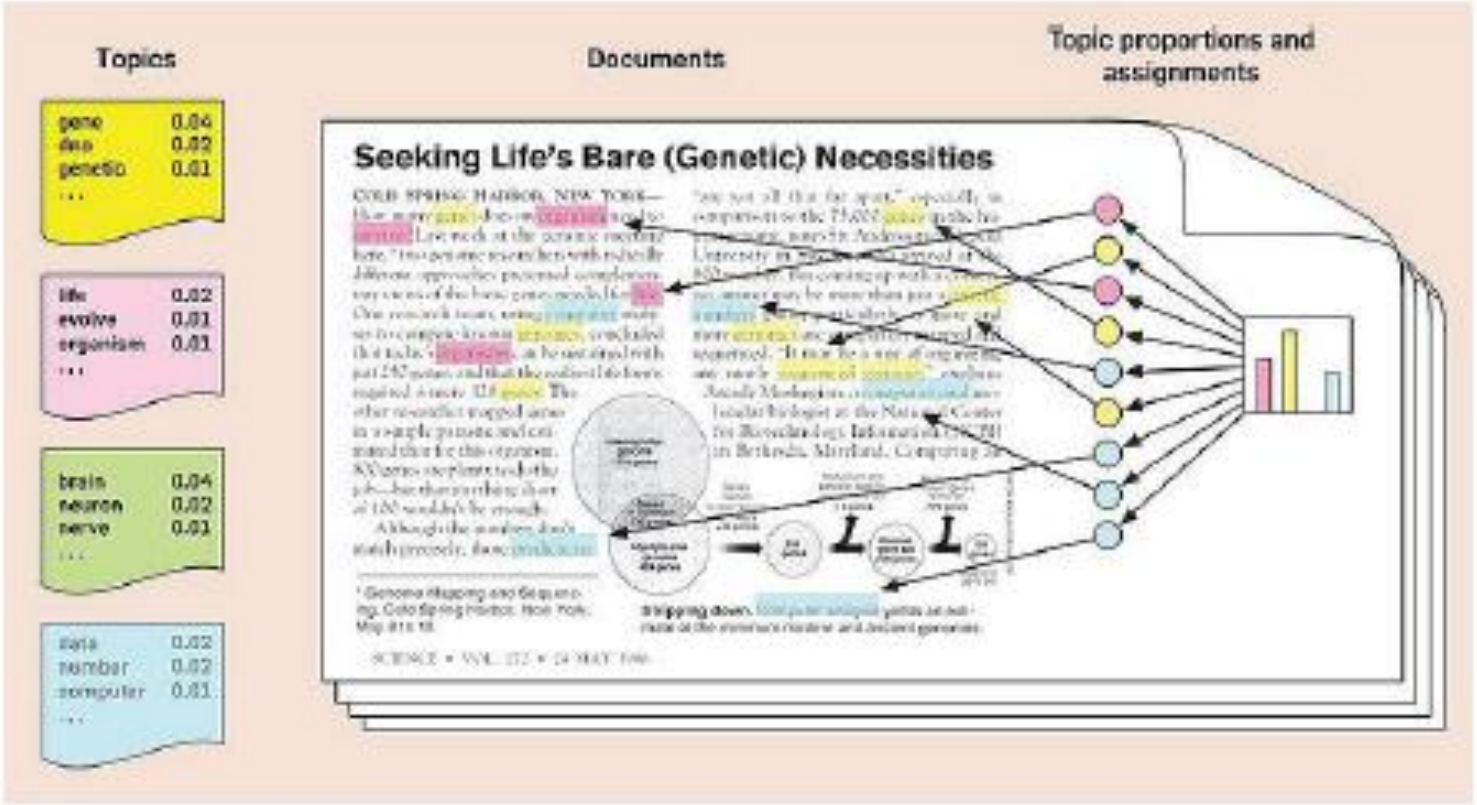
