BERT는 구글에서 공개한 Transformer 기반 사전 학습 언어 모델

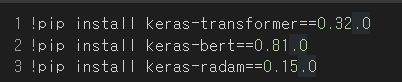


Multilingual BERT는 Google에서 발표한 BERT의 다국어 버전

104개 언어의 위키피디아 코퍼스를 모두 사용하여 pre-training을 진행

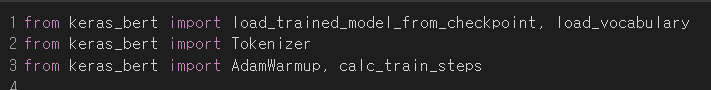
그 외에는 기존 BERT과 학습 과정이 다르지 않음

그럼에도 불구하고 Multilingual BERT는 다양한 언어 및 task에서 높은 일반화 성능을 보여줌

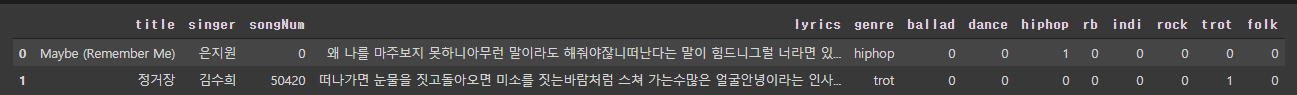


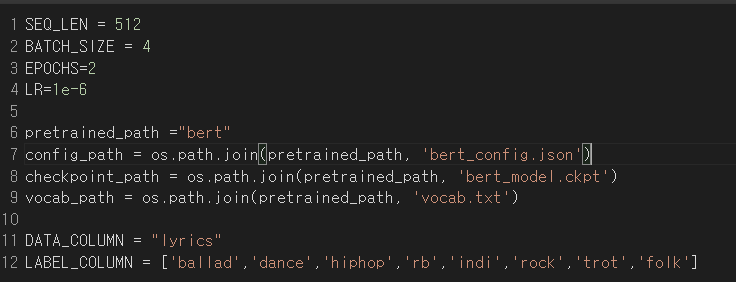
keras-bert 모듈 keras 에서 BERT 활용 쉽게 만들어줌

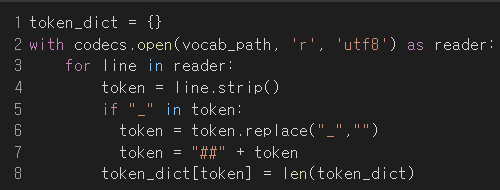
keras-radam 모듈 Adam optimizer 의 수정판



Bert 모형 활용에 필요한 모듈들

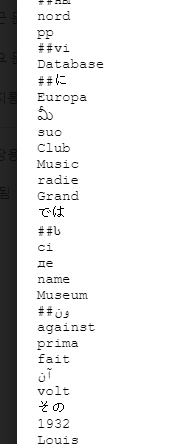






vocab.txt 에 있는 단어에 인덱스를 추가해주는 token\_dict 딕셔너리 생성

문장이 토큰화되고 그 다음에 인덱스로 변경되어 버트 신경망에 인풋으로 들어감



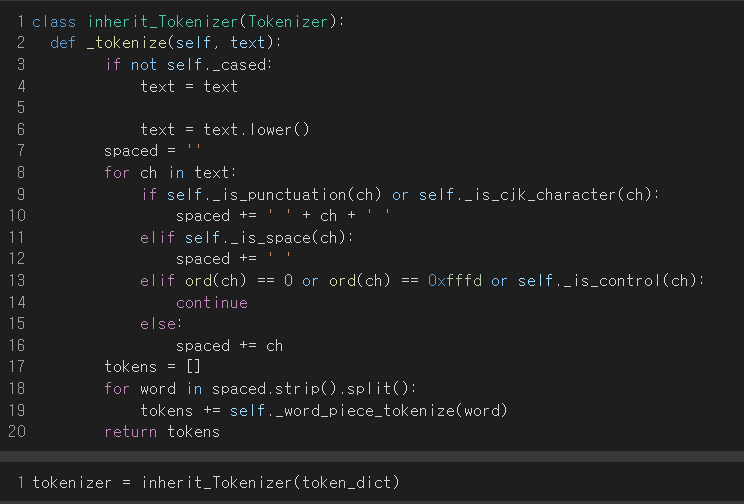
inherit\_Tokenizer라는 클래스(문장을 토큰화하는 기능)재정의 하고 \_tokenize 함수를 새로 작성

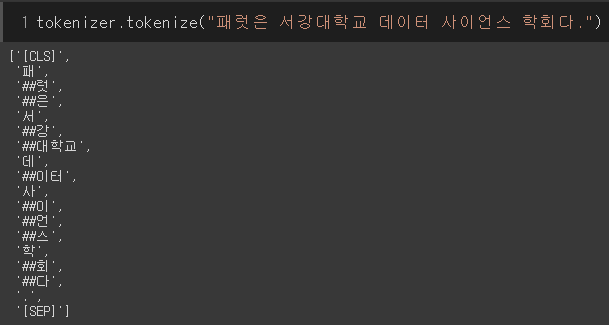
만약 원래 클래스를 그대로 사용하면 입력한 문장이 "완전자모분리"

"인공지능 할 수 있다" -> ㅇ,##ㅣ,##ㄴ,ㄱ,##ㅗ,##ㅇ,ㅈ,##ㅣ,##ㄴ,##ㅡ,##ㅇ

함수를 새로 만들면 -> "인,##공,##지,##능,할,수,있다"

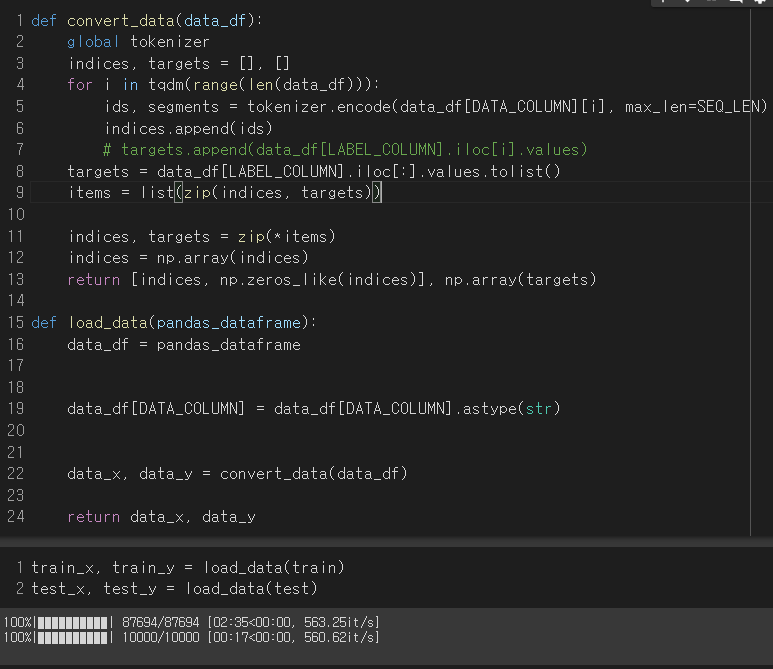
단어에 포함되면서 단어의 시작이 아닌 부분에는 ##가 붙는 것

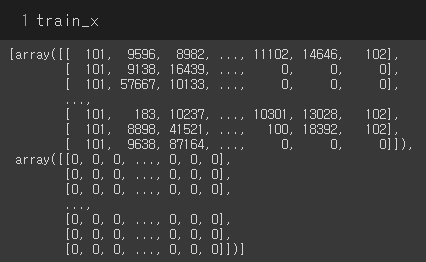


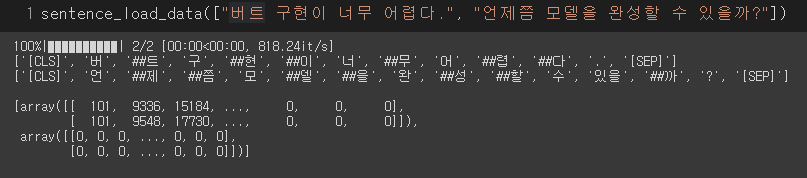


수집한 가사 데이터를 Bert 모형의 입력에 맞게 바꿔주는 함수 정의

tokenizer.encode 함수가 가사들을 토큰화해주고 토큰화 된 단어를 인덱스에 맞게 숫자로 바꿈







사전 학습된 버트 모델 인풋은

문장을 토큰화해서 이를 인덱스에 맞게 숫자로 바꾼 것과 앞문장인지 뒷문장인지 알려주는 토큰 임베딩

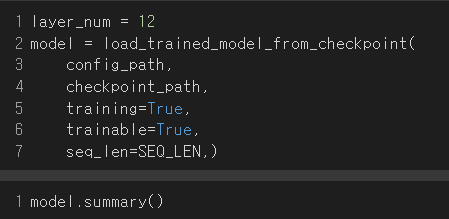
문장 하나를 가지고 훈련하므로 세그멘트 임베딩은 모두 0으로 통일

포지션 임베딩은 단어 순서에 따라 자동으로 부여됨

문장 안에 일부를 가려 예측하게 하는 마스킹은 사용하지 않음

