한국어 임베딩 (이기창 저. 2019)

Chapter 1. Introduction

Spring Semester Capstone Study

TEAM Kai.Lib

발표자 : 배세영

2020.04.08 (WED)

이기창님 DevFest 2019 Seoul PPT 참고(링크)

1. 임베딩이란

사람이 쓰는 자연어를 기계가 이해할 수 있는 숫자의 나열인
벡터(Vector) 로 바꾼 **결과, 또는 그 일련의 과정**

단어나 문장 각각을 벡터로 변환하여
벡터 공간(Vector Space) 으로
끼워 넣는다(embed) 는 의미

1. 임베딩이란

가장 간단한 형태의 임베딩?
단어의 빈도를 그대로 벡터로 사용하는 것!

단어-문서 행렬(Term-Document Matrix)

구분	IEEE	InterSpeech	CS	EESS
Kai.Lib	1	2	1	0
근본	1	1	1	0
캡스톤	0	1	2	3

InterSpeech : [2,1,1] / EESS : [0,2,3] : 유사도 Low

근본 : [1,1,1,0] / Kai.Lib : [1,2,1,0] : 유사도 High

2. 임베딩의 역할

- **단어-문장 간 관련도 계산**
  - 이전의 단어-문서 행렬은 가장 단순한 형태의 임베딩
  - 현업에서는 이보다 복잡한 형태의 임베딩 사용 (Word2Vec, Google. 2013.)

‘희망’ 이라는 단어의 Word2Vec 임베딩

$$[-0.00209\ -0.03918\ 0.02419\ \ldots\ 0.01715\ -0.04975\ 0.09300]$$

이렇게 단어를 수치화하면

단어 벡터들 사이의 유사도를 측정하는 것이 가능해진다!

