

BERT base cased

<Bert>

<https://wikidocs.net/115055>

<Tokenize & Embedding>

tokenize

```
16 tokenizer = BertTokenizer.from_pretrained('bio-bert', do_lower_case=False)
17
18
19 tokenized_texts = [tokenizer.tokenize(sent) for sent in sentences]
20
21 print (sentences[0])
22 print (tokenized_texts[0])
23 print (sentences[2])
24 print (tokenized_texts[2])
```

```
[CLS] gene is mutation. [SEP]
['[CLS]', 'gene', 'is', 'mutation', '.', '[SEP]']
[CLS] BLANK1 is BLANK2. [SEP]
['[CLS]', 'BLANK1', 'is', 'BLANK2', '.', '[SEP]']
```

```
1 max_len=10
2
3 # 입력 토큰의 최대 시퀀스 길이
4 MAX_LEN = max_len
5
6 # 토큰을 숫자 인덱스로 변환
7 input_ids = [tokenizer.convert_tokens_to_ids(x) for x in tokenized_texts]
8
9 # 문장을 MAX_LEN 길이에 맞게 자르고, 모자란 부분을 패딩 0으로 채움
10 input_ids = pad_sequences(input_ids, maxlen=MAX_LEN, dtype="long", truncating="post", padding="post")
11
12 print(input_ids[0])
13 print(input_ids[2])
```

```
[ 101 5563 1108 17894 117 102 0 0 0 0]
[ 101 5564 1108 17895 117 102 0 0 0 0]
```

vocab.txt에서 몇 번째에 있는 단어 인지에 의해 numbering 됩니다.

```

Sentence : gene is mutation.
Token : ['[CLS]', 'gene', 'is', 'mutation', '.', '[SEP]']
Number : [101, 5563, 1108, 17894, 117, 102]
Embed : tensor([[ 0.4554,  0.0531, -0.2298, ...,  0.0140,  0.0527, -0.1251], ➡ [CLS]
               [-0.6063,  0.0259, -1.1309, ..., -0.4874, -0.0078, -0.3163], ➡ gene
               [-0.9179,  0.3058,  0.5611, ...,  0.4600, -0.0942,  0.4065], ➡ is
               [ 0.2499, -0.8159,  0.6224, ...,  0.1767, -1.1238, -1.4970], ➡ mutation
               [-1.0499,  0.4114, -0.4734, ...,  0.5776, -1.4039, -0.4193], ➡ .
               [-0.2852,  0.0242,  0.2498, ...,  0.7378, -0.9040,  0.2615]], ➡ [SEP]
      grad_fn=<SelectBackward>)

Sentence : BLANK1 is BLANK2.
Token : ['[CLS]', 'BLANK1', 'is', 'BLANK2', '.', '[SEP]']
Number : [101, 5564, 1108, 17895, 117, 102]
Embed : tensor([[ 0.4554,  0.0531, -0.2298, ...,  0.0140,  0.0527, -0.1251],
               [-0.9248,  0.3694, -1.2119, ...,  0.1737,  0.9775,  0.4975],
               [-0.9179,  0.3058,  0.5611, ...,  0.4600, -0.0942,  0.4065],
               [ 0.7827,  1.1724,  0.4117, ...,  0.1014,  0.0028, -0.7993],
               [-1.0499,  0.4114, -0.4734, ...,  0.5776, -1.4039, -0.4193],
               [-0.2852,  0.0242,  0.2498, ...,  0.7378, -0.9040,  0.2615]],
      grad_fn=<SelectBackward>)

```

각 문장에서의 embedding 결과입니다. 각 단어 마다 768개의 숫자로 변환되었습니다.