

인공지능기초응용 II

6 주차 과제

인공지능응용

K2025029 금동환

목차

1. [첨부파일 6 쪽의 sentencePiece 토큰라이저 실습 및 실습화면 스샷](#)
2. [testSPM-ko.py 실습화면 스샷](#)

1. 첨부파일 6 쪽의 sentencePiece 토큰라이저 실습 및 실습화면 스샷

A. Train

```

1 import sentencepiece as spm
2
3 # train
4 spm.SentencePieceTrainer.train( Unresolved attribute reference 'train' for class 'SentencePieceTrainer'
5     '--input=kowiktext_20200920.test --model_prefix=kowiktext --vocab_size=8000 --character_coverage=0.9995'
6 )
7
8
9 # en/decoding
10 # sp = spm.SentencePieceProcessor()
11 # sp.load('kowiktext.model')
12 #
13 #
14 # text = "국민대학교 소프트웨어융합대학원 K2025029 김동환 인공지능기초응용II 6주차 과제입니다."
15 # print(sp.encode_as_pieces(text))
16 # print(sp.encode_as_ids(text))
17 #
18 # print('-----')
19 # print(sp.decode_ids(sp.encode_as_ids(text)))
20

```

```

trainer_interface.cc(560) LOG(INFO) Alphabet size=3823
trainer_interface.cc(561) LOG(INFO) Final character coverage=0.9995
trainer_interface.cc(592) LOG(INFO) Done! preprocessed 88139 sentences.
unigram_model_trainer.cc(265) LOG(INFO) Making suffix array...
unigram_model_trainer.cc(269) LOG(INFO) Extracting frequent sub strings... node_num=2366488
unigram_model_trainer.cc(312) LOG(INFO) Initialized 180248 seed sentencepieces
trainer_interface.cc(598) LOG(INFO) Tokenizing input sentences with whitespace: 88139
trainer_interface.cc(609) LOG(INFO) Done! 217827
unigram_model_trainer.cc(602) LOG(INFO) Using 217827 sentences for EM training
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=103270 obj=16.9724 num_tokens=560816 num_tokens/piece=5.43058
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=93076 obj=15.5238 num_tokens=564895 num_tokens/piece=6.06918
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=69800 obj=15.6226 num_tokens=588452 num_tokens/piece=8.43054
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=69761 obj=15.5641 num_tokens=588802 num_tokens/piece=8.44027
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=52318 obj=15.8524 num_tokens=620752 num_tokens/piece=11.865
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=52317 obj=15.7833 num_tokens=620779 num_tokens/piece=11.8657
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=39237 obj=16.1322 num_tokens=654691 num_tokens/piece=16.6856
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=39237 obj=16.0529 num_tokens=654704 num_tokens/piece=16.6859
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=29427 obj=16.4668 num_tokens=690084 num_tokens/piece=23.4507
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=29427 obj=16.3836 num_tokens=690223 num_tokens/piece=23.4554
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=22070 obj=16.8539 num_tokens=726830 num_tokens/piece=32.9329
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=22070 obj=16.7662 num_tokens=726840 num_tokens/piece=32.9334
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=16552 obj=17.2937 num_tokens=765662 num_tokens/piece=46.258
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=16552 obj=17.198 num_tokens=765674 num_tokens/piece=46.2587
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=12414 obj=17.787 num_tokens=806238 num_tokens/piece=64.9459
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=12414 obj=17.6807 num_tokens=806260 num_tokens/piece=64.9476
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=9310 obj=18.3622 num_tokens=851469 num_tokens/piece=91.4575
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=9310 obj=18.2309 num_tokens=851477 num_tokens/piece=91.4583
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=0 size=8800 obj=18.3716 num_tokens=859726 num_tokens/piece=97.6961
unigram_model_trainer.cc(618) LOG(INFO) EM sub_iter=1 size=8800 obj=18.3469 num_tokens=859730 num_tokens/piece=97.6966
trainer_interface.cc(687) LOG(INFO) Saving model: kowiktext.model
trainer_interface.cc(699) LOG(INFO) Saving vocabs: kowiktext.vocab

Process finished with exit code 0

```

B. En/decoding

```

1 import sentencepiece as spm
2
3 # train
4 # spm.SentencePieceTrainer.train(
5 #     '--input=kowiktext_20200920.test --model_prefix=kowiktext --vocab_size=8000 --character_coverage=0.9995'
6 # )
7
8 # en/decoding
9
10 sp = spm.SentencePieceProcessor()
11 sp.load('kowiktext.model')
12
13
14 text = "국민대학교 소프트웨어융합대학원 K2025029 김동환 인공지능기초응용II 6주차 과제입니다."
15 print(sp.encode_as_pieces(text))
16 print(sp.encode_as_ids(text))
17 print('-----')
18 print(sp.decode_pieces(sp.encode_as_pieces(text)))
19 print(sp.decode_ids(sp.encode_as_ids(text)))
20

```

Errors in the code:

