**NIA 24-38 수식 OCR 데이터**

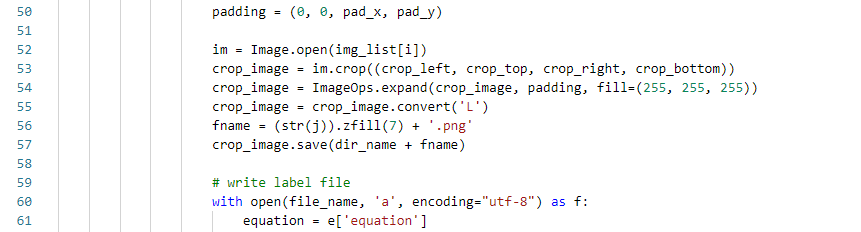
**데이터 전처리**

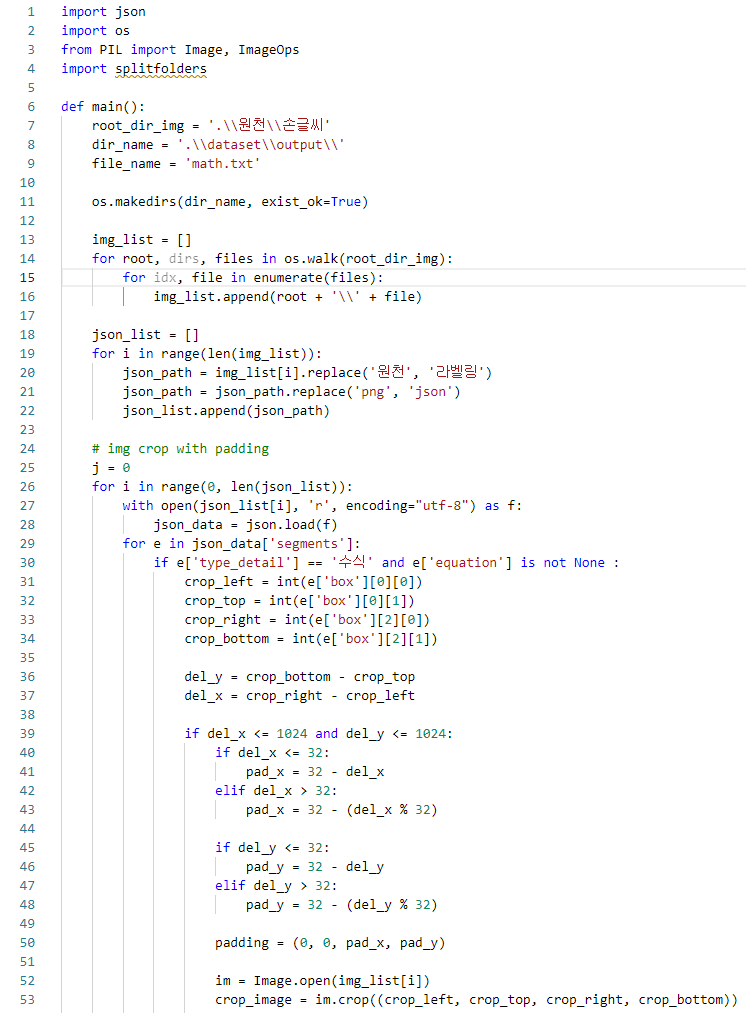


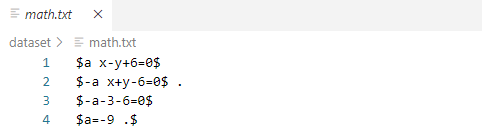
json파일 예시

라벨링 파일(json파일)의 **“type” : “수식/텍스트”** 의 **“type\_detail” : 수식**에해당하는 segments의 box좌표를 사용해 원천 이미지를 crop합니다. crop 과정에서 모델 학습을 위해 이미지의 가로 세로 길이가 32의 배수가 되도록 padding을 넣어 주었으며, 안정적인 학습을 위해 max length는 1,024로 제한했습니다. segments의 “equation” 은 학습을 위해 따로 math.txt 파일에 기록해 두었습니다. math.txt파일은 dataset폴더안에 들어있습니다.