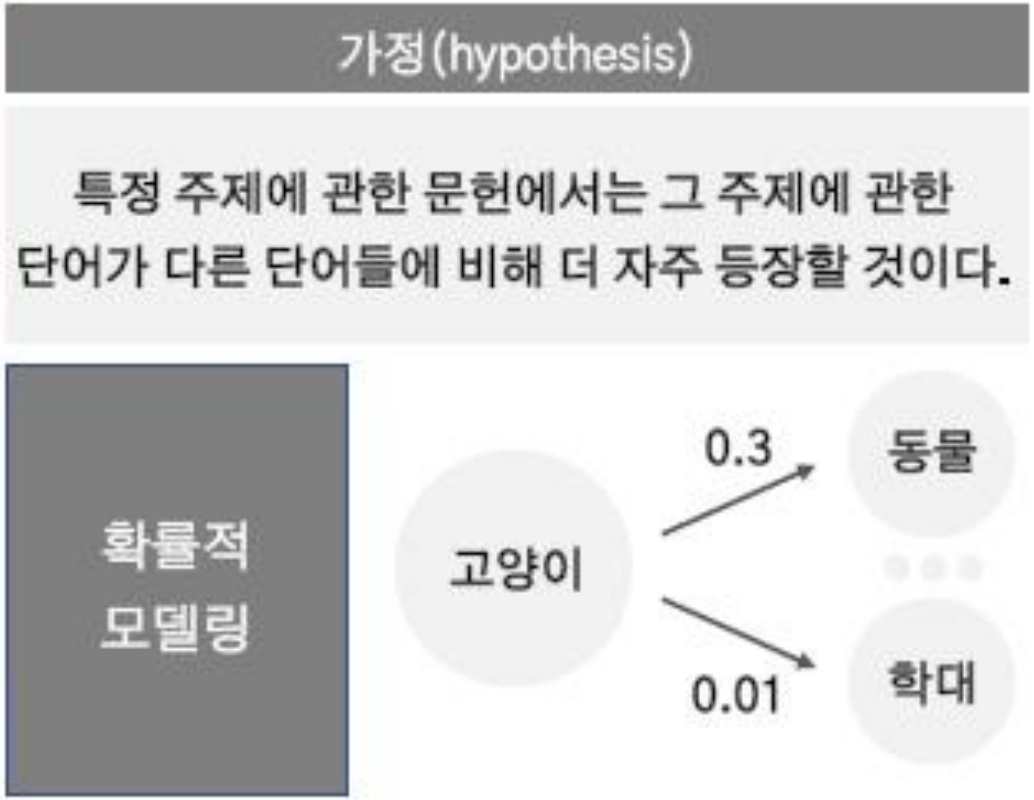
Topic Modeling

