17장 BERT
(Bidirectional Encoder Representations from Transformers)

BERT에 대해 알아봅시다!

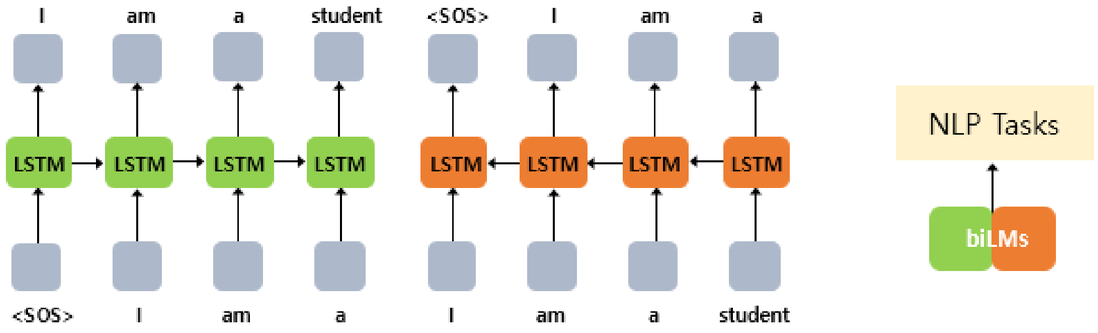
2023년도 동계인턴 스터디 5주차
박성호

17-1 Pre-Training까지의 흐름

Word2Vec, FastText, Glove는 다의어, 동음이의어를 구분하지 못한다

- 2015년 구글은 ‘Semi-supervised Sequence Learning’ 에서 레이블이 없는 데이터로 학습된 LSTM, 가중치가 랜덤으로 초기화 된 LSTM 두 가지를 비교
- 방대한 텍스트로 LSTM LM을 학습해놓으면 다른 태스크에서 높은 성능을 얻을 수 있다!

ELMo: Deep Contextual Word Embedding, AI2 & University of Washington, 2017



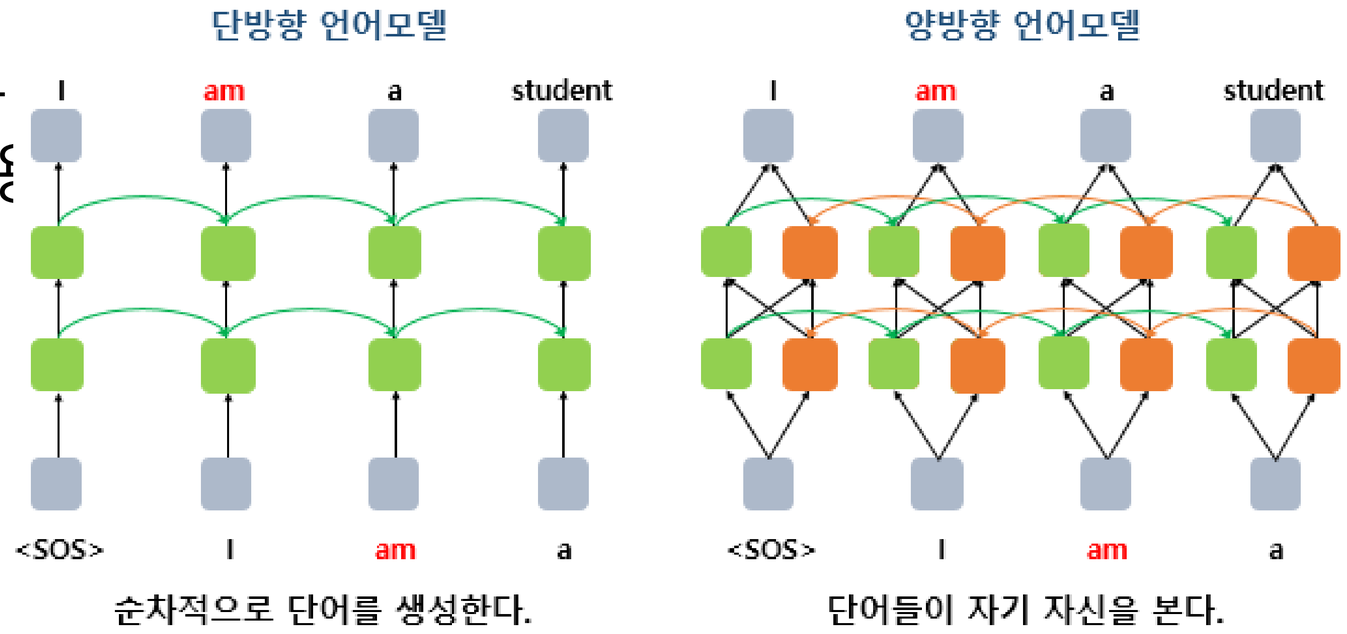
순방향 언어 모델과 역방향 언어 모델을 각각 훈련

사전 훈련된 임베딩에 사용

Transformer

- 변역기와 같은 인코더-디코더에서 LSTM을 뛰어넘은 Trm으로 사전학습
- 언어는 양방향으로 문맥을 가지므로 뒤의 정보도 필요하다.