- Line 11: `sp.load('kowiktext.model')` - Unresolved attribute reference 'load' for class 'SentencePieceProcessor'
- Line 15: `print(sp.encode_as_pieces(text))` - Unresolved attribute reference 'encode_as_pieces' for class 'SentencePieceProcessor'
- Line 16: `print(sp.encode_as_ids(text))` - Unresolved attribute reference 'encode_as_ids' for class 'SentencePieceProcessor'
- Line 18: `print(sp.decode_pieces(sp.encode_as_pieces(text)))` - Unresolved attribute reference 'encode_as_pieces' for class 'SentencePieceProcessor'
- Line 19: `print(sp.decode_ids(sp.encode_as_ids(text)))` - Unresolved attribute reference 'decode_ids' for class 'SentencePieceProcessor'

Console Output:

```

D:\MyProject\kmu\kmu-AI-II\.env\Scripts\python.exe D:\MyProject\kmu\kmu-AI-II\6강\test-SentencePiece.py
['_국민', '대학교', '_소', '프', '트', '웨', '어', '응', '합', '대학원', '_K', '20', '250', '29', '_금', '동', '환', '_', '인', '공', '지', '능', '기', '초', '응', '용', 'II', '6', '주', '차', '과', '제', '입', '니', '다', '.']
[1981, 703, 195, 205, 64, 1605, 43, 2597, 547, 2833, 414, 442, 4242, 1080, 852, 87, 548, 3, 25, 212, 30, 858, 35, 667, 1964, 175, 475, 475, 122, 55, 162, 3, 26, 73, 849, 5]
-----
국민대학교 소프트웨어융합대학원 K2025029 김동환 인공지능기초응용II 6주차 과제입니다.
국민대학교 소프트웨어융합대학원 K2025029 김동환 인공지능기초응용II 6주차 과제입니다.
Process finished with exit code 0

```

2. testSPM-ko.py 실습화면 스샷

The screenshot shows a code editor with the file `testSPM-ko.py` open. The code defines a `SentencePieceProcessor` and trains it on three datasets: `ITnews1000.txt`, `gileg.txt`, and `ko_wiki_text.txt`. It then loads the trained model and processes a test sentence: "행정안전부는 오는 20일부터 버스·전차 등 대중교통과 마트 등 대형시설 안의 개방형 약국에서 마스크 착용 의무를 해제한다. 전 세계적으로 '켓' 돌풍을 일으킨 생성형 인공지능(AI) 챗GPT 개발사인 오픈AI가 14일(현지시간) 더욱 강력해진 새로운 오픈AI GPT-4를 소개했다. '켓' 돌풍을 일으킨 생성형 인공지능(AI) 챗GPT 개발사인 오픈AI가 14일(현지시간) 더욱 강력해진 새로운 오픈AI GPT-4를 소개했다."

The console output shows the training progress and the final model state. It includes logs for adding meta-piece, normalizing sentences, character coverage, alphabet size, and the final preprocessing of 11990 sentences. The output also shows the number of tokens and the size of the vocabulary for each dataset.

```

1 import sentencepiece as spm
2 sp = spm.SentencePieceProcessor()
3
4 spm.SentencePieceTrainer.Train('--input=ITnews1000.txt --model_prefix=ITnews --vocab_size=8000')
5 #spm.SentencePieceTrainer.Train('--input=gileg.txt --model_prefix=gileg --vocab_size=16000')
6 #spm.SentencePieceTrainer.Train('--input=ko_wiki_text.txt --model_prefix=ko_wiki --vocab_size=64000')
7
8 sp.load('ITnews.model') Unresolved attribute reference 'load' for class 'SentencePieceProcessor'
9 #sp.load('gileg.model') dhkeum9886, 2025-04-01 오전 11:49 + 6강 업데이트
10 #sp.load('ko_wiki.model')
11
12 # encode: text => id
13 text='행정안전부는 오는 20일부터 버스·전차 등 대중교통과 마트 등 대형시설 안의 개방형 약국에서 마스크 착용 의무를 해제한다. 전 세계적으로 '켓' 돌풍을 일으킨 생성형 인공지능(AI) 챗GPT 개발사인 오픈AI가 14일(현지시간) 더욱 강력해진 새로운 오픈AI GPT-4를 소개했다. '켓' 돌풍을 일으킨 생성형 인공지능(AI) 챗GPT 개발사인 오픈AI가 14일(현지시간) 더욱 강력해진 새로운 오픈AI GPT-4를 소개했다.'
14 print(sp.encode_as_pieces(text)) Unresolved attribute reference 'encode_as_pieces' for class 'SentencePieceProcessor'
15 print(sp.encode_as_ids(text)) Unresolved attribute reference 'encode_as_ids' for class 'SentencePieceProcessor'
16
17 text2='전 세계적으로 '켓' 돌풍을 일으킨 생성형 인공지능(AI) 챗GPT 개발사인 오픈AI가 14일(현지시간) 더욱 강력해진 새로운 오픈AI GPT-4를 소개했다. '켓' 돌풍을 일으킨 생성형 인공지능(AI) 챗GPT 개발사인 오픈AI가 14일(현지시간) 더욱 강력해진 새로운 오픈AI GPT-4를 소개했다.'
18 print(sp.encode_as_pieces(text2)) Unresolved attribute reference 'encode_as_pieces' for class 'SentencePieceProcessor'
19 print(sp.encode_as_ids(text2)) Unresolved attribute reference 'encode_as_ids' for class 'SentencePieceProcessor'
20
21 # decode: id => text
22 #print(sp.decode_pieces([1212, 32, 10, 587, 446]))
23 #print(sp.decode_ids([1212, 32, 10, 587, 446]))
24

```

Process finished with exit code 0