hreeee@yonsei.ac.kr
강사 백혜림

Topic modeling

- **토픽모델링은(Topic Modeling)**은 문서의 주제를 발견하기 위한 통계적 모델 중 하나로, 문서 본문의 숨겨진 의미 구조를 발견하기 위해 사용되는 텍스트 마이닝 기법을 말함.



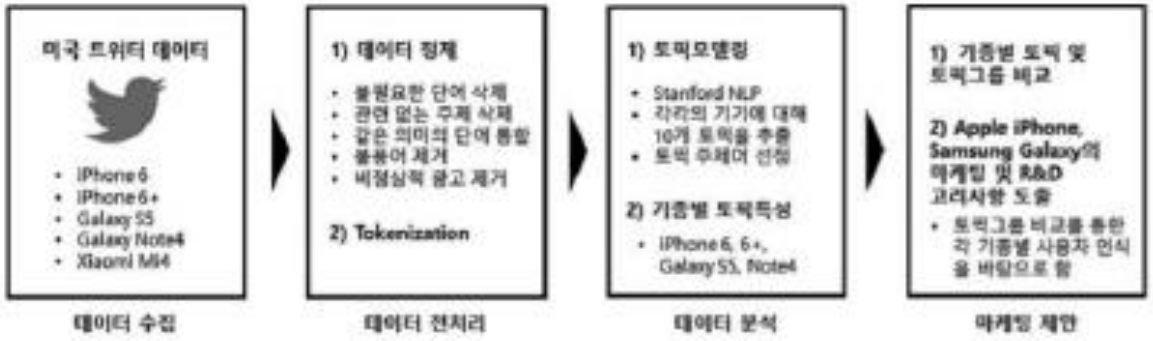


Topic modeling

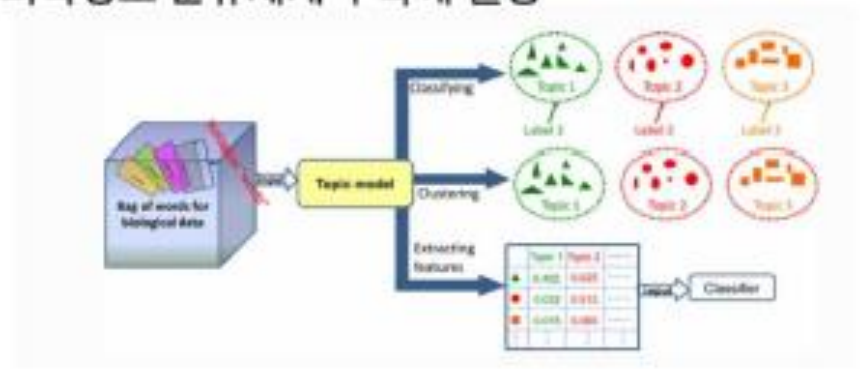
- 토픽모델링의 모델과 활용

토픽모델링(Topic Modeling)은 잠재 의미 분석, 잠재 디리클레 할당, 비순환 방향 그래프등 다양한 모델들이 있으며, 정보 검색 등 다양한 분야에 활용한다.

기업 의사결정 지원(주로 온라인 정보를 기반)에 활용



의학정보 분류체계 구축에 활용



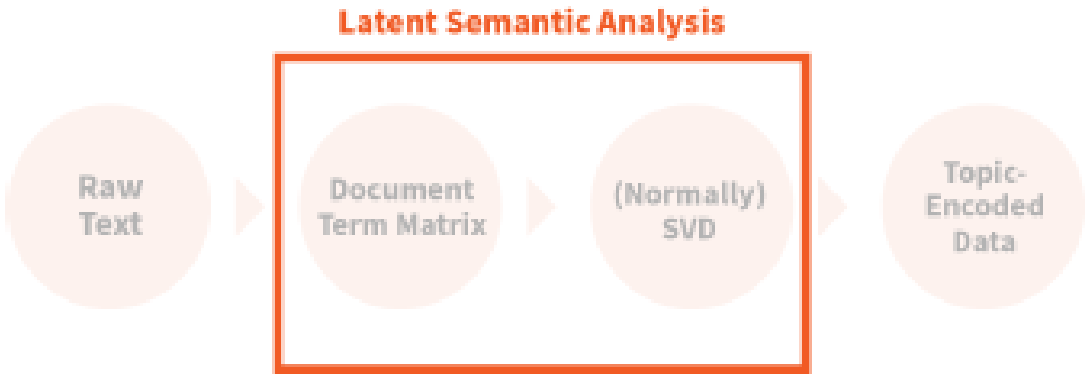
L S A

(Latent Senmantic Analysis)

LSA (Latent Senmantic Analysis)

- 잠재 의미 분석 (LSA)의 정의

잠재 의미 분석(Latent Semantic Analysis)은 문서 간의 관계 혹은 문서 내에서의 특정 단어의 관계를 사용하여 단어와 문장의 의미를 파악하고 이를 분석에 활용하는 것을 말함.



잠재 의미 분석은 기존 BoW를 활용한 DTM(Document-Term Matrix)의 문제점을 극복하기 위해 나타났으며, 주로 토픽 모델링(Topic Modeling)에서 좋은 성능을 보여줍니다.

LSA (Latent Senmantic Analysis)

- 특이값 분해(SVD; Singular Value Decomposition)

Truncated SVD는 Σ 행렬의 대각원소(특이값) 가운데 상위 t개만 골라낸 형태

이렇게 하면 행렬 A를 원복할 수 없게 되지만, 데이터 정보를 상당히 압축했음에도 행렬 A를 근사할 수 있다. 잠재의미분석이 이것을 사용한다.