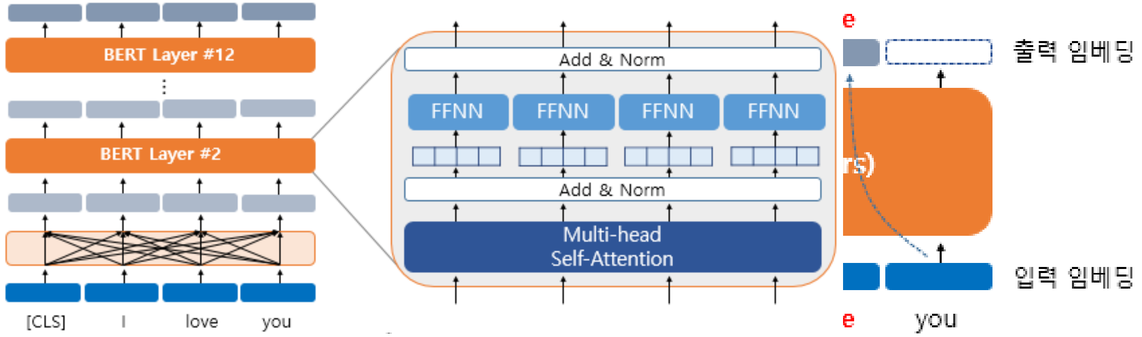
- 따라서 ELMo는 순방향과 역뱡향을
  갖는 두 개의 단방향 합친 LM을 이용



→ Masked Language Model의 등장!

17-2 BERT

레이블이 없는 방대한 데이터로 사전 훈련된 모델



- BooksCorpus(8억 단어)

- 다른 모델보다 뛰어난 성능

- Base는 L=12, D=768, A=12: 110M개의 파라미터
  Large는 L=24, D=1024, A=16: 340M개의 파라미터

WordPiece 토크나이저

	0
0	[PAD]
1	[unused0]
2	[unused1]
3	[unused2]
4	[unused3]
...	...
30517	## .
30518	## /
30519	## :
30520	## ?
30521	## ~

30522 rows × 1 columns

the sentence I want embbedings for.”

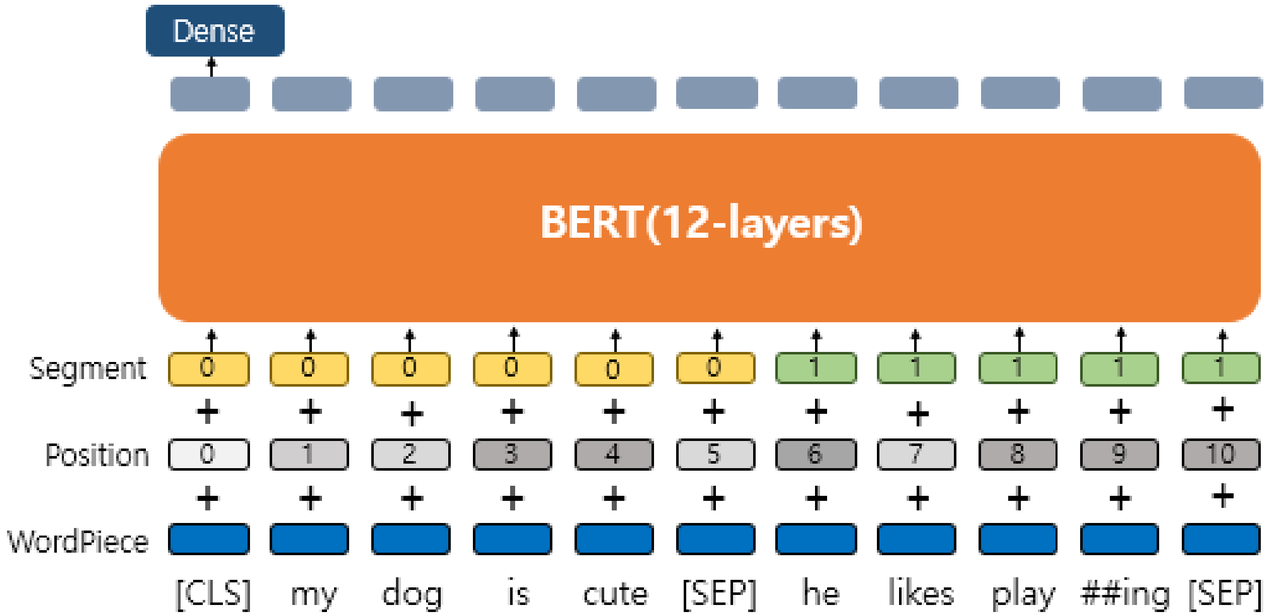
’, ‘is’, ‘the’, ‘sentence’, ‘i’, ‘want’, ‘em’, ‘##bed’, ‘##ding’, ‘##s’, ‘for’, ‘.’]