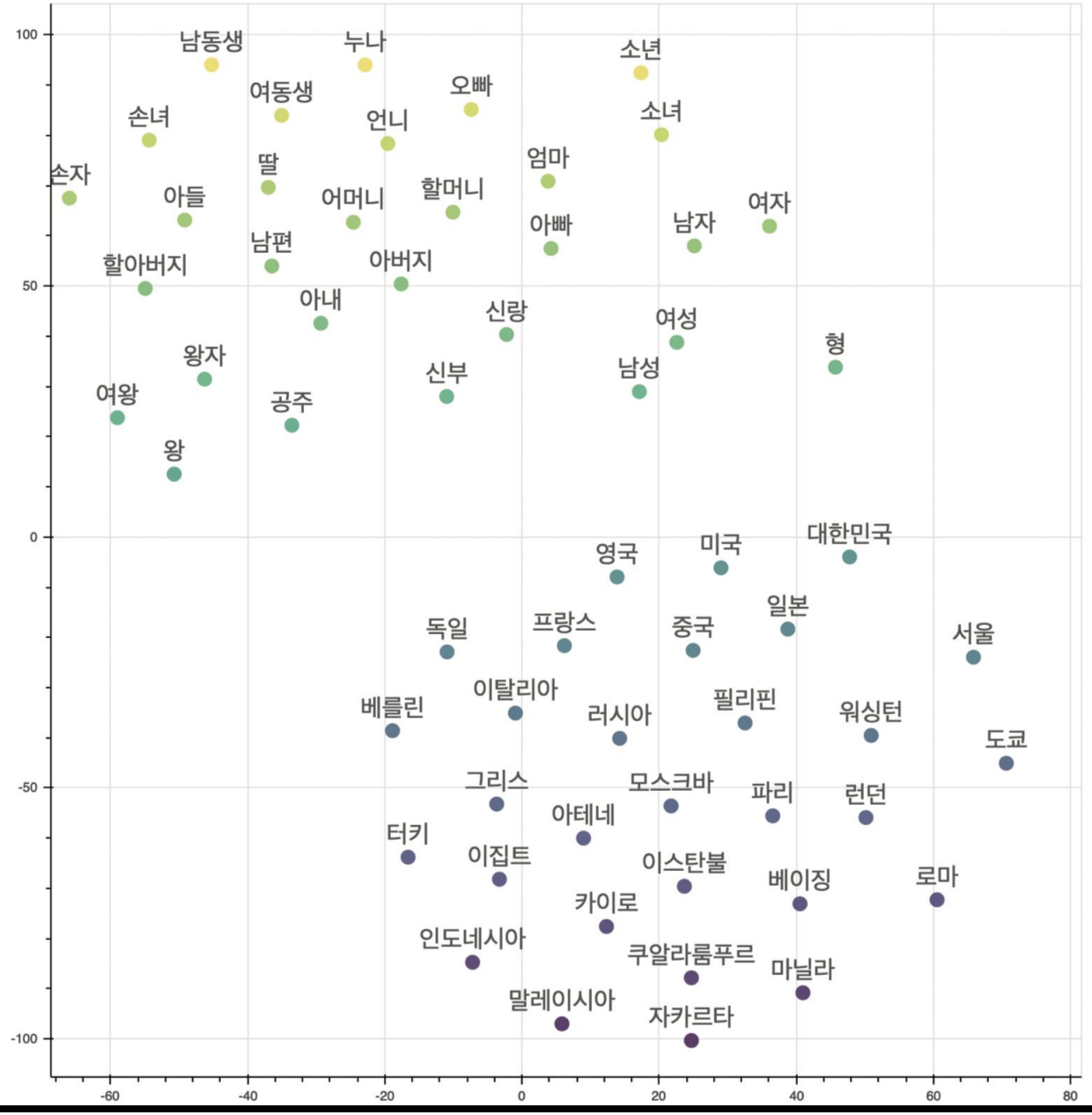
2. 임베딩의 역할

- **단어-문장 간 관련도 계산**
  - 이전의 단어-문서 행렬은 가장 단순한 형태의 임베딩
  - 현업에서는 이보다 복잡한 형태의 임베딩 사용 (Word2Vec, Google. 2013.)

Word2Vec 임베딩 단어들의 Cosine Similarity

희망	절망	학교	학생	가족	자동차
소망	체념	초등	대학생	아이	승용차
행복	고뇌	중학교	대학원생	부모	상용차
희망찬	절망감	고등학교	고학생	편부모	트럭
꿈	상실감	야학교	교직원	고달픈	대형트럭
열망	번민	중학	학부모	사랑	모터사이클

