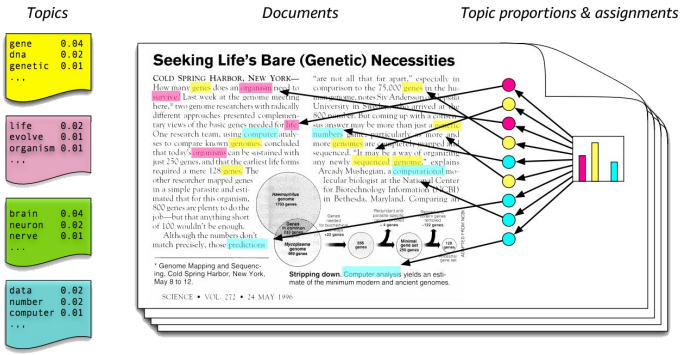
Insight campus SeSAC

토픽 모델링 (Topic Modeling)

실무형 인공지능 자연어처리

토픽모델링 (1)

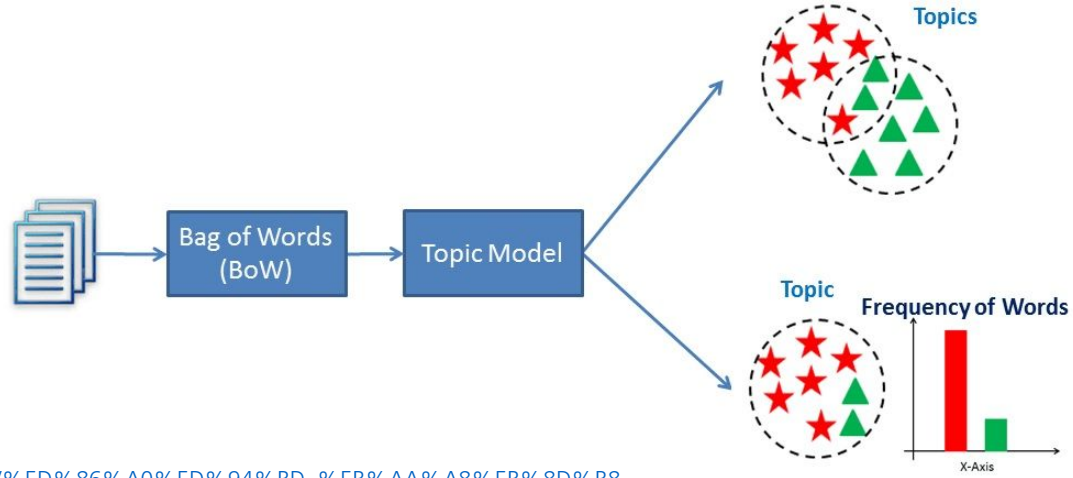
기계 학습 및 자연언어 처리 분야에서 토픽 모델(Topic model)이란 **문서 집합의 추상적인 "주제"를 발견하기 위한 통계적 모델 중 하나로, 텍스트 본문의 숨겨진 의미구조를 발견하기 위해 사용**되는 텍스트 마이닝 기법 중 하나이다. 특정 주제에 관한 문헌에서는 그 주제에 관한 단어가 다른 단어들에 비해 더 자주 등장할 것이다. 예를 들어 개에 대한 문서에서는 "개"와 "뼈다귀"라는 단어가 더 자주 등장하는 반면, 고양이에 대한 문서에서는 "고양이"와 "야옹"이 더 자주 등장할 것이고, "그", "~이다"와 같은 단어는 양쪽 모두에서 자주 등장할 것이다. 이렇게 함께 자주 등장하는 단어들은 대게 유사한 의미를 지니게 되는데 이를 잠재적인 "주제"로 정의할 수 있다. 즉, "개"와 "뼈다귀"를 하나의 주제로 묶고, "고양이"와 "야옹"을 또 다른 주제로 묶는 모형을 구상할 수 있는데 바로 이것이 토픽 모델의 개략적인 개념이다. 실제로 문헌 내에 어떤 주제가 들어있고, 주제 간의 비중이 어떤지는 문헌 집합 내의 단어 통계를 수학적으로 분석함으로써 알아낼 수 있다.



https://ko.wikipedia.org/wiki/%ED%86%A0%ED%94%BD_%EB%AA%A8%EB%8D%B8

토픽모델링 (2)