어 집합의 크기는 30,522개

‘##ding’은 4667번째 단어로 저장돼 있다.

’AD]: 0 [UNK]: 100 [CLS]: 101 [SEP]: 102 [MASK]: 103)

Position Embedding과 Segment Embedding

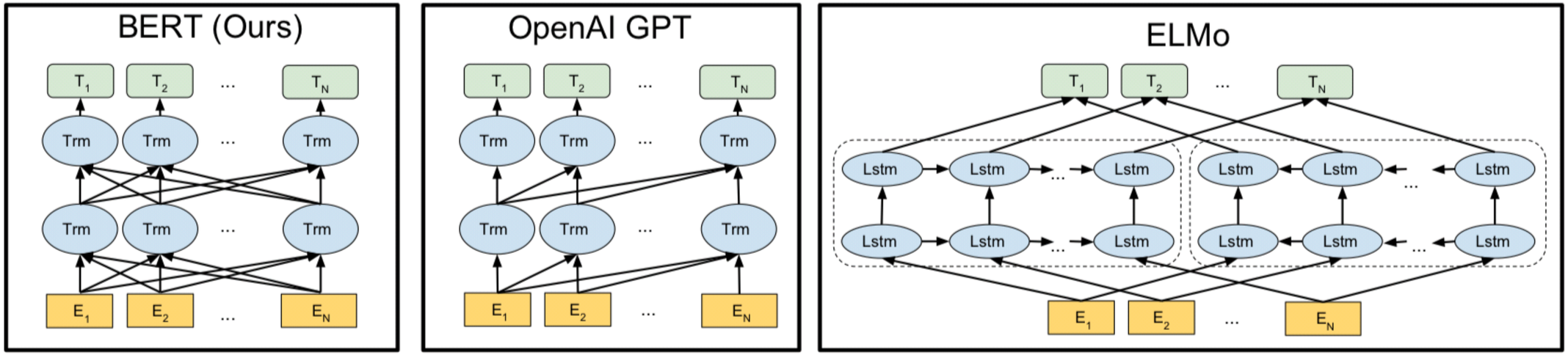


ㅏ 코사인 함수를 사용하여 위치에 따
: 학습을 통해서 얻는 포지션 임베딩

= 1)

문장은 실제로는 두 종류의 텍스트

BERT의 Pre-training



BERT의 Pre-training(2)

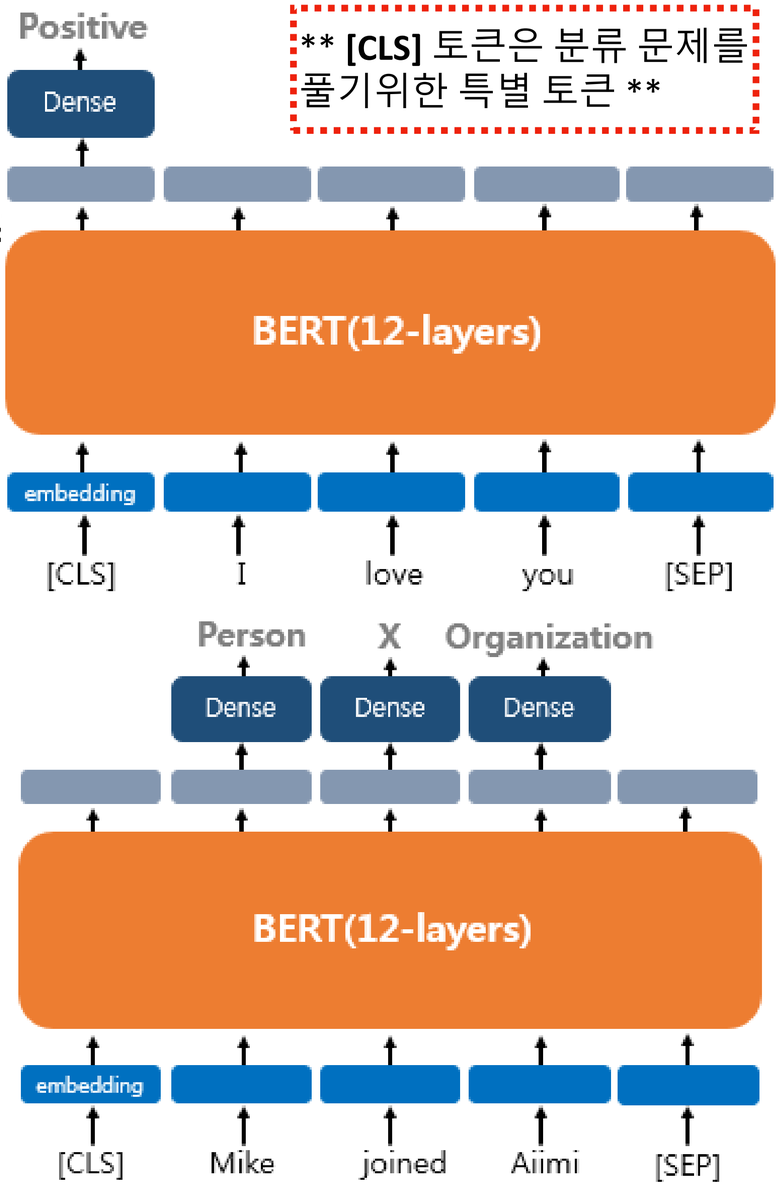
- Next Sentence Prediction(NLP):두 개의 문장이 이어지는지를 맞추는 방식으로 훈련
  50:50 비율로 실제 이어지는(IsNext), 랜덤으로 이어 붙인(NotNext) 문장 두 개를 주고
  [CLS] 토큰의 위치의 출력층에서 이진 분류 문제를 푼다. → QA, NLI에서 활용