2. 임베딩의 역할

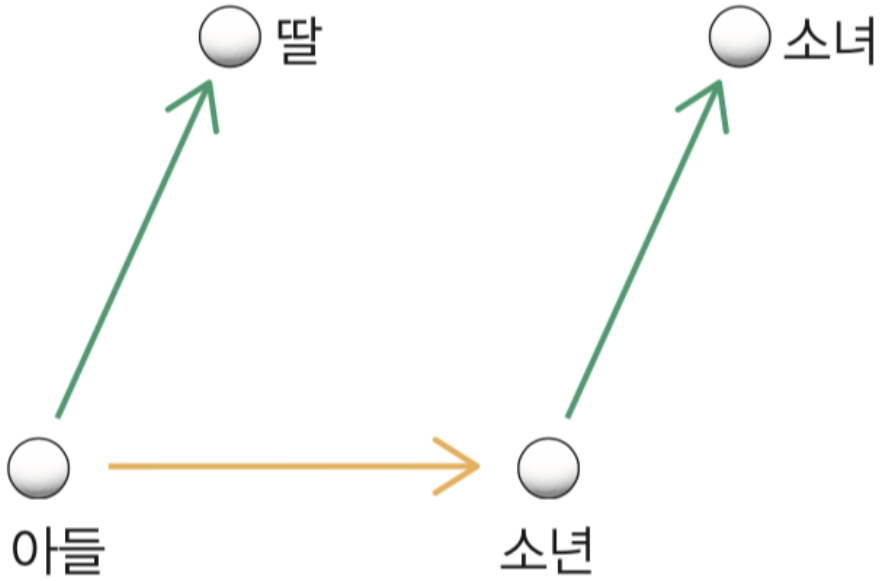

남동생
누나
소년
여동생
오빠
손녀
언니
소녀
엄마
딸
손자
아들
어머니
할머니
여자
아빠
남자
남편
할아버지
아버지
아내
신랑
여성
왕자
형
남성
신부
여왕
공주
왕
대한민국
영국
미국
일본
프랑스
독일
중국
서울
이탈리아
베를린
필리핀
러시아
워싱턴
도쿄
그리스
모스크바
파리
런던
아테네
터키
이집트
이스탄불
베이징
로마
카이로
인도네시아
쿠알라룸푸르
마닐라
말레이시아
자카르타
100
50
0
-50
-100
-60
-40
-20
0
20
40
60
80

2. 임베딩의 역할

- **의미적, 문법적 정보 함축**
  - 임베딩은 벡터이므로 사칙연산이 가능
  - 벡터들 간의 덧셈, 뺄셈을 통해 단어들 사이의 의미적, 문법적 관계를 도출해낼 수 있게 됨

Word2Vec 임베딩 단어들의 단어 유추 평가word analogy test

[아들 – 딸 + 소년 = 소녀]



2. 임베딩의 역할

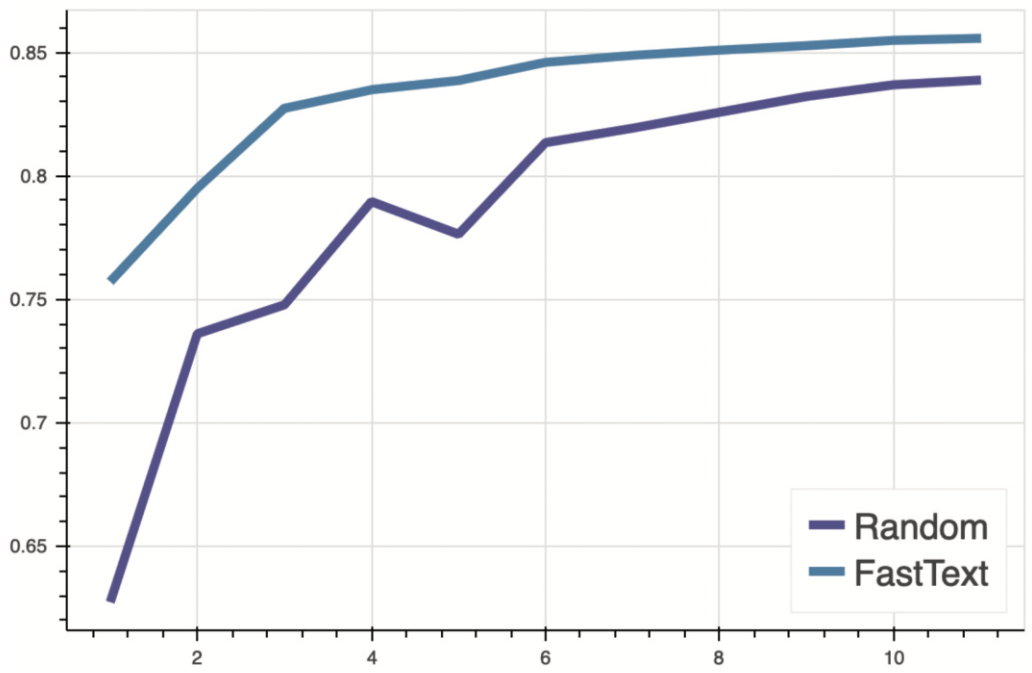
- **의미적, 문법적 정보 함축**
  - 임베딩은 벡터이므로 사칙연산이 가능
  - 벡터들 간의 덧셈, 뺄셈을 통해 단어들 사이의 의미적, 문법적 관계를 도출해낼 수 있게 됨

