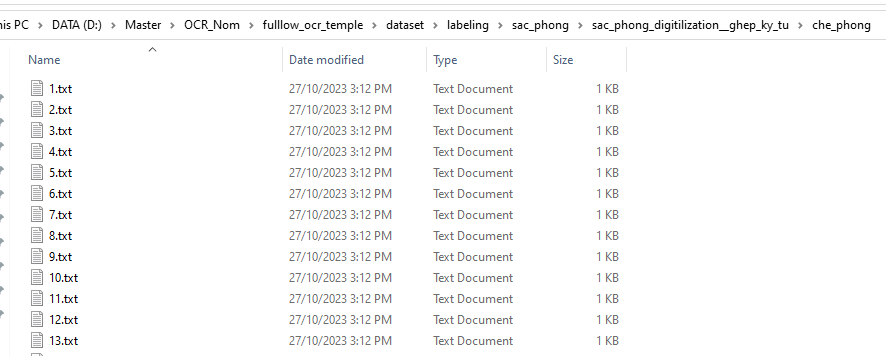
## Dữ liệu văn bản đã ghép các ký tự lại và xóa dấu câu

Dữ liệu chứa các văn bản được đánh số theo thứ tự của từng chế phong, sắc phong và chiếu.

A screenshot of a computer

Description automatically generated

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



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

A close up of a message

Description automatically generated

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

# Cách gán nhãn

## Cài đặt PPOCR

* Để cài đặt thư viện PPOCRLabel ta dùng Pip:

pip install PPOCRLabel

* Để kích hoạt PPOCTLabel, ta sử dụng lênh:

PPOCRLabel –lang ch

## Chọn thư mục cần gán nhãn

A screenshot of a computer

Description automatically generated

Hình : 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 .

A screenshot of a computer screen

Description automatically generated

## 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.

A screenshot of a computer

Description automatically generated

Hình : Nhận diện tự động

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

A screenshot of a computer

Description automatically generated

Hình : Chọn mô hình nhận diện tự động

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

A screenshot of a computer

Description automatically generated

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

## 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.

A computer screen shot of a program

Description automatically generated

Hình : 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)

## Gán nhãn cho phát hiện văn bản

Chọn nút “Gán nhãn đa giác”.

A screenshot of a computer screen

Description automatically generated

Hình : Chọn gán nhãn đa giác

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

A screenshot of a computer

Description automatically generated

Hình : 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.

## Gán nhãn cho nhận dạng văn bản

Hình ảnh sắc phong đã gán nhãn phần phát hiện văn bản.

A screenshot of a computer

Description automatically generated

Hình : Sắc phong đã gán nhãn phần phát hiên văn bản

A screenshot of a computer

Description automatically generated

Hình : 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.

A screenshot of a computer

Description automatically generated

**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 trị 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.

**A screenshot of a computer

Description automatically generated**

# Tham khảo

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