- *"Despite its **simplicity,** we demonstrate in Section 5.1 that pre-training towards this task is **very beneficial** to both QA and NLI. "*
  (Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova."BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" *arXiv:1810.04805*

BERT의 Fine-tuning (1)

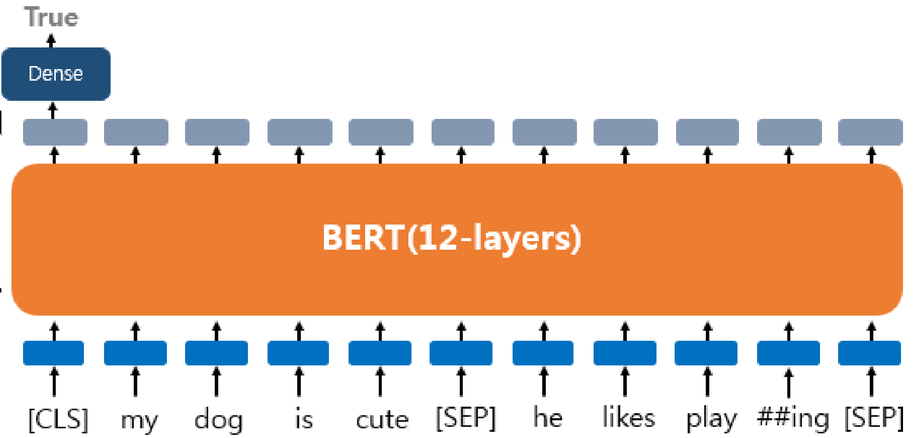
- [CLS] 토큰의 위치의 출력층에서 Dense layer를 추가하여 분류에 대한 예측

- 태깅 작업(PoS, NER 등)

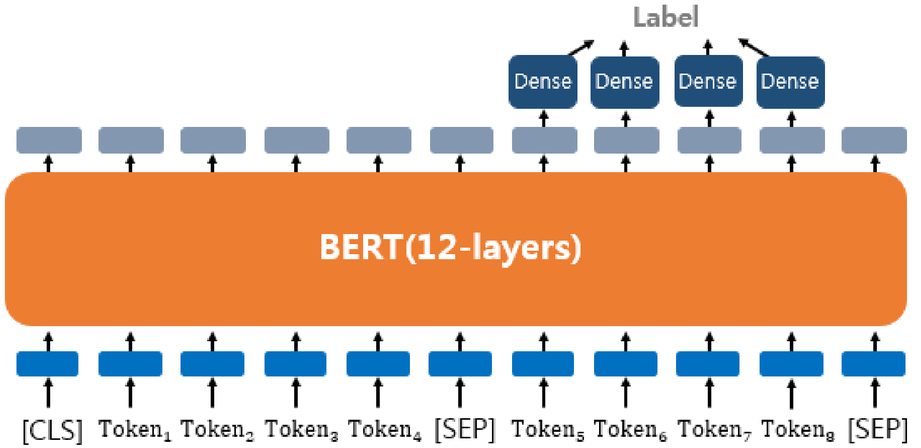


BERT의 Fine-tuning (2)

- 텍스트의 쌍에 대한 분류 또는 회귀 문
  NLI에서도 사용되는데 NLI는
  두 문장이 모순, 함의, 중립 관계인지 구
  [SEP] 토큰과 세그먼트 임베딩 사용



