**BÁO CÁO NỘI DUNG THỰC TẬP**

Nhóm trưởng: Mr Long

Người báo cáo: Nguyễn Hà Linh

MỤC LỤC

[1. **Xây dựng kế hoạch thực hiện (đơn vị theo task)**. 1](#_Toc143557695)

[**2.** **Link GitHub quản lý source code.** 1](#_Toc143557696)

[**3.** **Report công việc cho task** 1](#_Toc143557697)

[3.1 Chuẩn bị 1](#_Toc143557698)

[3.2 Làm quen 3](#_Toc143557699)

[3.3 Xây dựng mã 5](#_Toc143557700)

[3.4 Đánh giá dataset test 7](#_Toc143557701)

1. **Xây dựng kế hoạch thực hiện (đơn vị theo task)**.

|  |  |  |
| --- | --- | --- |
| **Công việc** | **Mô tả** | **Thời gian** |
| Chuẩn bị | Tìm hiểu vấn đề, chuẩn bị môi trường, tải dataset đánh giá, cài đặt PaddleOCR, cài đặt PaddleClas | 15/08 |
| Làm quen | Chạy thử các lệnh và mã với tập dataset mẫu | 16/08 |
| Xây dựng mã | Sử dụng thư viện paddle để xây dựng tool đọc text từ ảnh | 17/08 |
| Đánh giá | Dùng dataset đánh giá kết quả tool | 17/08 |
| Đẩy lên Github | Đẩy tool lên Github | 18/08 |
| Báo cáo | Viết báo cáo, nộp báo cáo | 18/08 |
| Đánh giá dataset | Dùng độ do Jaccard Similarity đánh giá từng ảnh | 21/08 |

