|  |  |
| --- | --- |
| 교육 제목 | AI 프로젝트 기반 S/W 전문가 양성과정 |
| 교육 일시 | 22.01.20. |
| 교육 장소 | 집 |
| **교육 내용** | |
| 오전 | 1. 어제 피드백 2. tensorflow의 subwordTextEncoder로 토큰화 3. sample 문장으로 encoding, decoding하기 tokenizer = tfds.deprecated.text.SubwordTextEncoder.build\_from\_corpus(train\_df["review"], target\_vocab\_size=2\*\*13) 4. encoding 🡪 부호화 tokenized\_string = tokenizer.encode(sample\_string) 5. decoding 🡪 복호화 original\_string = tokenizer.decode(tokenized\_string) 6. vocab size 확인 tokenizer.vocab\_size 7. 가정 설정문(assert) assert는 뒤의 조건이 True가 아니면 AssertError를 발생한다. |
| 오후 | 1. 데이터 전처리 - 데이터의 중복 제거 - NaN 결측치 제거 - 불용어(stopwords) 제거 - 한국어 토크나이저로 토큰화 - 사전 word\_to\_index 구성 - 텍스트 스트링을 사전 인덱스 스트링으로 변환 - x\_train, y\_train, x\_test, y\_test, word\_to\_index 리턴 2. Boost Camp 동영상 강의 시청 3. BOW 🡪 Bag of Words 문서 집합에서 단어 토큰을 생성하고 각 단어의 수를 세어 BOW 인코딩 벡터를 만든다. 4. 기본 from sklearn.feature\_extraction.text import CountVectorizer vector = CountVectorizer() # corpus로부터 각 단어의 빈도수를 기록 print("bag of words :", vector.fit\_transform(sentence).toarray()) # 각 단어의 인덱스가 어떻게 부여됐는지 보여줌 print("각 단어의 인덱스 :", vector.vocabulary\_) 5. 사용자 정의 불용어 vect = CountVectorizer(stop\_words = ["the", "a", "an", "is", "not"]) 6. CountVectorizer에서 자체 제공하는 불용어 vect = CountVectorizer(stop\_words = "english") 7. nltk에서 지원하는 불용어 !pip install nltk import nltk nltk.download("stopwords") from sklearn.feature\_extraction.text import CountVectorizer from nltk.corpus import stopwords sw = stopwords.words("english") vect = CountVectorizer(stop\_words = sw) 아래는 같음 8. 실습 |