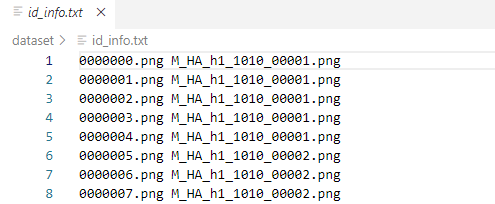
crop된 이미지의 파일명을 index로 저장했으며 (ex.0000001.png) 해당 파일의 equation정보는 math.txt파일의 각index 라인에 해당합니다.

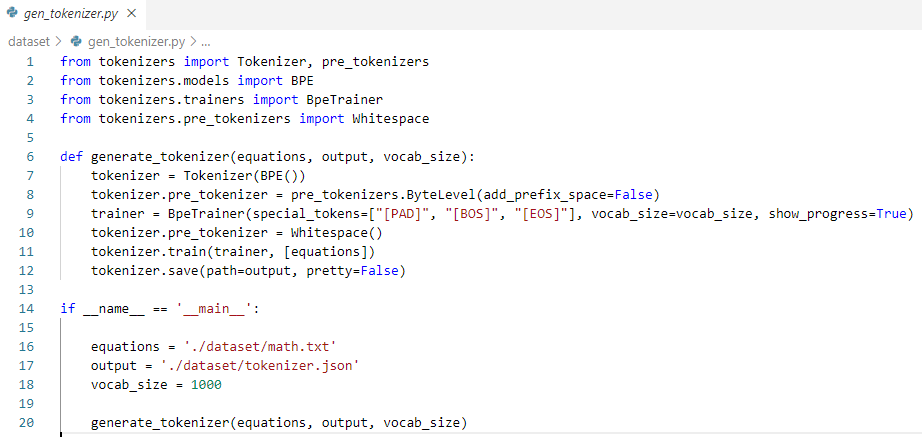




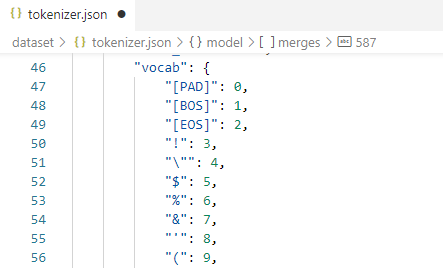


전처리 전후의 파일명은 id\_info.txt 파일에서 확인 가능합니다.

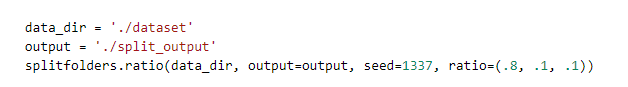


gen\_tokenizer.py

math.txt 파일로부터 학습을 위한 tokenizer 파일을 생성합니다.

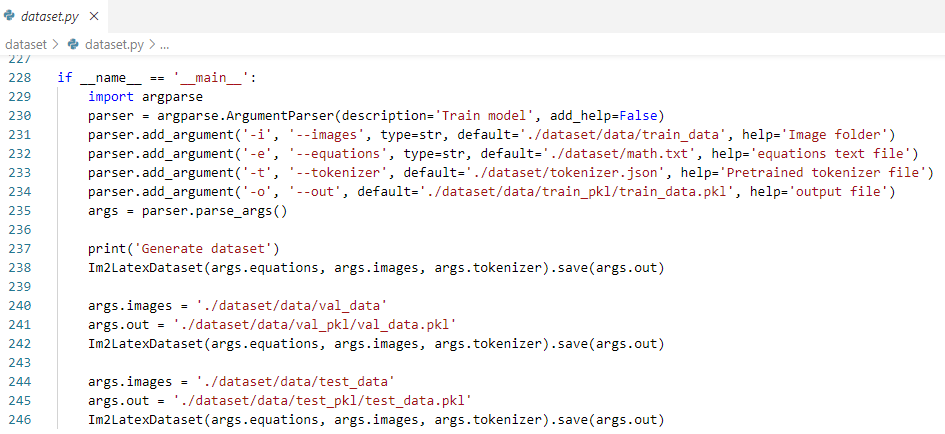


생성된 tokenizer 파일



데이터는 학습용, 검증용, 평가용으로 8:1:1 로 나누어집니다.

dataset.py



이미지 경로, math.txt, tokenizer 파일로부터 학습을 위한 pickle 파일을 생성합니다.

**NIA 유효성 검증**

위의 전처리 과정은 미리 수행하여 Docker 이미지 안에 넣어 두었으므로. 유효성 검증과정에서는 학습 및 평가 과정만 진행하셔도 상관없습니다.

**학습**

train.py



학습이 진행되며 일정 epoch마다 훈련된 모델을 저장합니다. 훈련된 모델은 평가 수행 시에 사용됩니다.

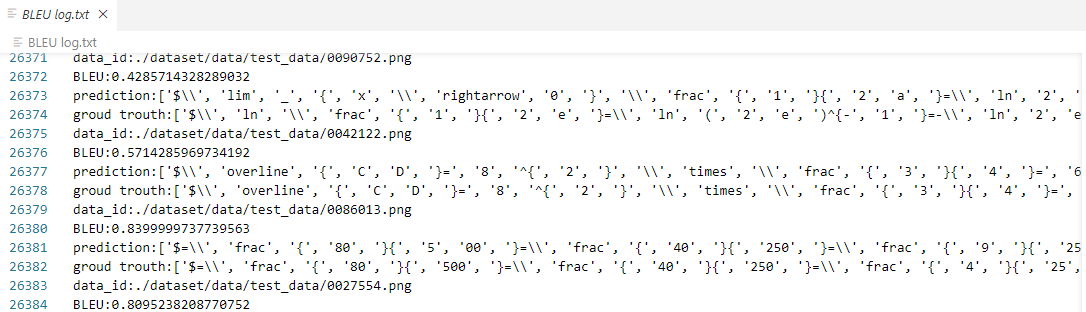
**평가**

eval.py

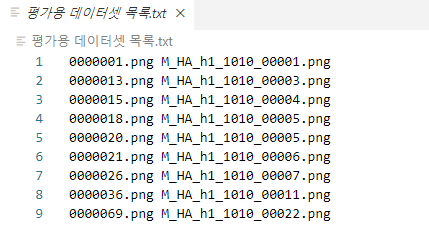


test데이터셋으로부터 평가를 수행하여 bleu score를 계산합니다.

eval.py 실행 시에 BLEU log.txt 파일이 생성되며, prediction은 모델의 예측 값, ground truth는 정답 라벨 값, data\_id는 예측에 사용된 이미지, BLEU는 prediction과 ground truth에 의해 계산된 점수입니다.



평가용 데이터셋 목록.txt 파일에서 평가에 사용된 데이터 파일명(전처리 된 데이터), 소스 이미지(crop한 원천데이터) 파일명을 확인할 수 있습니다.



**Docker 구동 명령어**

1. 도커 이미지 불러오기

docker load -i nia\_24-38\_ocr\_docker.tar

docker load -i nia\_24-38\_ocr\_handwritten\_docker.tar

2. 컨테이너 생성

docker run --gpus all -it --name [CONTAINER NAME] --rm --ipc=host [**IMAGE NAME**] /bin/bash

ex) docker run --gpus all -it --name NIA\_OCR --rm --ipc=host nia-24-38-ocr /bin/bash

3-1. 모델 학습 및 평가 (인쇄체)

디렉토리 이동: cd /NIA\_OCR\_printed

학습: python train.py

학습 결과 로그파일은 ‘유효성 증빙/038. 인쇄체 데이터/ 시험 결과/train\_log.txt’ 입니다.

평가: python eval.py

평과 결과 로그파일은 ‘유효성 증빙/038. 인쇄체 데이터/ 시험 결과/성능평가수행로그.txt’ 입니다.

코드 실행시 BLEU log.txt 파일이 생성됩니다. (시험 결과/evaluation\_log\_epoch30n4.txt 파일)

3-2. 모델 학습 및 평가 (손글씨)

디렉토리 이동: cd /NIA\_OCR\_printed

학습: python train.py

학습 결과 로그파일은 ‘유효성 증빙/038. 손글씨 데이터/ 시험 결과/train\_log.txt’ 입니다.

평가: python eval.py

평과 결과 로그파일은 ‘유효성 증빙/038. 손글씨 데이터/ 시험 결과/성능평가수행로그.txt’ 입니다.

코드 실행시 BLEU log.txt 파일이 생성됩니다. (시험 결과/evaluation\_log\_epoch30n4.txt 파일)