과제 #9

201904479 컴퓨터 전자 시스템 학과 임서연

1. 이 자료의 8~12쪽을 참고하여 다음 문장의 토큰수를 최대 128로 하고 tokenize 결과 (9쪽 형식), 인코딩 결과(11쪽 형식), attention mask(12쪽 형식)를 구하라.

문장: ‘여기서 주의할 점은 앞의 2 번과 뒤에 붙은 3 번은 원래 있던 단어가 아니라는 점입니다.’

Bert Tokenizer

```
['여기', '##서', '주의', '##할', '점', '##은', '앞', '##의', '2', '번', '##과', '뒤', '##에', '붙', '##은', '3', '번', '##은', '원래', '있', '##던', '단어', '##가', '아니', '##라는', '점', '##입니
```

Tokenizer padding

```
max_seq_len = 128
encoded_result = tokenizer.encode("여기서 주의할 점은 앞의 2 번과 뒤에 붙은 3 번은 원래 있던 단어가 아니라는 점입니다.",
max_length=max_seq_len, pad_to_max_length=True)
print(encoded_result)
print("길이 : ", len(encoded_result))
```

```
[2, 3776, 2112, 4924, 2085, 1540, 2073, 1388, 2079, 22, 1141, 2145, 873, 2170, 1181, 2073, 23, 1141, 2073, 5057, 1513, 2414, 5591, 2116, 3614, 23548, 1540, 12190, 18, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0,
길이 :  128
/usr/local/lib/python3.8/dist-packages/transformers/tokenization_utils_base.py:2336: FutureWarning: The `pad_to_max_length` argument is deprecated and will be removed in a future version, use `pad
  warnings.warn(


 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

padding=True` or `padding='longest'` to pad to the longest sequence in the batch, or use `padding='max_length'` to pad to a max length. In this case, you can give a specific length with `max_length
```

Attention Mask

```
tf.Tensor(
[[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]], shape=(1, 128), dtype=int32)
```

2. [마스크 실습] 이 자료의 15~16쪽을 참고하여 다음 문장에 대한 마스크 처리 결과를 구하라.
문장: '이 영화의 초점은 [MASK]다.'

```
inputs = tokenizer('이 영화의 초점은 [MASK]다.', return_tensors='tf')
print(inputs['input_ids'])
```

```
tf.Tensor([[   2 1504 3771 2079 6392 2073    4  809   18    3]], shape=(1, 10), dtype=int32)
```

```
[{'score': 0.05608559027314186,
  'token': 3657,
  'token_str': '하나',
  'sequence': '이 영화의 초점은 하나 다.'},
 {'score': 0.04044365510344505,
  'token': 1504,
  'token_str': '이',
  'sequence': '이 영화의 초점은 이 다.'},
 {'score': 0.03929238021373749,
  'token': 1417,
  'token_str': '었',
  'sequence': '이 영화의 초점은 었 다.'},
 {'score': 0.03794778138399124,
  'token': 5969,
  'token_str': '남녀',
  'sequence': '이 영화의 초점은 남녀 다.'},
 {'score': 0.029111366719007492,
  'token': 35,
  'token_str': '?',
  'sequence': '이 영화의 초점은? 다.'}]
```

3. [네이버 영화평] 이 자료의 17~25쪽의 프로그램들을 순차적으로 실행하여 영화평 분류기를 수행한다. 계산시간이 많이 걸리므로 훈련용, 테스트용 영화평을 각각 15000개와 5000개로 줄여서 실행한다. 훈련을 마친 후 25쪽과 유사하게 다음 영화평에 대한 결과를 구하라.
영화평: '시간은 그럭저럭 때울 수 있기도', '감독은 노력했지만 결과는 그다지...'

```
[{'label': 'LABEL_0', 'score': 0.913663923740387},
 {'label': 'LABEL_1', 'score': 0.08633603155612946}]
```

```
[{'label': 'LABEL_0', 'score': 0.9568098187446594},
 {'label': 'LABEL_1', 'score': 0.04319017007946968}]
```