transformers : huggingface.com/transformers/ <- 이 페이지 중요 사용하는 모델과 함수들 정의

mlm 말고 편집 거리 이용(edit\_distance?? 러시아에서 개발한 이런 함수가 있다고함)

문맥의존 철자 교정(범위를 어디까지할지 확립할 필요) :

1. 문법

2. 동음이의어 같은 느낌 발음 비슷한(peace, piece)

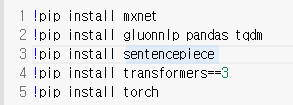
3. 타자 틀렸는데 단어가 있는 경우(from, form)

해봐야할 것 : 문장에서 [masked] 해서 이 단어 유추하는거 해봐야함. , etri kobert, klue/roberta-large (후달리면 base나 medium사용해도됨) 도 사용해봐야됨.

+ coourpus 몇개로 추가 학습시켜보기(기사 자료 오기전에)

참고

Skt-kobert tutorial : <https://colab.research.google.com/github/SKTBrain/KoBERT/blob/master/scripts/NSMC/naver_review_classifications_pytorch_kobert.ipynb>



여기서 transformers==3 해줘야함.. 최신 버전으로 하면 마지막 예측할 때 “Input: must be Tensor, not str” 에러 발생!!!!



: 데이터 파일 불러오기(위 train, 아래 test, 데이터 내용은 아래와 같음)

id document label

9976970 아 더빙.. 진짜 짜증나네요 목소리 0

3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나 1

10265843 너무재밓었다그래서보는것을추천한다 0

7156791 액션이 없는데도 재미 있는 몇안되는 영화 1

(네이버 영화리뷰를 가져왔으며, id값에 따라 리뷰평과 label은 부정(0), 긍정(1) 나타냄. )



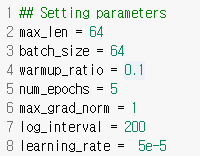
Train data와 test data 구분

field\_indices의 기본값이 [0, 1] ,

(위 예제는 텍스트가 1, 레이블이 2이기 때문에 위처럼 설정)

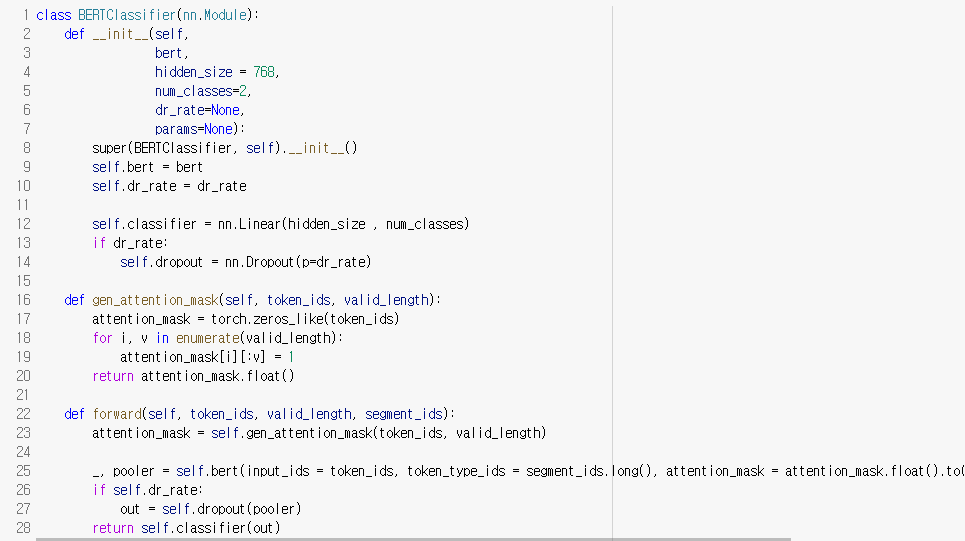
num\_discard\_samples는 데이터 상위 몇 개 row 제거할지 설정

(위 예제는 id document label이 적혀 있어서 1줄 삭제.)



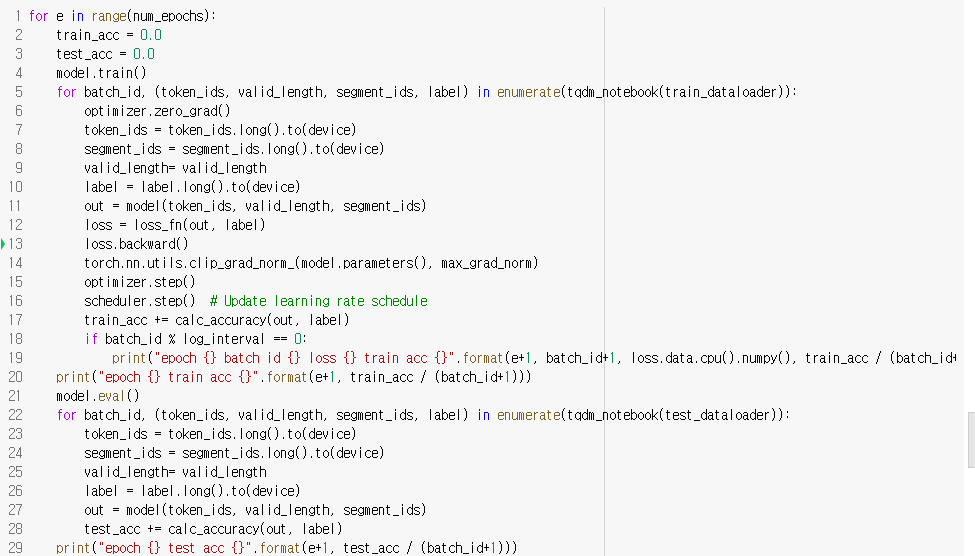
파라미터 설정

(max\_len : 텍스트 데이터 최대 길이, num\_epochs : 훈련 반복 횟수)

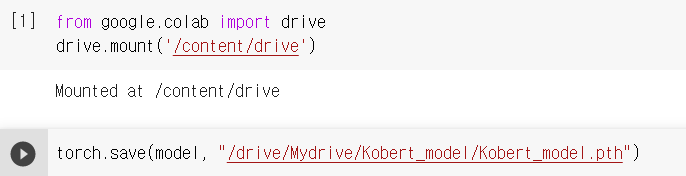


분류기 설정

Num\_class = 2 (2가지로 분류하겠다는 뜻.. 이진 분류기)



Train 및 test

학습시킨 model google drive에 저장!

불러올 때는 아래처럼…

*# 모델 클래스는 어딘가에 반드시 선언되어 있어야 합니다*

**model** **=** **torch.load(“/drive/Mydrive/Kobert\_model/Kobert\_model.pth”)**

Ps) 우리 졸과 주제는 Kobert에 korlex를 기반으로 학습 시킨 그래프 임베딩 데이터와 중앙기사 전처리 데이터를 추가 학습 시켜서 문맥 의존 철자 오류를 교정 시키는 것인데….

예를 들어, “병이 낳다” -> “병이 낫다”로 교정 시키는 건데

1. 중앙기사 데이터를 학습시킴으로써 어색한 문장을 찾고 어색한 문장들 중에서

Korlex를 이용해서 단어의 연관성을 찾아 어색한 문장을 교정하는 것인가?

아니면..

1. 틀린 문장을 가르쳐주고 이상한 단어를 찾아 교정하는 것인가?