คู่มือการใช้งาน Tesseract 5

1. ติดตั้งระบบปฏิบัติการ unix หรือใช้ virtual machine ที่ลงระบบปฏิบัติการ unix หรือติดตั้ง wsl บนเครื่องที่มีระบบปฏิบัติการ windows
2. ติดตั้ง tesseract 5 บนระบบปฏิบัติการ unix ตามคู่มือการติดตั้ง https://tesseract-ocr.github.io/tessdoc/Compiling-%E2%80%93-GitInstallation.html

การติดตั้งเครื่องมือในการเทรนฟอนท์ใหม่ให้ model

1. โคลนโฟลเดอร์ที่เก็บเครื่องมือสำหรับเทรนโมเดลจาก https://github.com/tesseract-ocr/tesstrain.git
2. สร้างโฟลเดอร์สำหรับเก็บ pretrained ตัวอักษรของภาษาชื่อ langdata โดยโคลนมาจากเว็บ https://github.com/tesseract-ocr/langdata.git (เลือกภาษาที่ต้องการอย่างเดียวได้) หากเป็นภาษาไทยให้ดาวน์โหลดไฟล์ radical-stroke.txt ด้วย
3. ดาวน์โหลดไฟล์ pretrained ของภาษา (.traineddata) จากเว็บ https://github.com/tesseract-ocr/tessdata โดยนำไปเก็บไว้ในโฟล์เดอร์ที่มีชื่อว่า /usr/local/share/tessdata
4. ติดตั้งฟอนท์ที่ต้องการลงในระบบปฏิบัติการ
5. สร้างโฟลเดอร์สำหรับเก็บ dataset ของภาษาที่ต้องการเทรนเพิ่มเติม ชื่อ data ในโฟลเดอร์ของ git ที่โคลนมา
6. สร้างโฟลเดอร์สำหรับเก็บ dataset ของฟอนท์ที่ต้องการเทรนด้วยชื่อ font-name-ground-truth (แทน font-name ด้วยชื่อของฟอนท์ที่ต้องการ)
7. สร้าง dataset ในการเทรนโมเดล ประกอบด้วย .gt.txt ซึ่งเก็บข้อความที่แท้จริง .box เก็บพิกัดของอักขระภายในภาพ และ .tif เป็นไฟล์ภาพที่มีข้อความจาก .gt.txt ในรูปแบบของฟอนท์ที่ต้องการเทรน เก็บทั้งสามไฟล์ไว้ใน ไว้ในโฟลเดอร์ font-name-ground-truth โดยใช้คำสั่ง text2image ซึ่งสามารถกำหนดฟอนท์ที่ต้องการ ข้อความ ตำแหน่งของไฟล์ output ขนาดของภาพ และชุด unichar ได้
8. เทรนโมเดลโดยใช้คำสั่ง make training MODEL_NAME=font-name START_MODEL=base-langaude TESSDATA=../../../../usr/local/share/tessdata MAX_ITERATIONS=1000

9. ในการ combine 2 model (model font และ pretrain model) ให้นำโมเดลของฟอนท์ไปใส่ใน folder ของ tessdata จากนั้นใช้คำสั่ง combine_tessdata -o tessdata/eng.traineddata \ tessdata/font-name.traineddata
10. เพิ่มเติม สามารถศึกษาวิธีการใช้งานฟังก์ชันอื่นๆของ tesseract 5 ได้จาก : https://github.com/tesseract-ocr/tesseract/tree/main/doc

แหล่งที่มา :

Gabriel Garcia. (2022). *Training Tesseract 5 for a New Font.* From: https://github.com/astutejoe/tesseract_tutorial