SVD 예시

$$A = U\Sigma V^T$$

$$\begin{bmatrix} 2 & 3 \\ 1 & 4 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5.47 & 0 & 0 & 0 \\ 0 & 0.37 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.40 & 0.91 \\ -0.91 & 0.40 \end{bmatrix}$$

대각행렬 : q개의 토픽 각각이 전체 말뭉치내에서 얼마나 중요한지 나타내는 가중치

truncated SVD 예시

$$A' = U_1\Sigma_1 V_1^T$$

해당 토픽에 대한 문서들의 분포 정보

$$\begin{bmatrix} 1.79 & 4.08 \\ 1.27 & 2.89 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.82 \\ 0.58 \\ 0 \\ 0 \end{bmatrix} [5.47] [0.40 \quad 0.91]$$ q

각각 해당 토픽에 대한 단어들의 분포정보

LSA (Latent Senmantic Analysis)

- 특이값 분해(SVD; Singular Value Decomposition)

Truncated SVD는 Σ 행렬의 대각원소(특이값) 가운데 상위 t개만 골라낸 형태

이렇게 하면 행렬 A를 원복할 수 없게 되지만, 데이터 정보를 상당히 압축했음에도 행렬 A를 근사할 수 있다. 잠재의미분석이 이것을 사용한다.

SVD 예시

$$A = U\Sigma V^T$$

$$\begin{bmatrix} 2 & 3 \\ 1 & 4 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5.47 & 0 & 0 & 0 \\ 0 & 0.37 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.40 & 0.91 \\ -0.91 & 0.40 \end{bmatrix}$$

truncated SVD 예시

$$A' = U_1\Sigma_1 V_1^T$$

$$\begin{bmatrix} 1.79 & 4.08 \\ 1.27 & 2.89 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.82 \\ 0.58 \\ 0 \\ 0 \end{bmatrix} [5.47] [0.40 \quad 0.91]$$

LSA Example

doc1 : 나,는,학교,에,가,ㄴ,다

doc2 : 학교,에,가,는,영희

doc3 : 나,는,영희,는,좋,다

- 잠재 의미 분석 (LSA)의 정의

잠재의미분석이란 위와 같은 **단어-문서행렬(Word-Document Matrix)**, **단어-문맥행렬(window based co-occurrence matrix)** 등 입력 데이터에 특이값 분해를 수행해 데이터의 차원수를 줄여 계산 효율성을 키우는 한편 행간에 숨어있는(latent) 의미를 이끌어내기 위한 방법론

-	doc1	doc2	doc3
나	1	0	0
는	1	1	2
학교	1	1	0
에	1	1	0
가	1	1	0
ㄴ	1	0	0
다	1	0	1
영희	0	1	1
좋	0	0	1

LSA Example

$$A = U\Sigma V^T$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -0.17 & 0.27 & -0.40 \\ -0.63 & -0.41 & -0.03 \\ -0.32 & 0.37 & 0.21 \\ -0.32 & 0.37 & 0.21 \\ -0.32 & 0.37 & 0.21 \\ -0.17 & 0.27 & -0.40 \\ -0.33 & -0.12 & -0.52 \\ -0.30 & -0.29 & 0.49 \\ -0.15 & -0.39 & -0.13 \end{bmatrix} \begin{bmatrix} 3.61 & 0 & 0 \\ 0 & 2.04 & 0 \\ 0 & 0 & 1.34 \end{bmatrix} \begin{bmatrix} -0.63 & -0.53 & -0.57 \\ 0.56 & 0.20 & -0.80 \\ -0.54 & 0.83 & -0.17 \end{bmatrix}$$