텍스트에 숨겨져 있는 주제들을 찾아내기 위한 통계추론에 기반한 분석 기법을 말한다. 개별 문서는 다수의 주제, 혹은 토픽을 다룰 수 있다는 점을 전제로 수집된 텍스트를 토픽의 확률적 혼합체로 간주하고, 각 **토픽을 추출된 키워드들의 분포로 나타냄으로써 텍스트 내의 구조를 파악**한다.
토픽모델링을 이해하려면 통계에 대한 사전지식이 필요하다.



https://ko.wikipedia.org/wiki/%ED%86%A0%ED%94%BD_%EB%AA%A8%EB%8D%B8

토픽모델링의 활용

- 사회 문제를 다루고 있는 대용량 **뉴스기사로부터** 토픽분석을 적용하여 사회적 이슈에 관한 키워드를 도출하는 시스템을 제안

- 트위터 데이터를 대상으로 **SNS상에서의 주요 이슈를 추출**하는 트위터 이슈 트래킹 시스템을 제안

- 국토교통, 안전, 정보통신기술, 건설과 철강산업 등의 분야에도 토픽 모델링을 적용하여 **미래 핵심 기술과 이슈를 발견하고 트랜드를 분석**하여 경제적, 사회적 부가가치를 창출하고 국가 전략 및 정책 수립 반영하는데에 활용

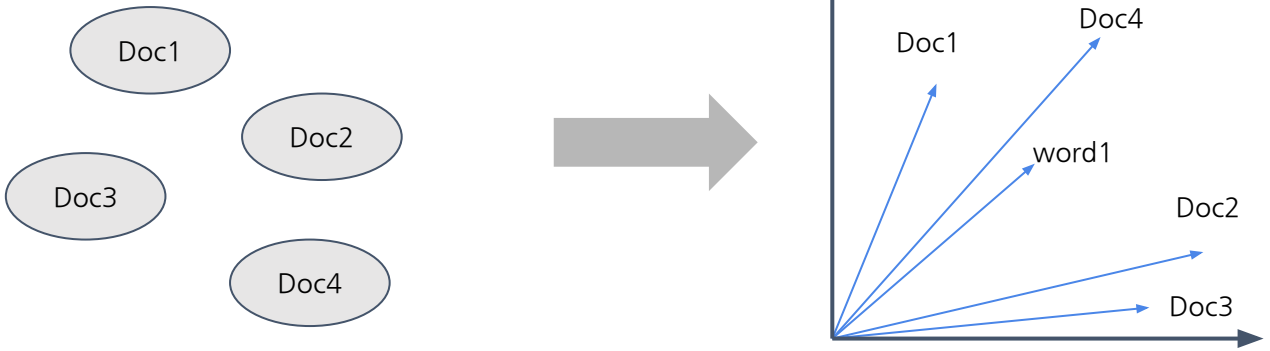
- **토픽을 도출하고 그것을 보면서 사회와 시대를 이해**하고 이를 바탕으로 의사결정을 하고 계획을 세울 수 있는 분석이 토픽분석

1 잠재 의미 분석 (LSA)

Latent Semantic Analysis

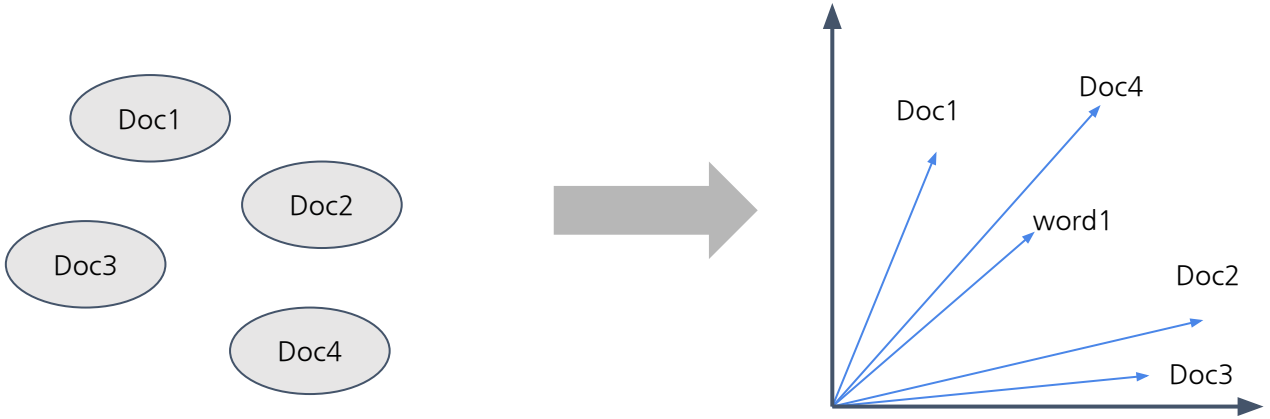
잠재 의미 분석 (LSA, Latent semantic analysis)

