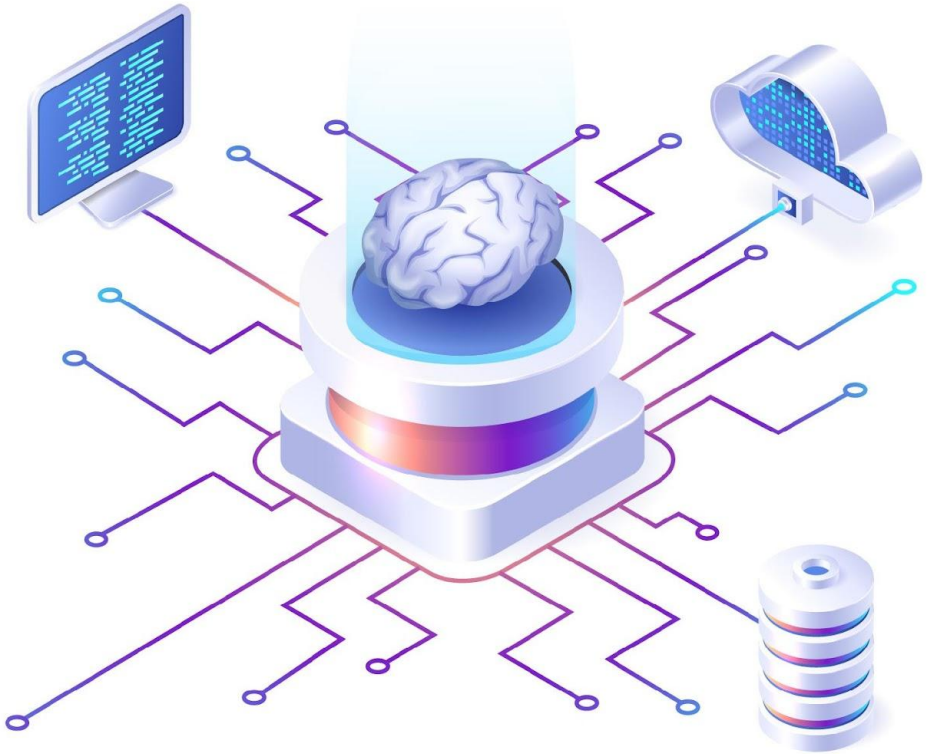
Insight campus SeSAC

표현(Representation)

실무형 인공지능 자연어 처리



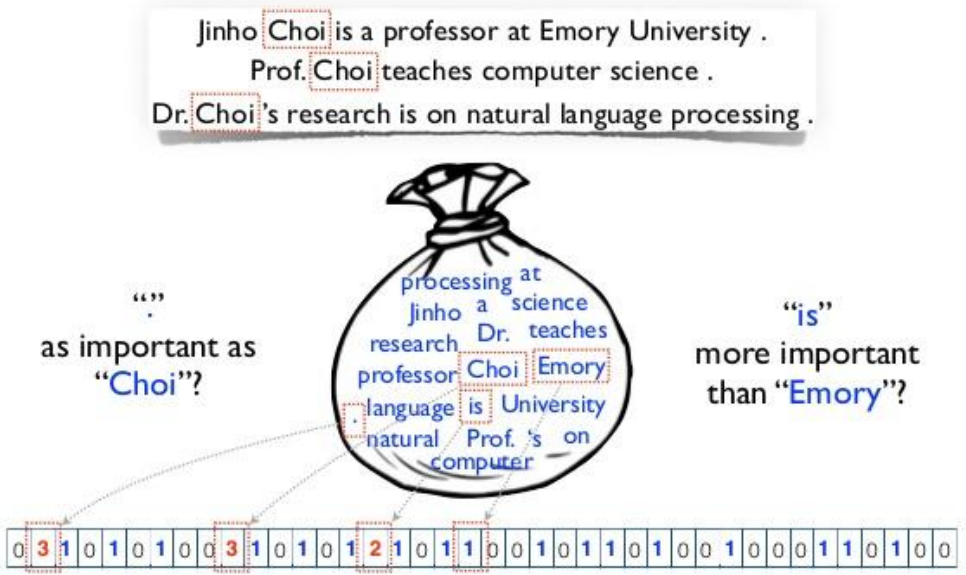
1

BoW (단어 주머니)

Bag of Words

BoW (Bag of Words)

BoW(Bag of Words) : 문서 내 단어 출현 순서는 무시. 빈도수만 기반으로 문서를 표현하는 방법



Insight campus SeSAC

BoW 생성 방법

문서1: 오늘 동물원에서 코끼리를 봤어
문서2:오늘 동물원에서 원숭이에게 사과를 줬어

Step1. 각 토큰에 고유 인덱스 부여

오늘	0
동물원에서	1
코끼리를	2
봤어	3
원숭이에게	4
사과를	5
줬어	6

Step2. 각 인덱스 위치에 토큰 등장 횟수를 기록

	오늘	동물원에서	코끼리를	봤어	원숭이에게	사과를	줬어
문서1	1	1	1	1	0	0	0

	오늘	동물원에서	코끼리를	봤어	원숭이에게	사과를	줬어
문서2	1	1	0	0	1	1	1

한계

- 단어의 **순서를 고려 하지 않음**
- **BoW 는 Spare** 함. 벡터 공간의 낭비, 연산 비효율성 초래
- **단어 빈도수가 중요도를 바로 의미 하지 않음.** 단어가 자주 등장한다고 중요한 단어는 아님.
- 전처리가 매우 중요함. 같은 의미의 다른 단어 표현이 있을 경우 다른것으로 인식될 수 있음.
  (뉴스와 같이 정제된 어휘를 사용하는 매체는 좋으나, 소셜에서는 활용하기 어려움)

2

TDM (단어-문서 행렬)

Term-Document Matrix

TDM (Term-Document Matrix)

BoW(Bag of Words) 중 하나
문서에 등장하는 각 단어 빈도를 행렬로 표현한 것

문서1: 동물원 코끼리
문서2: 동물원 원숭이 바나나
문서3: 엄마 코끼리 아기 코끼리
문서4: 원숭이 바나나 코끼리 바나나

	동물원	코끼리	원숭이	바나나	엄마	아기
문서1	1	1	0	0	0	0
문서2	1	0	1	1	0	0
문서3	0	2	0	0	1	1
문서4	0	1	1	2	0	0

Insight campus SeSAC

TDM vs DTM

	Tweet 1	Tweet 2	Tweet 3	…	Tweet N
Term 1	0	0	0	0	0
Term 2	1	1	0	0	0
Term 3	1	0	0	0	0
…	0	0	3	1	1
Term M	0	0	0	1	0

Term Document Matrix (TDM)

	Term 1	Term 2	Term 3	…	Term M
Tweet 1	0	1	1	0	0
Tweet 2	0	1	0	0	0
Tweet 3	0	0	0	3	0
…	0	0	0	1	1
Tweet N	0	0	0	1	0

Document Term Matrix (DTM)

TDM (Term-Document Matrix) 의 한계

- 단어의 **순서를 고려 하지 않음**
- **TDM 는 Spare** 함. 벡터 공간의 낭비, 연산 비효율성 초래
- **단어 빈도수가 중요도를 바로 의미 하지 않음.** the와 같은 단어는 빈번하게 등장하고 TDM에서 중요한 단어로 판단 될 수 있음
  => 이를 보완 하기 위하여 TF - IDF 를 사용

감사합니다.

Insight campus SeSAC

daniel@fins.ai