BERT base cased

<Bert>
https://wikidocs.net/115055

<Tokenize & Embedding>

tokenize

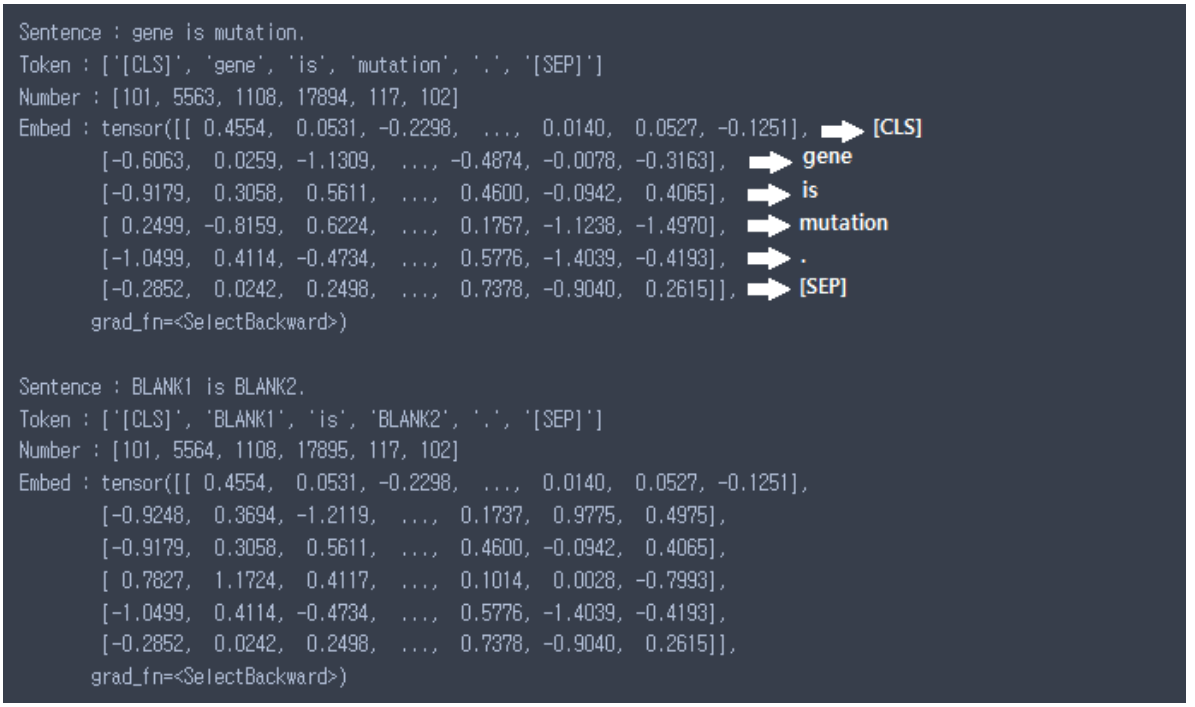
```
tokenizer = BertTokenizer.from_pretrained('bio-bert', do_lower_case=False)


tokenized_texts = [tokenizer.tokenize(sent) for sent in sentences]

print (sentences[0])
print (tokenized_texts[0])
print (sentences[2])
print (tokenized_texts[2])
```

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

```
max_len=10

# 입력 토큰의 최대 시퀀스 길이
MAX_LEN = max_len

# 토큰을 숫자 인덱스로 변환
input_ids = [tokenizer.convert_tokens_to_ids(x) for x in tokenized_texts]

# 문장을 MAX_LEN 길이에 맞게 자르고, 모자란 부분을 패딩 0으로 채움
input_ids = pad_sequences(input_ids, maxlen=MAX_LEN, dtype="long", truncating="post", padding="post")

print(input_ids[0])
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.



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