- 잠재 의미 분석(LSA, Latent Semantic Analysis)은 토픽모델링 방법 중 하나
- BoW에 기반한 TDM이나 TF-IDF가 빈도로 단어 중요도를 판단하고 있어 의미를 고려하지 못한다는 단점이 있음
- LSA는 동일한 의미를 공유하는 단어들은 같은 텍스트에서 발생(co-occurrence)한다고 가정하는 벡터 기반 방법
- TDM 내에 잠재된(Latent) 의미를 이끌어 내는 방법



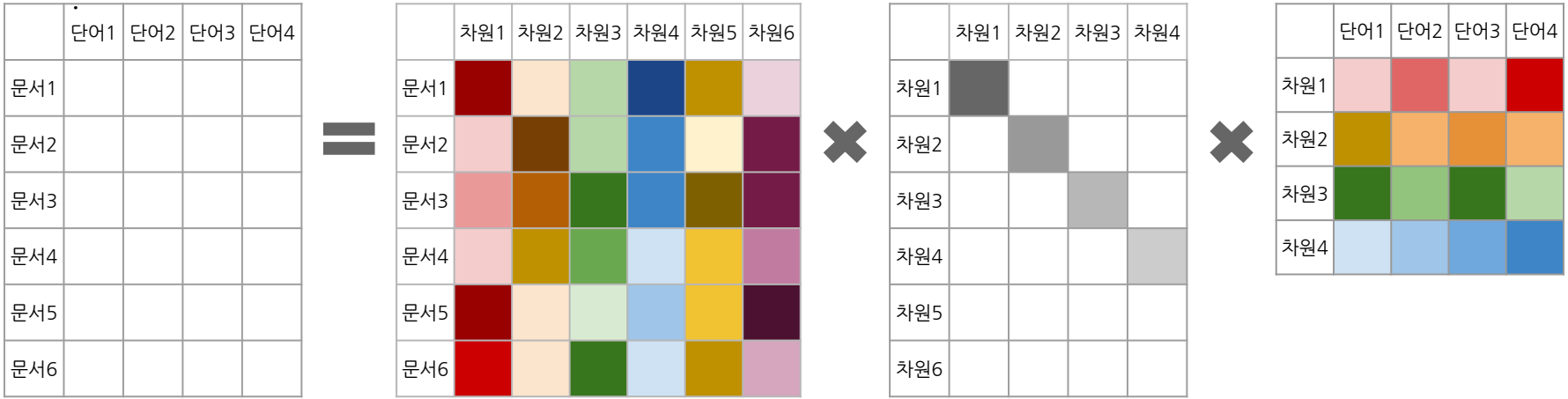
잠재 의미 분석 (LSA, Latent semantic analysis)

- 대량의 텍스트 문서에서 발생하는 단어들 간의 연관관계를 분석함으로써 잠재적인 의미 구조를 도출
- 문서 집합 내에서 연관성, 즉 **동시출현(co-occurrence)빈도가 높은 단어들을 기준**으로 유사한 문서를 추출
  - co-occurrence 정보를 이용한다는 것은 단어의 '형태(morphology)가 아닌 의미(semantic)'를 이용한다는 뜻이다. 예를 들어 '배'라는 단어는 같은 문장에 co-occur 하는 동사가 '타다' 인지 '먹다' 인지에 따라 의미가 갈라지게 된다.