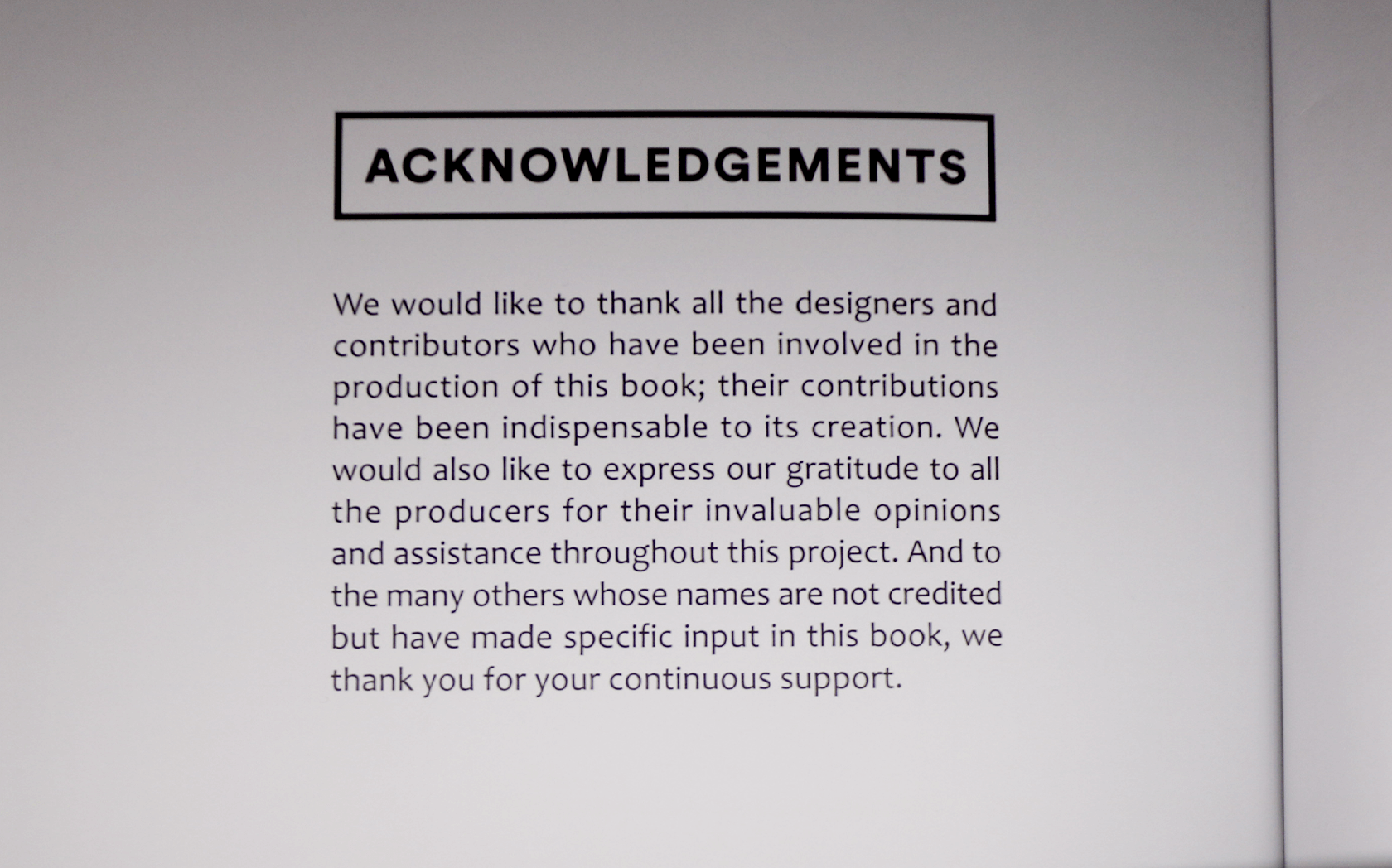
1. **Link GitHub quản lý source code.**

Link GitHub: <https://github.com/linheri06/the_text_recognition1>

1. **Report công việc cho task**
   1. Chuẩn bị
2. Tìm hiểu vấn đề

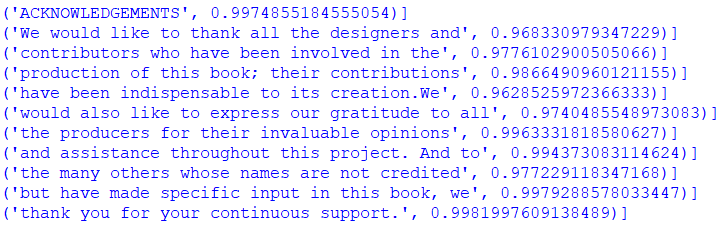
- Xác định yêu cầu: Sử dụng thư viện paddle để xây dựng tool đọc text từ ảnh

+ Input: Ảnh đầu vào chứa text



+Output: Text được đọc:

* Mã đầu ra:



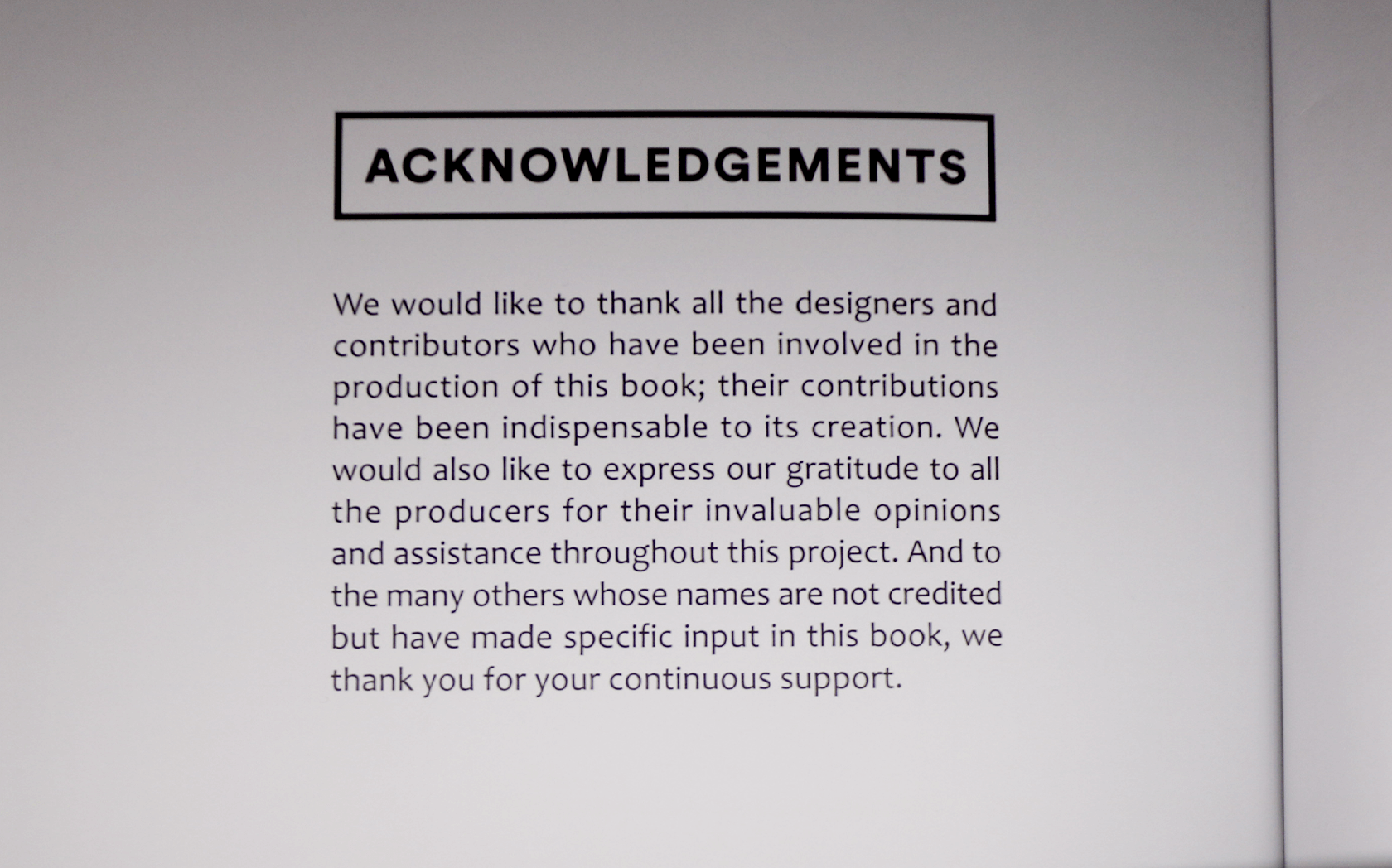
* Mô tả:

A close up of a book

Description automatically generated

- Các vấn đề: Sau khi quan sát dataset đánh giá kết quả thấy đầu vào có 2 trường hợp:

+ Trường hợp ảnh đầu vào có bố cục chữ rõ ràng:



* Sử dụng PP-OCR

+ Trường hợp ảnh đầu vào có bố cục chữ không rõ ràng hoặc dạng bảng:

A white clock on a wall

Description automatically generated A table with numbers and letters

Description automatically generated

* Sử dụng PP-Structure

- Tìm hiểu về thư viện Paddle: Chức năng, cách sử dụng, mở rộng,…

1. Chuẩn bị

- Tải dataset đánh giá

- Chuẩn bị môi trường:

* PaddlePaddle 2.5.1
* Python 3.8

- Cài đặt PaddleOCR 2.7.0.2

- Cài đặt PaddleClas 2.5.1

* 1. Làm quen

1. PaddleOCR:

- Dòng lệnh:

+ Lệnh:

paddleocr --image\_dir ./imgs\_en/img\_12.jpg --use\_angle\_cls true --lang en --use\_gpu false

+ Kết quả:

A screenshot of a computer screen

Description automatically generated

- Mã (python):

A screen shot of a computer

Description automatically generated

+ Kết quả:

A blue text on a white background

Description automatically generated

A close up of a book

Description automatically generated

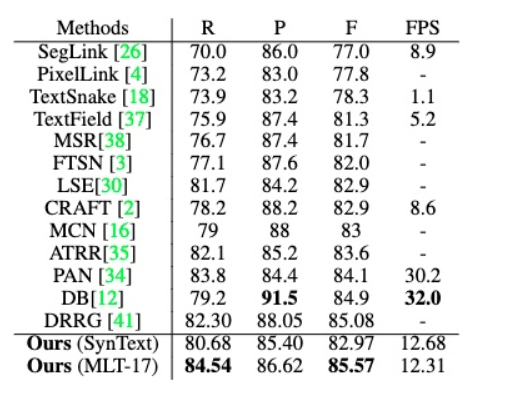
1. PaddleClas:

- Dòng lệnh:

paddleocr --image\_dir=ppstructure/docs/table/1.png --type=structure --image\_orientation=true

- Mã (python):

+Đầu vào



+Mã:

A screen shot of a computer code

Description automatically generated

+Kết quả:

A screenshot of a table

Description automatically generated

* 1. Xây dựng mã
* Căn cứ vào các dạng đầu vào để suwe dụng thư viện:

+ PaddleOCR: Đầu vào không cần phải phân tích bố cục chữ hay nhận dạng bảng

+ PaddleClas: Đầu vào cần phân tích bố cục chũ hoặc nhận dạng bảng

* Xây dựng 2 đoạn mã ứng với 2 thư viện:

+ PaddleOCR:

A screenshot of a computer program

Description automatically generated

+ PaddleClas:

A screenshot of a computer program

Description automatically generated

* 1. Đánh giá dataset test

1. Giới thiệu độ đo Jaccard Similarity.

Jaccard Similarity: Độ đo này tính tỷ lệ số lượng token chung giữa hai chuỗi văn bản so với tổng số token trong hai chuỗi. Nó đơn giản và thường được sử dụng trong các tình huống đếm số lượng phần tử chung trong hai tập hợp.

1. Quy trình

- Nhận xét:

+ Trong thư mục “scene” (chứa datadet test) có 1 file “word.xml” chứa thông tin về các từ trong các ảnh.

=> Đưa về 1 từ điển với key: tên tệp ảnh; value: mảng string chứa các từ trong ảnh. (Đây ứng với quá trình lấy nhãn của data)

+ Duyệt qua từng ảnh trong thư mục: đưa output thành 1 mảng string chứa các từ trong ảnh => so sánh với value từ điển (ứng với key) : dùng độ do Jaccard Similarity.

- Chi tiết:

+ Gán nhãn cho tệp ảnh:

A white background with many text

Description automatically generated with medium confidence

*Thông tin ứng với 1 ảnh trong “word.xml”*

*A close up of text

Description automatically generated*

*Thông tin sau khi lấy ra*

A screenshot of a computer program

Description automatically generated

*Code gán nhãn: tạo 1 từ điển với key: tên tập; value: mảng string chứa các từ*

+Hàm tính độ đo:

A close-up of a computer code

Description automatically generated

+ Duyệt qua từng ảnh trong thư mục dataset test:

A computer screen shot of a computer code

Description automatically generated

*Code ứng với thư mục “ryoungt\_05.08.2002”*

Tương tự với các thư mục “ryoungt\_13.08.2002”, “sml\_01.08.2002”.

A computer code with black text

Description automatically generated

*Kết quả của 1 lần duyệt*

+ Độ đo đánh giá: trung bình cộng của độ do các tệp ảnh:

