OCR Văn bản Tây Du Ký Bản Trung – Việt

## Hướng dẫn cài đặt

### Tesseract OCR

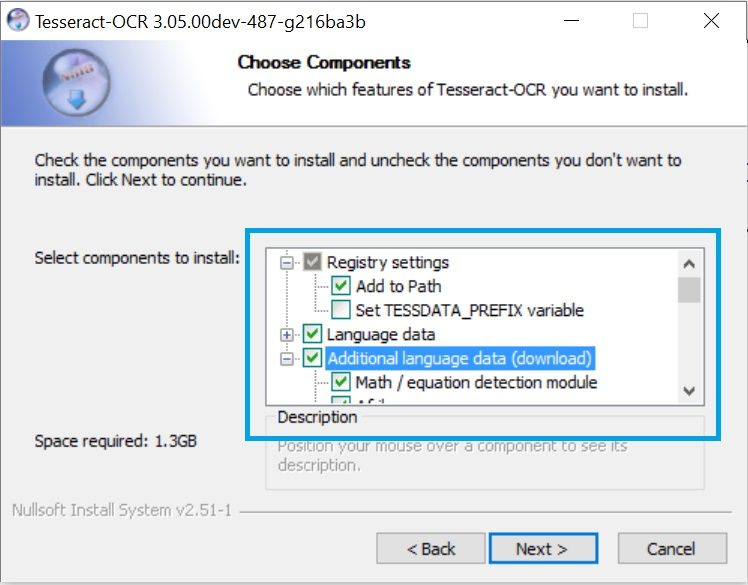
Cài đặt dựa vào hướng dẫn sau:

<https://tesseract-ocr.github.io/tessdoc/Installation.html>

Hoặc

<https://miai.vn/2019/08/22/ocr-nhan-dang-van-ban-tieng-viet-voi-tesseract-ocr/>

Đối với window, trong khi cài đặt cần chọn gói ngôn ngữ tiến trung (chi\_tra) và tiếng việt (vie)



Hoặc download gói ngôn ngữ bổ sung theo hướng dẫn ở link 2.

### Cài đặt các thư viện cần thiết:

Cài đặt các thư viện trong file requirements.txt đi kèm bằng cách:

pip install -r requirements.txt

## Hướng dẫn chạy code

Để OCR file pdf ra text, chạy file *ocr\_utils*.py

Với các thông số sau:

-l, --lang: ngôn ngữ của văn bản. Chọn 1 trong “viet”, “trung”

-p, --pdf\_path: đường dẫn đến file văn bản

-fp, --first\_page: trang bắt đầu

-np, --np: số trang muốn OCR

Để OCR bản tiếng Trung:

python main.py -l trung -p Book-Tay\_Du\_Ky-Trung.pdf -fp 1 -np 20

Để OCR bảng tiếng Việt:

python main.py -l viet -p Book-Tay\_Du\_Ky-Viet.pdf -fp 48 -np 20

Số trang bắt đầu câu chuyện của tiếng Việt là trang 48

## Tổng quan về văn bản:

## Bản tiếng Việt

Tác giả: Ngô Thừa Ân

Dịch giả: Như Sơn, Mai Xuân Hải và Phương  
Oanh

Nhà xuất bản Văn học hà Nội – 1988

Dịch từ nguyên bản tiếng Trung Quốc – Nhân dân văn học xuất bản xã – 1972

Khổ sách: 13cm x 19cm

Số tập: 10

Ngày bắt đầu: 19-01-2013

Ngày hoàn thành: 21-6-2013

Scan: 4DHN

Scan những trang bị rách: ICT

Hình minh họa: Thien\_Tue

Nguồn: http://www.e-thuvien.com

Số trang: 3435

### Bản tiếng Trung:

Title: 西遊記

Author: Cheng'en Wu

Release date: December 22, 2007

[eBook #23962]

Language: Chinese

Credits: Produced by Leong Joana Kit Ieng

\*\*\* START OF THE PROJECT GUTENBERG EBOOK 西遊記 \*\*\*

Produced by Leong Joana Kit Ieng

## Khó khăn trong quá trình OCR

### Bản tiếng Việt:

Ảnh gốc từ file PDF sẽ được chuyển về trắng đen nhằm tăng độ chính xác cho việc OCR.