- 질문과 본문이라는
  두 개의 텍스트의 쌍을 입력받아
  본문의 일부분을 추출해서 질문에 답변
  ex)강우가 떨어지도록 영향을 주는 것은
  "기상학에서 강우는 대기 수증기가
  응결되어 중력의 영향을 받고 떨어지는



17-4 한국어 BERT의 MLM

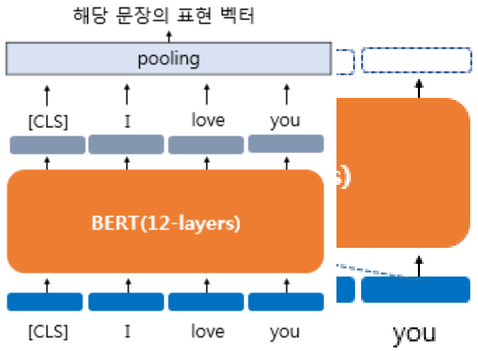
```
from transformers import TFBertForMaskedLM
from transformers import AutoTokenizer
from transformers import FillMaskPipeline

model = TFBertForMaskedLM.from_pretrained('klue/bert-base', from_pt=True)
tokenizer = AutoTokenizer.from_pretrained("klue/bert-base")
pip = FillMaskPipeline(model=model, tokenizer=tokenizer)
pip('축구는 정말 재미있는 [MASK]다.')
```

```
Some weights of the PyTorch model were not used when initializing the TF 2.0 mo
- This IS expected if you are initializing TFBertForMaskedLM from a PyTorch moc
- This IS NOT expected if you are initializing TFBertForMaskedLM from a PyTorch
All the weights of TFBertForMaskedLM were initialized from the PyTorch model.
If your task is similar to the task the model of the checkpoint was trained on,
[{'score': 0.8963516354560852,
  'token': 4559,
  'token_str': '스포츠',
  'sequence': '축구는 정말 재미있는 스포츠 다.'},
 {'score': 0.02595745585858822,
  'token': 568,
  'token_str': '거',
  'sequence': '축구는 정말 재미있는 거 다.'},
 {'score': 0.010033913888037205,
  'token': 3682,
  'token_str': '경기',
  'sequence': '축구는 정말 재미있는 경기 다.'},
 {'score': 0.007924334146082401,
  'token': 4713,
  'token_str': '축구',
  'sequence': '축구는 정말 재미있는 축구 다.'},
 {'score': 0.007844174280762672,
  'token': 5845,
  'token_str': '놀이',
  'sequence': '축구는 정말 재미있는 놀이 다.'}]
```

17-7 Sentence BERT(SBERT)

문장 임베딩을 얻을 수 있는 BERT



터를 얻는 세 가지 방법

| 출력 벡터

| 출력 벡터에 평균 풀링을 수행한 벡터 (모든 단어 의미를 골고

3. 모든 단어의 출력 벡터에 맥스 풀링을 수행한 벡터 (모든 단어 의미를 다르게 반영)

18-9 Facebook AI Similarity Search(FAISS)

효율적인 유사도 검색을 위한 라이브러리

- 수십억 데이터 셋에서 GPU를 사용한 첨단 k-selection algorithm 보다 8.5배나 빠른,
- 기존 SQL의 데이터 처리 방식(해시 기반, 1D 간격 검색)은 고차원 벡터의 딥러닝에서
  비효율적 →FAISS: 메모리에 최적화된 여러 유사성 검색 방법, **최신 GPU 구현 제공**
- 고정된 메모리 사용량으로 정확도와 검색 시간의 최적화된 균형을 자동 튜닝 메커니즘
- CPU 친화적인 알고리즘(힙 정렬 등)이 아닌 효율적인 타일링 전략, 데이터 샤딩 + 벤치를 Lua로 인코딩하면 같은 정확도 손실 없이 속도가 향상된다는 사실을 발견

Faiss: a bunch of the painful work of paying attention to tasking engineering details.

인덱스 객체

- 결론: IndexFlatL2 > IndexHNSWFlat > IndexFlatIP

ch를 위한 전처리를 한 Index를 생

- 대부분의 index 기법에서는 벡터 분포를 분석하기 위해 학습 단계가 필요
- 학습 단계까지 마치면, index에 대해서 add / search 연산이 가능
  - add: index에 vector들을 추가함
  - search: index에서 query 행렬과 유사한 벡터를 검색해서 k개 반환(기본 k-NN 서치)
    D는 nearest-neighbor와 query간의 거리
    I는 nearest-neighbor의 database index

..?🤔

인덱스 가이드 라인

How big is the dataset?

a few essential questions that can help in the choice of an index. They are

cate:

Then: `"Flat"`

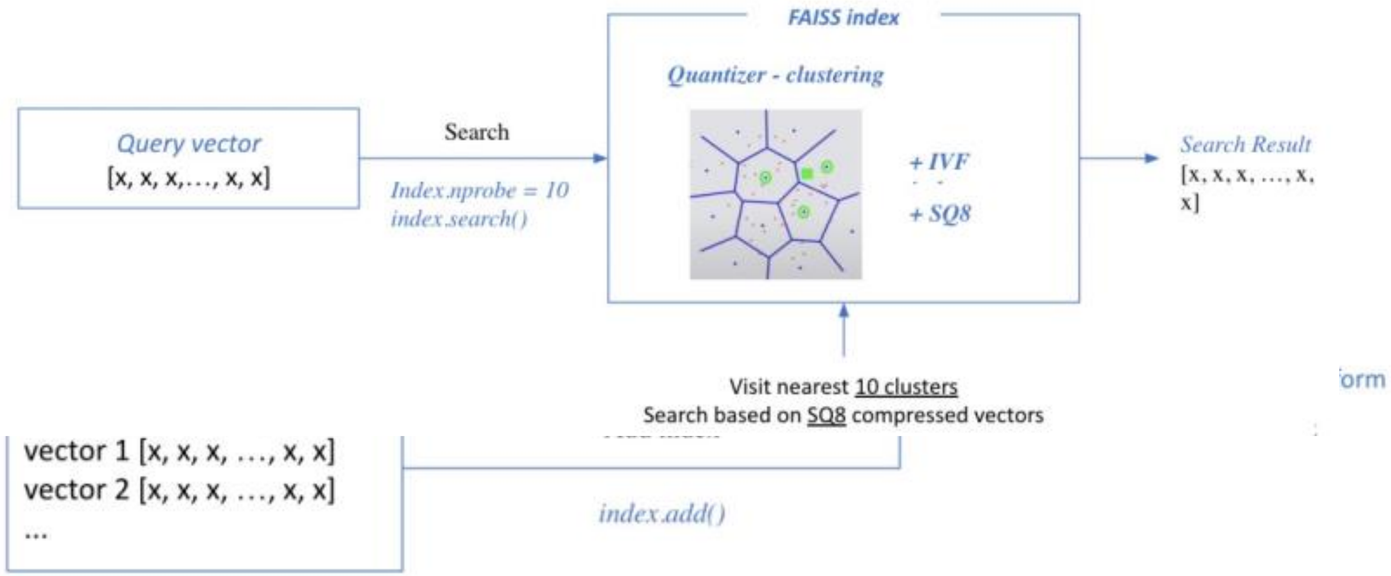
- the `index_factory` string for each of them.
- if there are parameters, we indicate them as the corresponding `ParameterSpace` argument.

InVerted File index

dictionary	posting						
apple	1	2	5	27	80	127	.....
boy	1	2	4	11	23	92	.....
cat	1	2	4	5	6	12	.....
.....							
eagle	1	0	1	0	1	1	
.....							

→ 0이 메모리를 차지하지 않아 효율적

그림으로 보는 FAISS



FAISS + KPF-SBERT(1)

```
query = '문학비평의 기본적인 이론'
dist1, id1 = search(query, index=index)

dist2, id2 = search(query, index=index2)

print(f'문장 분리 후 임베딩 결과 \ndist:{dist1[0]}, \nid:{id1}\n\n')
print(f'문장 분리하지 않고 임베딩 결과 \ndist:{dist2[0]}, \nid:{id2}')

print('\n문장 분리 후 결과')
for cnt, id in enumerate(id1):
  print(f'{cnt+1}. {lecDescSentenceList[lecIDList.index(str(id))]}')

print('\n\n문장 분리하지 않은 결과')
for cnt, id in enumerate(id2):
  print(f'{cnt+1}. {lecDescSentenceList[lecIDList.index(str(id))]}')
```