LSA Example

$$A' = U_2 \Sigma_2 V_2^T$$

$$\begin{bmatrix} 0.71 & 0.44 & -0.09 \\ 0.97 & 1.04 & 1.99 \\ 1.15 & 0.76 & 0.04 \\ 1.15 & 0.76 & 0.04 \\ 1.15 & 0.76 & 0.04 \\ 0.71 & 0.45 & -0.09 \\ 0.62 & 0.58 & 0.88 \\ 0.36 & 0.45 & 1.11 \\ -0.09 & 0.14 & 0.97 \end{bmatrix} = \begin{bmatrix} -0.17 & 0.27 \\ -0.63 & -0.41 \\ -0.32 & 0.37 \\ -0.32 & 0.37 \\ -0.32 & 0.37 \\ -0.17 & 0.27 \\ -0.33 & -0.12 \\ -0.30 & -0.29 \\ -0.15 & -0.39 \end{bmatrix} \begin{bmatrix} 3.61 & 0 \\ 0 & 2.04 \end{bmatrix} \begin{bmatrix} -0.63 & -0.53 & -0.57 \\ 0.56 & 0.20 & -0.80 \end{bmatrix}$$

LSA Example

$$round(A^{'}) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

LSA Example

$$A' = U_2 \Sigma_2 V_2^T$$

$$X_1 = \begin{bmatrix} -2.28 & -1.90 & -2.07 \\ 1.14 & 0.42 & -1.64 \end{bmatrix}$$

LSA Example

$$U_2^\top \times A' = U_2^\top \times U_2 \Sigma_2 V_2^T$$

$$X_1 = \begin{bmatrix} -2.28 & -1.90 & -2.07 \\ 1.14 & 0.42 & -1.64 \end{bmatrix}$$

LSA Example

$$U_2^\top A' = \Sigma_2 V_2^T$$

$$X_1 = \begin{bmatrix} -2.28 & -1.90 & -2.07 \\ 1.14 & 0.42 & -1.64 \end{bmatrix}$$

LSA Example

$$U_2^\top A' = \Sigma_2 V_2^T$$

$$X_1 = \begin{bmatrix} -2.28 & -1.90 & -2.07 \\ 1.14 & 0.42 & -1.64 \end{bmatrix}$$

문서(문장)

단어(변수)

A에서는 9개로 표현, 단 2개로 표현가능

LSA Example

반대로 해봅시다

$$A' \times V_2 = U_2 \Sigma_2 V_2^T \times V_2$$

$$X_2 = \begin{bmatrix} -0.63 & 0.56 \\ -2.30 & -0.84 \\ -1.16 & 0.76 \\ -1.16 & 0.76 \\ -1.16 & 0.76 \\ -0.63 & 0.56 \\ -1.20 & -0.24 \\ -1.10 & -0.60 \\ -0.57 & -0.80 \end{bmatrix}$$

LSA Example

A에서는 3개로 표현, 단 2개로 표현가능

문서(문장)

반대로 해봅시다

$$A' V_2 = U_2 \Sigma_2$$

$$X_2 = \begin{bmatrix} -0.63 & 0.56 \\ -2.30 & -0.84 \\ -1.16 & 0.76 \\ -1.16 & 0.76 \\ -1.16 & 0.76 \\ -0.63 & 0.56 \\ -1.20 & -0.24 \\ -1.10 & -0.60 \\ -0.57 & -0.80 \end{bmatrix}$$

단어(변수)

LSA (Latent Senmantic Analysis)

- **장점**
- 단어와 문맥 간의 내재적인의미를 효과적으로 보전할 수 있어, 결과적으로 문서간 유사도 측정등 모델의 성능 향상에 도움을 줌.
- 입력 데이터의 노이즈 제거, sparsity를 줄이게 돼 효과가 좋음
- But, 새로운 문서나 단어가 추가되면 아예 처음부터 작업을 새로 시작해야 함.
- 이 때문에, 최근에는 word2vec등 뉴럴네트워크 기반의 representation방법론이 각광받고 있다

L D A

(Latent Dirichlet Allocation)