LSA 활용 (1) - 토픽모델링

- 토픽을 추출된 키워드들의 분포로 나타냄으로써 텍스트 내의 구조를 파악



문서-단어행렬
m x n = 6 x 4

좌특이벡터 (문서벡터)
m x m = 6 x 6

특이값
m x n = 6 x 4

우특이벡터 (단어벡터)
n x n = 4 x 4

LSA 활용 (1) - 토픽모델링

- 토픽을 추출된 키워드들의 분포로 나타냄으로써 텍스트 내의 구조를 파악

문서-단어행렬
m x n = 6 x 4

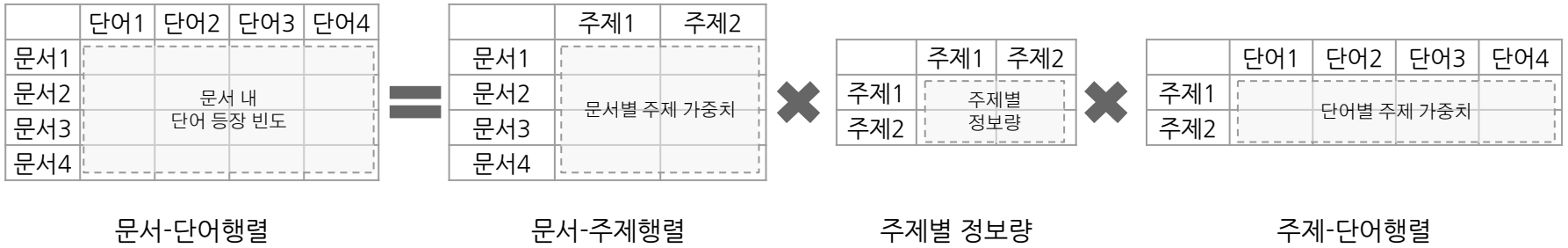
좌특이벡터 (문서벡터)
m x m = 6 x 6

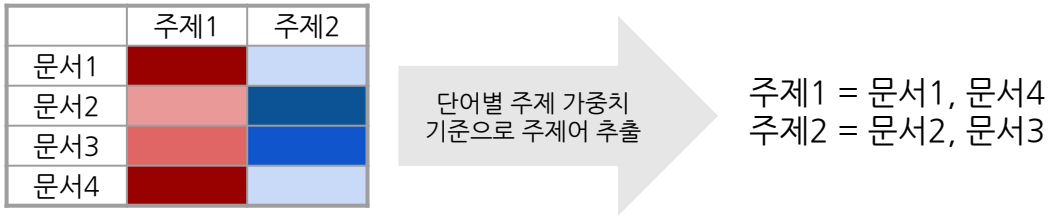
특이값
m x n = 6 x 4

우특이벡터 (단어벡터)
n x n = 4 x 4

LSA 활용 (1) - 토픽모델링

- 토픽을 추출된 키워드들의 분포로 나타냄으로써 텍스트 내의 구조를 파악





LSA 활용 (2) - 밀집 벡터 생성

- TDM (문서-단어 행렬)은 sparse 함
- LSA를 활용하여 의미를 보존하며 밀집벡터(dense vector)를 생성할수 있음

Insight campus SeSAC

LSA 활용 (3) - 단어 간 유사도 측정

- *U* 는 k차원의 단어 벡터를 의미함
- *U* 는 잠재 의미에 기반한 단어 벡터
- 단어 벡터간 코사인 유사도를 측정하여 유사도 측정이 가능

	단어1	단어2	단어3	단어4
문서1	문서 내 단어 등장 빈도			
문서2				
문서3				
문서4				

문서-단어행렬

=

	차원1	차원2
문서1	단어별 주제 가중치	
문서2		
문서3		
문서4		

문서벡터행렬

✖

	차원1	차원2
차원1	특이값	
차원2		

✖

	단어1	단어2	단어3	단어4
차원1	문서별 주제 가중치			
차원2				

단어벡터행렬

	단어1	단어2	단어3	단어4
차원1	단어1 벡터	단어2 벡터	단어3 벡터	단어4 벡터
차원2				

LSA 활용 (4) - 문서 간 유사도 측정

- V^T 는 토픽별 문서의 벡터를 의미함
- V^T 는 잠재 의미에 기반한 단어 벡터
- 문서 벡터간 코사인 유사도를 측정하여 유사도 측정이 가능

잠재 의미 분석 절차


토픽모델링
(Topic modeling)
문서(Document) / 단어(Term)
TDM 생성 (문서-단어 행렬)
특이값 분해 (SVD)
토픽모델링 (n개 토픽)
예. 5개 토픽모델링
벡터 활용
단어간 유사도 분석
단어 벡터(U 벡터)간 유사도 분석
문서간 유사도분석
문서 벡터(V^T 벡터)간 유사도 분석
문서-단어간 유사도분석
문서-단어간 유사도분석