FAISS + KPF-SBERT(2)

```
문장 분리 후 임베딩 결과
dist:[28.742184 26.578941 24.796602 23.00883  22.779123],
id:[2150387001, 2150551501, 2150501501, 2150039801, 2150501601]


문장 분리하지 않고 임베딩 결과
dist:[21.127104 19.709488 17.87293  16.896643 16.37863 ],
id:[2150551501, 2150387001, 2150265801, 2150393701, 2150307801]

문장 분리 후 결과
1. 문학의 개념, 장르별 특성, 문예사조, 문학비평방법 등 문학의 기초이론을 다룬다.문학뿐 아니라 영화, TV드라마, 연극 등 다양한 문화예술텍스트를 감상하고 분석한다.
2. 이 과목은 간명하게 문학비평의 기본적인 이론과 그 구체적인 적용방법을 학습하기 위해 개설되었다. 오늘날 대표적인 여러 종류의 비평 이론을 학습하는 것은 물론이고,
3. 1. 문학비평 위주로 진행한다.2. 오늘날 비평의 의미, 역할의 필요, 중요성을 확인한다.3. 비평의 역사와 역할을 확인하고 오늘날 비평의 다양한 실제와 현장을 검토한
4. 이번 수업은 수동적 학습자로서가 아닌 능동적으로 창작할 수 있는 방법을 익히고 실천해 보는 역동적인 방식의 수업입니다. Engaged Learning ( 강의 중심 수업을
5. 'Engaged Learning' 수업이란 학습자가 수업에서 학습한 내용을 강의실 밖의 다양한 문제 상황에 적용할 수 있도록 “문제 정의하기”, “아이디어 도출”, “해결 방안


문장 분리하지 않은 결과
1. 이 과목은 간명하게 문학비평의 기본적인 이론과 그 구체적인 적용방법을 학습하기 위해 개설되었다. 오늘날 대표적인 여러 종류의 비평 이론을 학습하는 것은 물론이고,
2. 문학의 개념, 장르별 특성, 문예사조, 문학비평방법 등 문학의 기초이론을 다룬다.문학뿐 아니라 영화, TV드라마, 연극 등 다양한 문화예술텍스트를 감상하고 분석한다.
3. 독일 질풍노도 문학운동의 선구자인 하만과 헤르더의 생애와 사상을 탐색한다.
4. 단편 소설에 중점을 둔 창작론 탐구와 다양한 작품을 읽고 합평하는 창작실습 과정
5. 논리는 모든 생각, 모든 학문의 기본이다.  본 과목은 철학과에서 개설한 교양과목으로서 논리학의 기초를 학습하여 논리적으로 생각하는 방법을 훈련하고 논리적인 언어
```

FAISS + KPF-SBERT(3)

```
query = '예술과의 융합적 지점'
```

```
문장 분리 후 임베딩 결과
dist:[27.735529 21.464443 21.464443 19.515085 19.507273],
id:[2150537501, 2150542602, 2150542601, 2150182301, 2150031701]


문장 분리하지 않고 임베딩 결과
dist:[16.241972 15.387587 14.269231 13.645694 13.034503],
id:[2150223601, 2150691801, 2150032601, 2150569301, 2150307801]

문장 분리 후 결과
1. 1) 18세기 이후 프랑스 문학 사조 및 역사를 짚어봅니다.2) 세기별로 프랑스의 주요 작가와 작품세계를 이야기합니다.3) 문학사와 문학, 예술과의 융합적 지점들을 찾아봅니다.4) '
2. <현대문학사>는 역사와 문학이라는 두 가지 문제가 결합된 내용을 가지고 현대문학의 흐름이 어떻게 전개되었나를 검토하는 과목이라 할 수 있다. 현대문학은 100년의 역사적 진통 속
3. <현대문학사>는 역사와 문학이라는 두 가지 문제가 결합된 내용을 가지고 현대문학의 흐름이 어떻게 전개되었나를 검토하는 과목이라 할 수 있다. 현대문학은 100년의 역사적 진통 속
4. 이 수업은 우리 시대의 시급한 문제적 사안 중의 하나인 생태, 환경 위기에 관한 철학적, 예술적 고찰을 시도한다. 생태철학과 생태예술의 논의에 있어 전제되어야 할 주요한 논제는
5. 디지털 시대의 촬영자를 위한 궁극의 기초 II작품의 분석을 통해서 촬영상의 예술적.기술적 문제점 해결에 초점을 둔다. 동시에 앞으로 촬영할 다양한 스크립트에 대한 비주얼 컨셉을


문장 분리하지 않은 결과
1. 이번 강의는 현대 윤리에서 가장 중요하게 거론되는 정의(justice) 문제를 비판적으로 다루는 데 목표를 둔다. 근대 윤리학은 정치와 윤리를 분리시키려 했다. 하지만 현대 윤리에서
2. 다양하고 흥미로운 사운드를 활용한 작품창작
3. 영화나 OTT와 같은 콘텐츠 제작은 반드시 협동 작업이 필요하며 그에 따른 예술적 고민과 기술적 문제에 대한 해결 방법들이 동반 됩니다.독단적 실행 또는 결과론적인 방법에만 그치
4. 이 과목은 방언의 개념 및 그 연구 방법론에 대한 지식을 습득하고, 이를 바탕으로 방언과 문화의 상호 관계에 대해 연구하는 데 목적이 있다.
5. 논리는 모든 생각, 모든 학문의 기본이다.  본 과목은 철학과에서 개설한 교양과목으로서 논리학의 기초를 학습하여 논리적으로 생각하는 방법을 훈련하고 논리적인 언어 능력과 추론 능
```

```
df_result = df[df['강의 설명'].str.contains('예술과의 융합적 지점')]['강의 설명'].tolist()
print(df_result)
```

```
["1) 18세기 이후 프랑스 문학 사조 및 역사를 짚어봅니다.2) 세기별로 프랑스의 주요 작가와 작품세계를 이야기합니다.3) 문학사와 문학, 예술과의 융합적 지점들을 찾아봅니다.
```