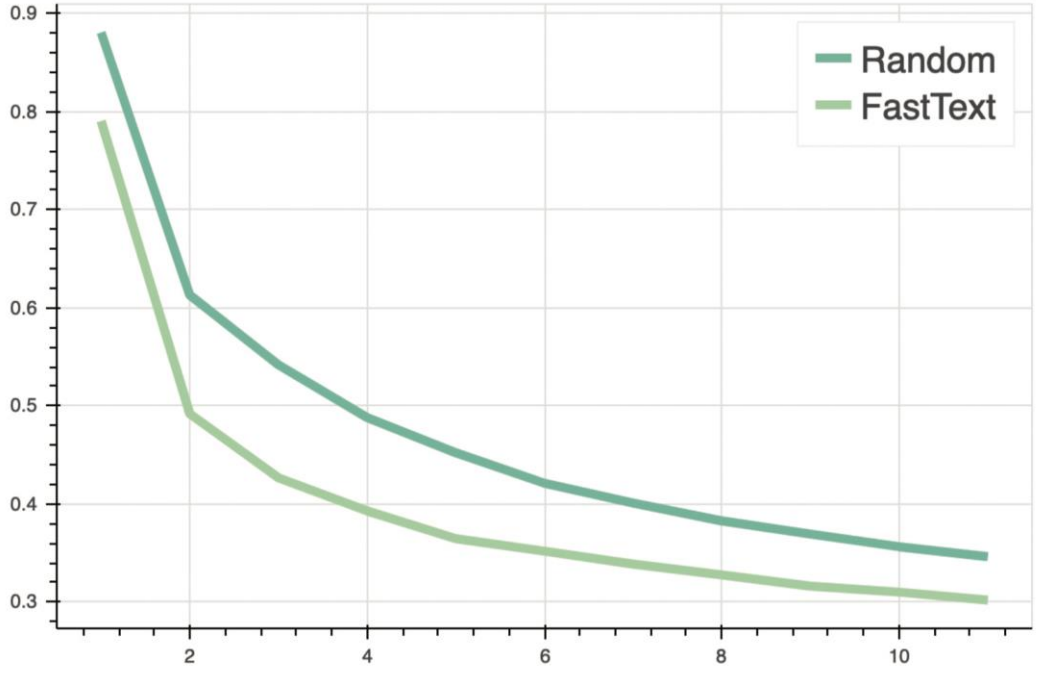
단어1	단어2	단어3	결과
아들	딸	소년	소녀
아들	딸	아빠	엄마
아들	딸	남성	여성
남동생	여동생	소년	소녀
남동생	여동생	아빠	엄마
남동생	여동생	남성	여성
신랑	신부	왕	여왕
신랑	신부	손자	손녀
신랑	신부	아빠	엄마

2. 임베딩의 역할

- 전이 학습(Transfer Learning)
  - 임베딩을 다른 딥러닝 모델의 입력값으로 사용하는 기법
  - 임베딩에 의미적, 문법적 정보 등이 포함되어 있으므로 성능과 학습 속도가 향상됨

FastText 임베딩 전이학습 시 모델 성능과 학습 손실loss





3. 임베딩 기법의 역사와 종류

- 통계 기반 -> 신경망(Neural Network) 기반
- 단어 수준 -> 문장 수준
- 룰 -> End-to-End -> Fine Tuning
- 임베딩의 종류와 성능
  - 행렬 분해factorization
  - 예측 기반
  - 토픽 기반
  - 임베딩 성능 평가

3. 임베딩 기법의 역사와 종류

- **통계 기반에서 신경망Neural Network 기반으로**
  - 전통적 임베딩은 주로 말뭉치의 통계량을 직접적으로 활용
  - 잠재 의미 분석(Latent Semantic Analysis)

잠재 의미 분석?