잠재 의미 분석 한계

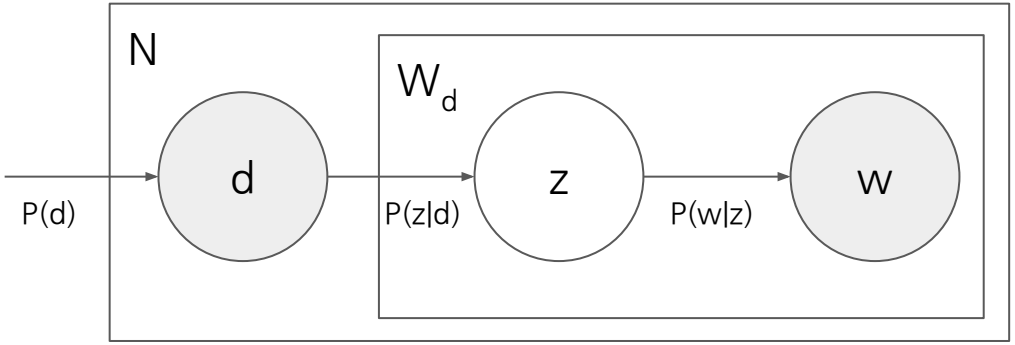
- 잠재 의미 분석은 다의어(polysemy) 문제에 한계가 있음
  - 확률적 잠재 의미 분석(pLSA)는 다의어(polysemy)를 다룰 수 있음
- 잠재 디리클레 할당(LDA)는 LSA의 한계를 극복

2 확률적 잠재 의미 분석 (pLSA)

Probabilistic Latent Semantic Analysis

확률적 잠재 의미 분석 (pLSA)

- 토픽모델링을 위해 잠재 의미 분석에서 사용하는 **특이값 분해가 아닌 확률적 방법을 사용**
- 토픽 모델링 가정 **"문서는 여러 주제로 구성되어 있고, 각 주제는 단어 집합으로 구성된다."**



N : 문서집합

W_d : 문서내 단어집합

d : 문서

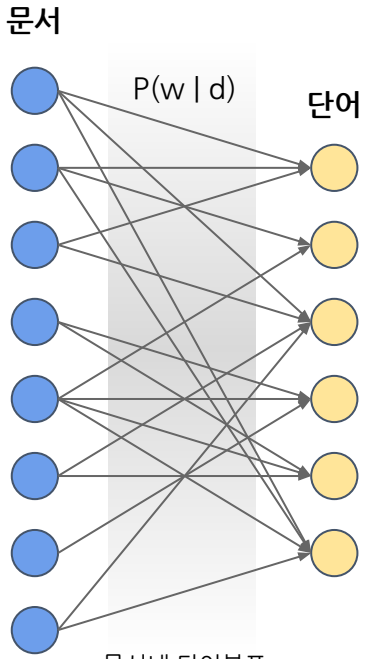
z : 주제

w : 단어

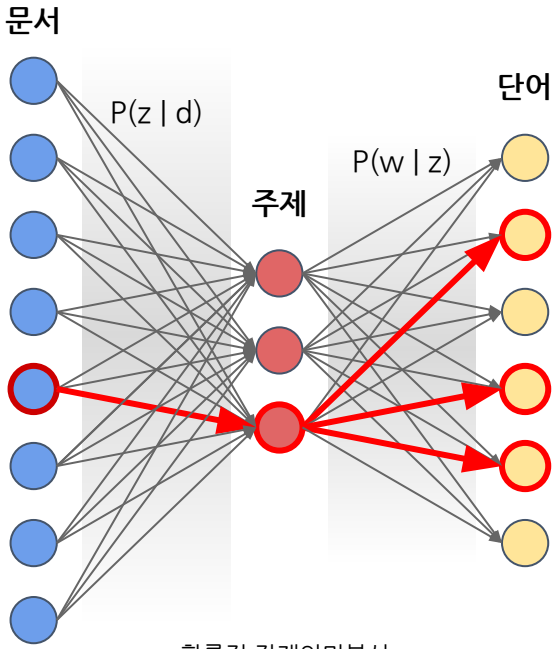
문서(d)가 주어지면 주제(z)는 확률 P(z|d)로 문서 내에 존재한다.

주제(z)가 주어지면 단어(w)는 확률 P(w|z)로 주제 내에 존재한다.

확률적 잠재 의미 분석 (pLSA)



문서내 단어분포



확률적 잠재의미분석

확률적 잠재 의미 분석의 한계

- 모델 P(D)에 대한 매개 변수가 없기 때문에 새 문서에 확률을 할당하는 방법을 모른다
- 확률적 잠재 의미 분석의 파라미터는 분석할 문서 수에 따라 선형적으로 증가하기 때문에 오버피팅 되기 쉽다.
- 확률적 잠재 의미 분석은 거의 사용되지 않기 때문에 코드를 찾아보기 어렵다. 일반적으로 잠재 의미 분석보다 성능이 좋은 모델을 찾고자 할때 이어서 살펴볼 잠재 디리클레 할당을 사용한다.

감사합니다.

Insight campus SeSAC

daniel@fins.ai