L2 vs Inner Product (L2)

```
query = '예술과의 융합적 지점'
```

```
======================L2======================
[[29.299747 38.047016 38.92013  38.92013  40.270866 42.96111  43.249775
  43.683834 43.84935  44.5434  ]]
[[2150537501 2150032601 2150542602 2150542601 2150215101 2150223601
  2150501101 2150031701 2150052101 2150016501]]
1. 1) 18세기 이후 프랑스 문학 사조 및 역사를 짚어봅니다.2) 세기별로 프랑스의 주요 작가와 작품세계를
2. 영화나 OTT와 같은 콘텐츠 제작은 반드시 협동 작업이 필요하며 그에 따른 예술적 고민과 기술적 문제에
3. <현대문학사>는 역사와 문학이라는 두 가지 문제가 결합된 내용을 가지고 현대문학의 흐름이 어떻게 전개
4. <현대문학사>는 역사와 문학이라는 두 가지 문제가 결합된 내용을 가지고 현대문학의 흐름이 어떻게 전개
5. 한국미술사는 과거 인간들이 남긴 미술작품을 통해 그 역사적인 전개를 이해하는 학문의 하나로 어떤 작품
6. 이번 강의는 현대 윤리에서 가장 중요하게 거론되는 정의(justice) 문제를 비판적으로 다루는 데 목표
7. 본 강의는 스토리텔링 이론의 기초적인 학습을 선행하는 과목입니다. 메타버스시대의 현실문학과 가상문학
8. 디지털 시대의 촬영자를 위한 궁극의 기초 II작품의 분석을 통해서 촬영상의 예술적.기술적 문제점 해결
9. 고대 동아시아에서는 불교나 유교와 같은 종교 및 사상체계를 배경으로 조각과 회화, 공예의 다양한 작품
10. 이 수업에서는 20세기 신학의 고전이라고 할 수 있는 가톨릭 신학자 한스 큉과 개신교 신학자 볼프하르
```

L2 vs Inner Product (Inner Product)

```
query = '예술과의 융합적 지점'
```

```
==================IP==================
[[27.735529 21.464443 21.464443 19.515085 19.507273 19.496973 19.015625
  17.85282  16.920736 16.48478 ]]
[[2150537501 2150542601 2150542602 2150182301 2150031701 2150223601
  2150673301 2150151901 2150395001 2150359001]]
1. 1) 18세기 이후 프랑스 문학 사조 및 역사를 짚어봅니다.2) 세기별로 프랑스의 주요 작가와 작품세계를 이야기합니다.3) 문학사와 문
2. <현대문학사>는 역사와 문학이라는 두 가지 문제가 결합된 내용을 가지고 현대문학의 흐름이 어떻게 전개되었나를 검토하는 과목이라 할
3. <현대문학사>는 역사와 문학이라는 두 가지 문제가 결합된 내용을 가지고 현대문학의 흐름이 어떻게 전개되었나를 검토하는 과목이라 할
4. 이 수업은 우리 시대의 시급한 문제적 사안 중의 하나인 생태, 환경 위기에 관한 철학적, 예술적 고찰을 시도한다. 생태철학과 생태예술
5. 디지털 시대의 촬영자를 위한 궁극의 기초 II작품의 분석을 통해서 촬영상의 예술적.기술적 문제점 해결에 초점을 둔다. 동시에 앞으로
6. 이번 강의는 현대 윤리에서 가장 중요하게 거론되는 정의(justice) 문제를 비판적으로 다루는 데 목표를 둔다. 근대 윤리학은 정치와
7. 4차산업혁명시대의 융·복합기술은 완성도 높은 서사(이야기)가 구축되어주지 않으면, 단지 기술에 불과할 뿐입니다.지금은 한류콘텐츠의
8. 전면 대면수업 합니다.그러나 개강 후 학교의 상황판단에 따라서 수업 형태가 변경될 수도 있습니다.20세기 이후 영미권에서 전개된 주요
9. *Engaged Learning  수업으로 소설 창작을 위한 요소들과 관련한 문제 정의-아이디어 도출-해결방안 적용의 3단계 과정의 그룹 토
10. 중세는 고대와 근대 사이에 낀 암흑시대로 평가받는다. 이 시기의 철학은 이성보다 믿음을 중시하고 그 자체가 아니라 종교의 시녀로서
```

감사합니다.