크기가 큰 행렬에 특이값 분해Singular Value Decomposition 등의 기법을 적용
차원을 축소하는 방법

이때 대상이 되는 '크기가 큰' 행렬은 **단어-문서 행렬**을 비롯하여
TF-IDF Term Frequency-Inverse Document Frequency,
단어-문맥 행렬 Word-Context Matrix,
점별 상호 정보량 행렬 Pointwise Mutual Information Matrix 등 여러 가지가 될 수 있다.

3. 임베딩 기법의 역사와 종류

- **통계 기반에서 신경망Neural Network 기반으로**
  - 전통적 임베딩은 주로 말뭉치의 통계량을 직접적으로 활용
  - 잠재 의미 분석(Latent Semantic Analysis)

	문서 1	문서 2	문서 3	문서 4
단어 1	2	0	0	0
단어 2	0	1	0	1
단어 3	0	0	0	3
단어 4	1	0	1	2

	주제 1	주제 2
단어 1	0.42	1.92
단어 2	1.03	-0.29
단어 3	2.88	-0.69
단어 4	2.29	0.64

	문서 1	문서 2	문서 3	문서 4
주제 1	0.81	0.27	0.59	3.70
주제 2	2.08	-0.13	0.30	-0.49

3. 임베딩 기법의 역사와 종류

- **단어 수준에서 문장 수준으로**
  - 2017년 이전의 임베딩 기법들은 주로 단어 수준 (NPLM, Word2Vec, Glove, FastTex, Swivel)
  - 단어 수준 임베딩은 동음이의어를 분간하기 어려움
    :: 단어의 형태가 같다면 동일한 단어로 판단하고, 모든 문맥 정보를 해당 단어 벡터에 투영하기 때문

  - 2018년 초 ELMo(Embeddings from Language Models) 발표
  - BERT(Bidirectional Encoder Representations from Transformer)
  - GPT(Generative Pre-Training)

  - 문장 수준 임베딩은 개별 단어가 아닌 단어 시퀀스 전체의 문맥적 의미를 함축
  - 단어 임베딩 기법보다 전이학습 효과가 높음

3. 임베딩 기법의 역사와 종류

- 룰 기반 -> End-to-End -> Pre-Train/Fine Tuning
  - 언어학적 지식을 사용하는 룰 기반 임베딩이 1990년대의 초기 방식

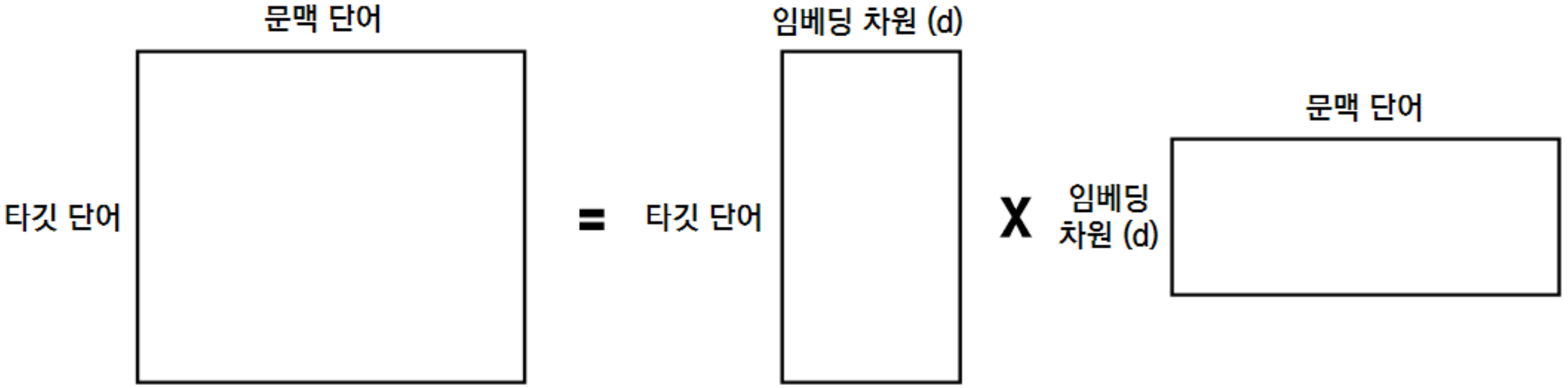
  - 2000년대 중반 이후 자연어 처리 분야에서도 딥러닝 모델이 주목받기 시작하며 입출력 사이의 관계에 사람이 전혀 개입하지 않고 모델 스스로 이해하도록 유도하는 End-to-End 모델 등장
  - 기계번역에서 널리 쓰였던 Seq2Seq 모델이 대표적