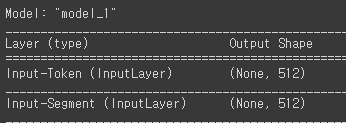


구글 깃허브에서 다운 받았던 사전학습된 모델을 코랩으로 로드



총 12층의 트랜스포머 층이 있음

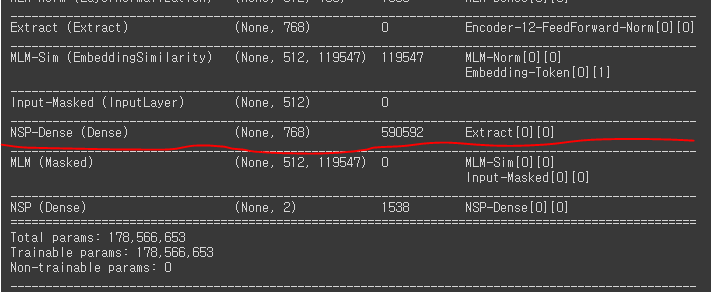
불러온 사전 학습 모델을 목적에 맞게 변형



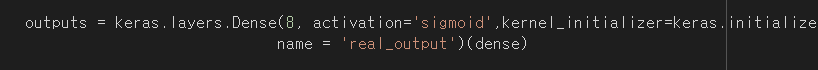


input : 토큰 벡터, 세그멘트 벡터

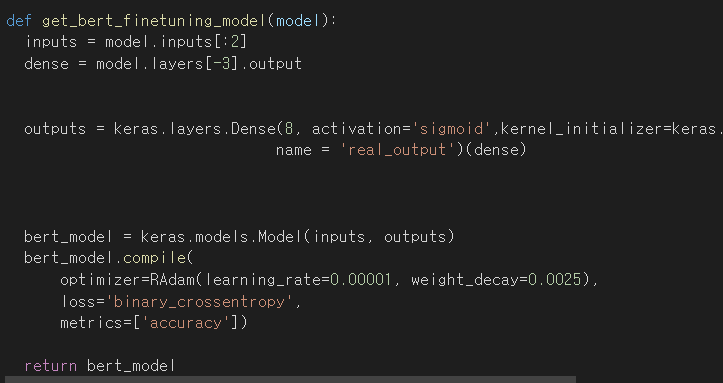
inputs = model.inputs[:2]

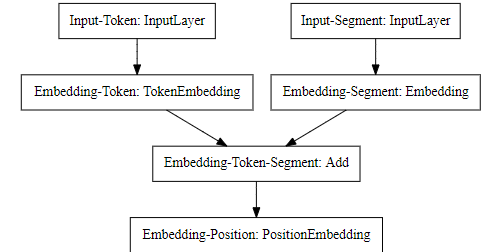


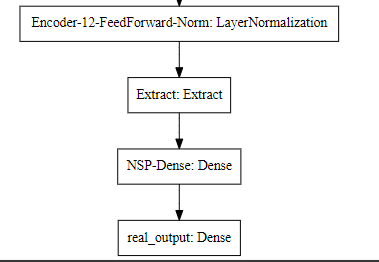




output



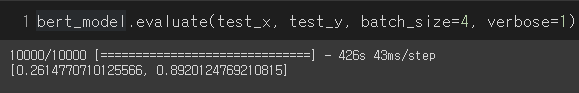




bert\_model을 get\_bert\_finetuning\_model 함수로 부름

bert\_model.fit으로 버트 모형 파인튜닝 시작





텐서플로우 2 에서 돌아가지 않음

노래에 한국어만 있는 것은 아니지만 한국어로 학습된 버트 모형을 이용해보기 (ETRI 에서 만든 한국어로 학습된 BERT 모형을 사용하면 정확도가 더 올라간다고 함)