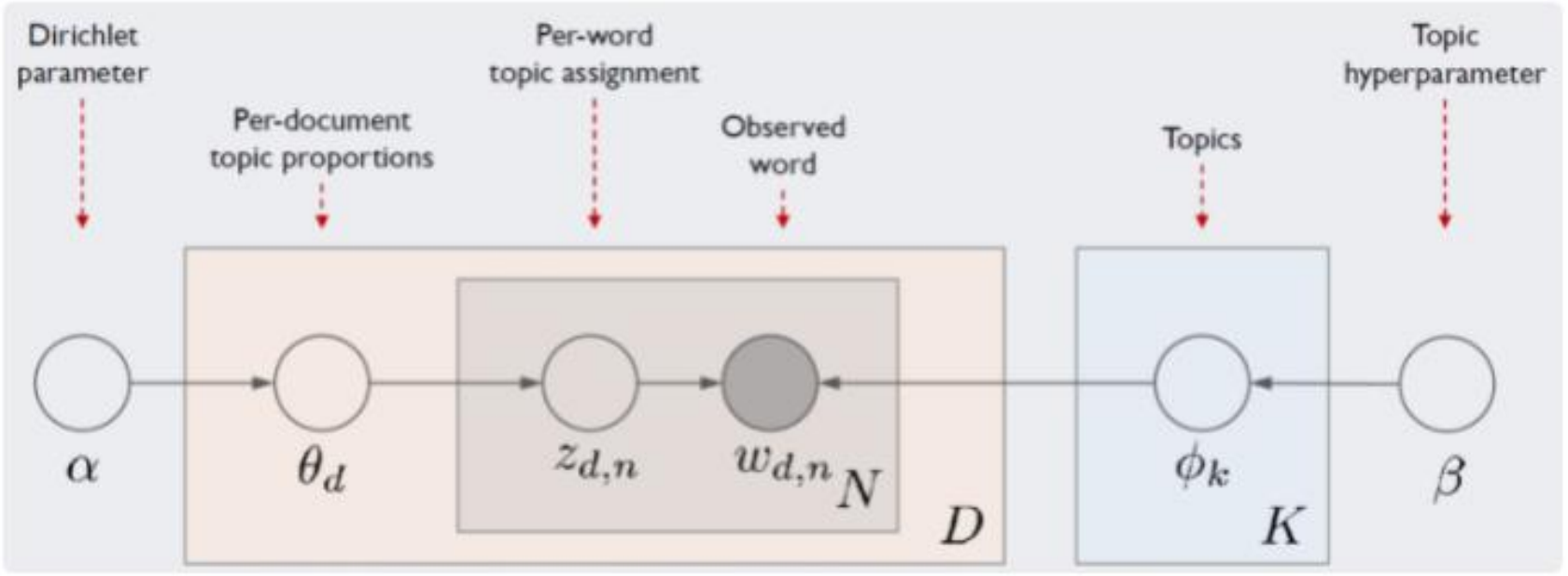
LDA (Latent Dirichlet allocation)