  - 2018년 ELMo 모델이 제안된 이후 자연어 처리 모델은 End-to-End에서 Pre-Train/Fine Tuning 방식으로 발전
  - 먼저 대규모 말뭉치로 임베딩을 만들고(Pre-Train), 활용을 원하는 구체적 문제에 맞는 소규모 데이터로 임베딩을 포함한 모델 전체를 업데이트(Fine Tuning, Transfer Learning)

3. 임베딩 기법의 역사와 종류

- **임베딩 기법의 종류 : 행렬 분해factorization 기반**
  - 말뭉치 정보가 들어 있는 원본 행렬을 두 개 이상의 작은 행렬로 분해
  - 분해한 행렬을 서로 더하거나(sum) 이어 붙여(concatenate) 임베딩으로 활용

  - Glove, Swivel 등이 이에 해당



3. 임베딩 기법의 역사와 종류

- **임베딩 기법의 종류 : 예측 기반**
  - 어떤 단어 주변에 특정 단어가 나타날지 예측
  - 이전 단어들이 주어졌을 때 다음 단어를 예측
  - 문장 내 일부 단어를 지우고 해당 단어가 무엇일지 예측

  - 위와 같은 예측 과정에서 학습하는 임베딩 기법으로, 신경망 기반임

  - Word2Vec, FastText, BERT, ELMo, GPT 등이 이에 해당

3. 임베딩 기법의 역사와 종류

- **임베딩 기법의 종류 : 토픽Topic 기반**
  - 주어진 문제에 잠재된 주제(Latent Topic)를 추론(Interference)하는 방식으로 임베딩을 수행
  - 잠재 디리클레 할당(LDA, Latent Dirichlet Allocation)이 이에 해당
  - LDA는 학습이 완료되면 각 문서가 어떤 주제 분포를 갖는지 확률 벡터 형태로 반환
  - 주제 파악 Task에서 사용하는 특수한 임베딩 기법이라고 판단, 본 캡스톤 주제와는 큰 관련 없을 것

4. 주요 용어

- 주로 다루는 데이터는 text 형태의 자연어

- 말뭉치 : 특정한 목적을 가지고 수집한 표본(sample)
  :: 표본이 아무리 커도 무한한 자연어의 일부일 뿐이므로

- 컬렉션collection : 말뭉치에 속한 각각의 집합
  :: 한국어 위키백과, 네이버 영화 리뷰, ...

- 문장sentence : 데이터의 기본 단위
  :: . ! ? 등으로 구분된 문자열을 문장으로 취급

- 문서document : 생각이나 감정, 정보를 공유하는 문장의 집합
  :: 단락paragragh와 굳이 구분하지 않음,
  :: 개행문자 \n으로 구분된 문자열을 문서로 취급

4. 주요 용어

- 토큰token : 문장을 이루는 가장 작은 단위
  :: 단어word, 형태소morpheme, 서브워드subword라고도 함

- 토크나이즈tokenize : 문장을 토큰 시퀀스로 분석하는 과정

- 어휘 집합vocabulary : 말뭉치에 있는 모든 문서를 문장으로 나누고 여기에 토크나이즈를 실시한 후 중복을 제거한 토큰들의 집합

- 미등록 단어unknown word : 어휘 집합에 없는 토큰