KoBERT 기반 감성분석 Fine-tuning Flow

■ 1. Corpus 수집 및 정제

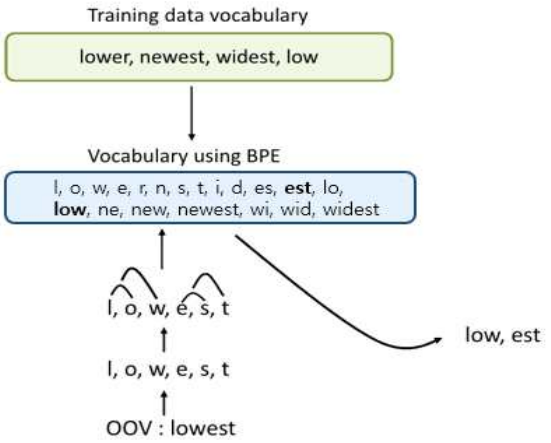
- 데이터 (Corpus - 문장/단어) 수집 및 정제(전처리)(개행문자, 특수문자, 중복 표현, 조사 제거, 띄어쓰기 보정 등)

input	ex) "나는밥을먹었습니다ㅎㅎㅎㅎㅎㅎㅎ"
output	ex) "나는 밥을 먹었습니다ㅎㅎ"

- Kobert : 한국어 위키 문장(5M) / 단어(54M)

■ 2. Tokenizing

- Subword 기반 분절 - BPE(Byte Pair Encoding) 기반 Tokenizer(=SentencePiece)가 위 문장들을 학습
- BPE는 글자(charcter) 단위에서 점차적으로 단어 집합(vocabulary)을 만들어 내는 **Bottom-up 방식의 접근 방법으로 OOV(Out of Vocabulary, 희귀 단어, 신조어) 문제를 완화 시켜 줌**
- 참고링크 : https://wikidocs.net/22592 : BPE 알고리즘



	Tokenizer
input	**문장/단어 Corpus들** ex) "삼성 주식은 정말 좋습니다", "삼성 반도체는 정말 좋은 것 같아"
output	**Tokenizer를 통해 만들어진 Vocabulary** ex)'삼','성','주','식','은','반','도','체','는','정','말'.'좋','습','니','다','은','것','같','아','삼성','주식','반도체','정말'

■ 3. 학습된 Tokenizer를 통해 학습 데이터 로드 및 변환

- 기존에 한국어 위키 문장을 Tokenizing 하고 학습한 Tokenizer 로드
- 학습하고자 하는 데이터를 로드 및 Tokenizing

	Tokenizer
case 1	**학습 데이터 중 OOV (신조어/은어 - 조합형) - 기존에 '갓삼성', '오뚜기' 가 학습 되어 있는 경우** tokenizer vocab : ['갓','삼','성','삼성','오','뚜','기','오뚜기'] tokenizer input : "갓오뚜기" tokenizer output : '갓_', '_오뚜기' **결과 : Subword에 존재하는 경우니 해당 조합을 학습**
case 2	**학습 데이터 중 OOV (신조어/은어 - 신규형) - 기존에 '헬뚜기' 관련 학습 안되어 있는 경우** tokenizer vocab : ['갓','삼','성','삼성','오','뚜','기','오뚜기'] tokenizer input : '헬뚜기' tokenizer output : '헬_', '_뚜_', '_기' **결과 : 한글자씩 찢으며, 대신 이 Sequence를 학습함**

- 지난번 피드백, Vocab (단어 사전) 구성 (ex-신조어 '갓뚜기'가 단어 사전에 들어 있으면 좋겠다) -> **없어도 가능하다**
- Fine-tuning 만으로도 해당 신조어를 잘 학습하여 충분한 성능을 낼수 있으며, 해당 Fine-Tuning 후 이용이 통상적인 방법 (다양한 블로그 및 오픈채팅방 답변))

■ 4. BERT 모델 문장 기반 학습 (Fine-tuning)

- 기존에 한국어 위키 문장을 학습 한 Weight 로드
- Embedding layer ~ Dense까지 End-to-End 학습 / Embedding layer는 학습한 단어들의 의미적, 문맥적 의미를 가짐

	Tokenizer
input	Tokenizer에 의해 정수인코딩 된 input 에 embedding layer를 거친 실수형 벡터
output	클래스 (긍정, 부정, 중립)

■ 5. 추론

- 테스트 데이터 3. Tokenizer 변환 후 4. 학습된 모델에 Predict ==> 긍/부/중립 여부 판단

■ 향후 방향

1. 신규 Corpus 구축 (많은 시간이 소요 될것으로 보임)

- 수집된 Corpus들을(전처리 까지 완료 된) 일반적으로 잘 공개 안함 / 전처리된 Vocab으로 Tokenizer 재학습 필요
- 한국어 위키(오픈데이터) 수집 가능 / EDA등을 거치며 전처리 필요 / 많은 데이터 필요, 전처리 다 처리할 수 있을지 의문
- Vocab 자체를 가지고 있으면 기본 + 기업에 맞게 Corpus 단 부터 수정하며 더 맞춤형으로 가능하긴 함

2. 기존에 Pre-trained된 모델 Fine-tuning 사용

- 다른 최신 BERT KcELECTRA, KcBERT 실험
- 두 모델 모두 KoBERT에 비해 큰 Vocab Size를 가지고 있어 더 좋은 성능이 예상됨. KcELECTRA (32000 vocab)
- 등장한지 1년 안팎 모델들로 아직 레퍼런스와 Star 수가 적어 실제 사용 가능한지는 테스트 예정

- (1) KcELECTRA (https://github.com/Beomi/KcELECTRA)
  - 기존 정제된 데이터로 학습한 모델 들에 비해, 신조어, 오탈자 등의 표현을 학습한 모델. (네이버 뉴스 댓글, 대댓글 등)
-(2) KcBERT (https://github.com/Beomi/KcBERT)