- 잠재 디리클레 할당 (LDA)의 정의

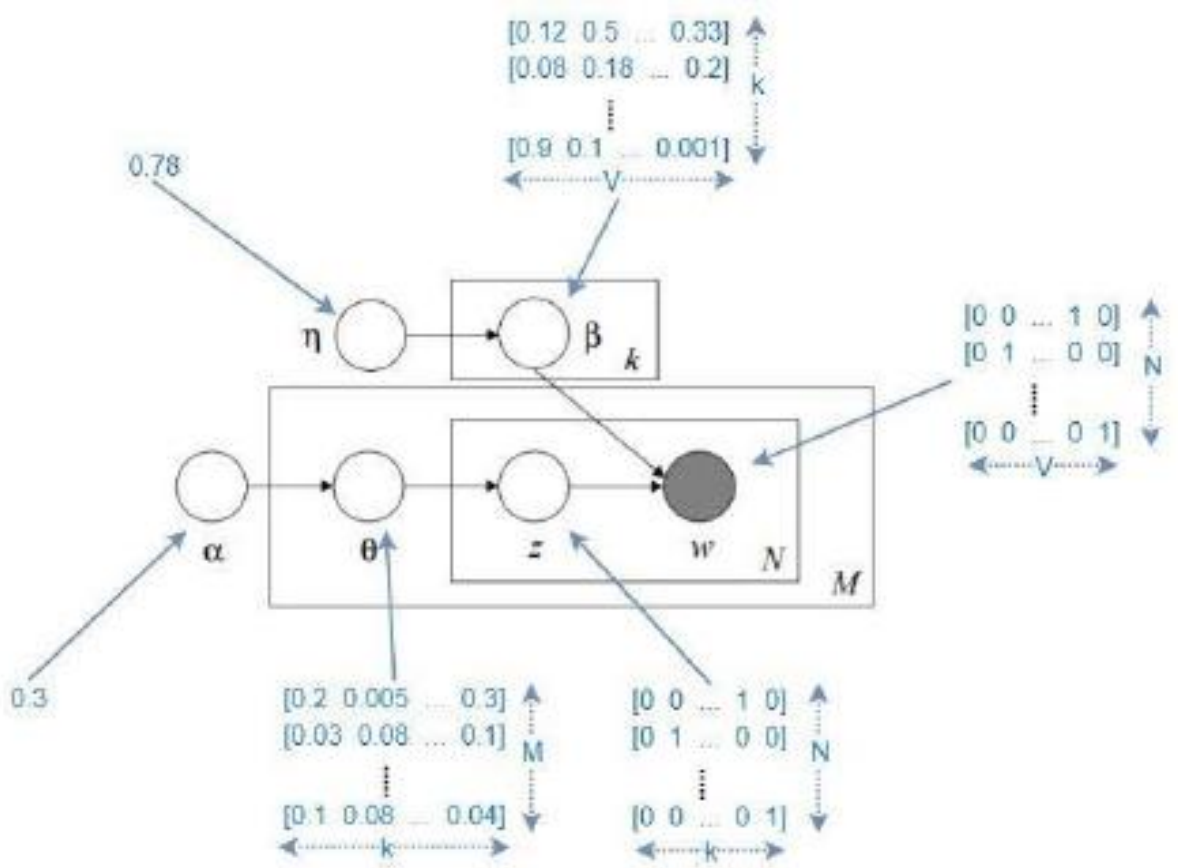
자연어 처리에서 잠재 디리클레 할당(Latent Dirichlet allocation, LDA)은 주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는지를 서술하는 대한 확률적 토픽 모델 기법 중 하다

- 잠재 디리클레 할당 (LDA)의 특징

잠재 디리클레 할당에서 각 주제와 단어의 분포는 디리클레 확률 분포를 따른다. 또한, 각 단어가 어떤 주제에 속하는지를 나타내는 잠재 z가 존재하며, 이는 다항 분포를 따른다.



LDA (Latent Dirichlet allocation)



$$p(\phi_k \mid \beta) \sim Dir(\beta)$$

$$p(\theta_d \mid \alpha) \sim Dir(\alpha)$$

$$p(z_{d,n} \mid \theta_d) \sim Multinomial(\theta_d)$$

$$p(z_{d,n} = k \mid \theta_d) = (\theta_d)_k$$

$$p(w_{d,n} \mid z_{d,n}, \phi) \sim Multinomial(\phi_{z_{d,n}})$$

$$p(w_{d,n} = v \mid z_{d,n}, \phi_1, \ldots, \phi_K) = (\phi_{z_{d,n}})_v$$

LDA (Latent Dirichlet allocation)

d번째 문서 i번째 단어의 토픽 $z_{\{d,j\}}$ 가 j번째 할당될 확률

$$p(z_{d,i}=j|z_{-i},w) = \frac{n_{d,k}+\alpha_j}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)} \times \frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j}+\beta_j)} = AB$$

표기	내용
$n_{d,k}$	k번째 토픽에 할당된 d번째 문서의 단어 빈도
$v_{k,w_{d,n}}$	전체 말뭉치에서 k번째 토픽에 할당된 단어 $w_{d,n}$의 빈도
$w_{d,n}$	d번째 문서에 n번째로 등장한 단어
α	문서의 토픽 분포 생성을 위한 디리클레 분포 파라메터
β	토픽의 단어 분포 생성을 위한 디리클레 분포 파라메터
K	사용자가 지정하는 토픽 수
V	말뭉치에 등장하는 전체 단어 수
A	d번째 문서가 k번째 토픽과 맺고 있는 연관성 정도
B	d번째 문서의 n번째 단어($w_{d,n}$)가 k번째 토픽과 맺고 있는 연관성 정도

LDA (Latent Dirichlet allocation)