Ảnh PDF gốc:

A page of a text

AI-generated content may be incorrect.

Ảnh PDF sau khi chuyển về trắng đen:

A black text on a white background

AI-generated content may be incorrect.

### Văn bản bị xuống dòng

Sau khi OCR văn bản thu được có nhiều đoạn tự động xuống hàng mặc dù còn đang nằm giữa câu. Nguyên nhân có thể là do định dạng của văn bản trong file pdf

Văn bản trước khi xử lý:

A screenshot of a computer screen

AI-generated content may be incorrect.

Văn bản sau khi xử lý:

A black screen with white text

AI-generated content may be incorrect.Phương pháp xử lý:

Phân tíchh từng dòng trong văn bản xem thuộc về thơ hay đoạn văn. Nếu là đoạn văn thì sẽ nối các câu lại với nhau. Nếu là thơ thì sẽ giữ nguyên mỗi dòng một câu. Để xác định dòng là thơ hay không sẽ dựa vào độ dài của chữ trong câu (số chữ ít hơn 9) và nếu bắt đầu dòng là chữ viết hoa thì sẽ xem là thơ. Ngược lại xem là câu trong đoạn văn.

Có thể sử dụng phương pháp sử lý khác bằng các tắt phương pháp xử lý hiện tại trong file main.py -pt False

### Văn bản chứa các ký tự lỗi

Do trong văn bản có chứa các ký tự nhằm mục đích chú thích ([1]), khiến cho việc OCR đưa ra kết quả không chính xác ở các dòng có chứa ký tự trên.

Ví dụ:

Văn bản gốc:

A text on a page

AI-generated content may be incorrect.

Văn bản bị lỗi:

*lò 6m...*

*Một hôm tảng đá nứt đôi, sinh ra một quả trứng đá tròn, to bằng quả cầu, gặp gió hóa thành một con khỉ đá, đủ cả mặt mũi chân tay. Con khỉ đá liền học bò học chạy, vái lạy bốn phương trời làm kinh động cả ai.*

*thượng ở.*

*Tiêu! cùng các vị tiên. Nhìn thấy tia hào quang chói lọi. Thượng đế lập tức ra lệnh Thiên lý nhãn, Thuận phong nhĩ, mở cửa Nam thiên xem xét. Hai.*

*lúc ấy đang ngự trị ở điện Linh.*

*tướng vâng chỉ ra xem, nhìn nghe đích thực, một lát quay về báo:*

Phương pháp xử lý: Chưa có, do mỗi ký tự sẽ đưa ra một kết quả khác nhau khi OCR.

### Sai chính tả

Có thể còn chứa sai sót về chính tả trong quá trình OCR.

Phương pháp: Sửa chính tả bằng mô hình “bmd1905/vietnamese-correction-v2” lấy từ Hugging Face.

### Bảng tiếng Trung:

Văn pdf bản gốc:

A black text on a white background

AI-generated content may be incorrect.

Văn bản pdf tiếng trung không cần xử lý đưa về đen trắng.

### Thông tin ở đầu trang và cuối trang bị lẫn vào nội dung

Văn bản sau khi OCR vẫn còn thông tin nằm ở đầu và cuối trang, có dạng chữ latin và không liên quan đến nội dung tác phẩm. Cần xóa bỏ các thông tin đó để văn bản thu được “sạch“hơn.  
Văn bản trước khi xử lý:

A screenshot of a computer screen

AI-generated content may be incorrect.

Văn bản sau khi sử lý:

A black background with many small white symbols

AI-generated content may be incorrect.

Phương pháp xử lý: Thông tin nằm trong header và footer được ngăn cách với nội dung bằng 2 lần xuống hàng (‘\n\n’). Dựa vào đó có thể bỏ đi thông tin nằm trong 2 phần trên.

### Vấn đề về dấu

Dấu 」đôi khi bị nhầm sang dấu ] khiến cho các dấu trong văn bản trở nên không đồng nhất.

Phương pháp: hiện chưa có phương pháp

### Sai chính tả

Có thể còn chứa sai sót về chính tả trong quá trình OCR

Phương pháp: hiện chưa có cách để kiểm tra