Topic Modeling

토픽 모델링의 이해

- 토픽 모델링은 문서들에 잠재되어 있는 공통된 토픽(주제)들을 추출해 내는 기법을 의미합니다.

- 공통된 유사성을 도출한다는 측면에서 문서 군집화/유사도와 비슷한 기법일 수 있지만 토픽 모델링은 문서들이 가지는 주요 토픽의 분포도와 개별 토픽이 어떤 의미인지를 제공하는 특징을 가지고 있습니다.

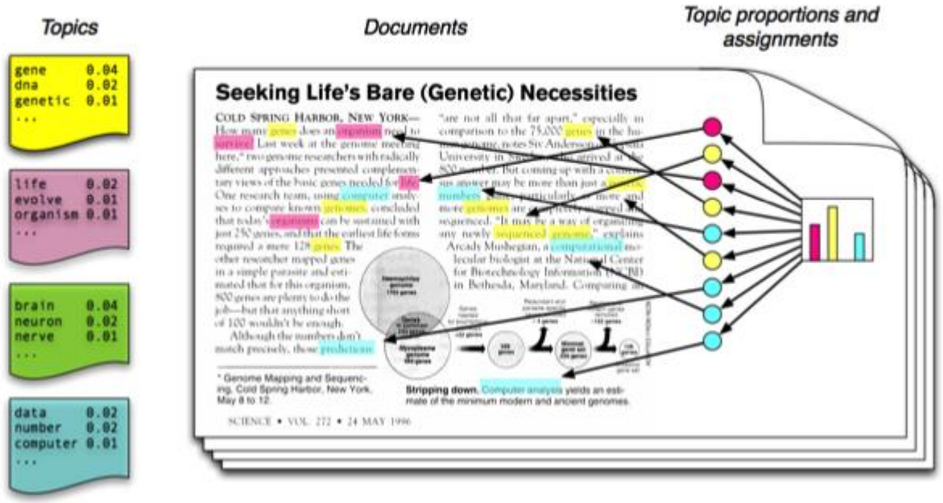


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

토픽 모델링의 이해

나는 바나나와 오렌지를 좋아한다.

개구리와 물고기는 연못에 산다

나는 시금치와 사과 주스를 먹는다

고양이와 강아지는 푹신하다

내 고양이는 케일을 좋아한다.

토픽 모델링의 이해

나는 바나나와 오렌지를 좋아한다.

개구리와 물고기는 연못에 산다

나는 시금치와 사과 주스를 먹는다

고양이와 강아지는 푹신하다

내 고양이는 케일을 좋아한다.

토픽 추출

토픽 A

바나나 40%
오렌지 30%
케일 30%,
아침밥 30%
…

토픽 B

고양이 30%
강아지 20%
개구리 10%
물고기 10%
…

토픽 모델링의 이해

나는 바나나와 오렌지를 좋아한다.

개구리와 물고기는 연못에 산다

나는 시금치와 사과 주스를 먹는다

고양이와 강아지는 푹신하다

내 고양이는 케일을 좋아한다.

토픽 추출

나는 바나나와 오렌지를 좋아한다.

토픽 A

개구리와 물고기는 연못에 산다

토픽 B

나는 시금치와 사과 주스를 먹는다

토픽 A

고양이와 강아지는 푹신하다

토픽 B

내 고양이는 케일을 좋아한다.

토픽 A
+
토픽 B

토픽 A

바나나 40%
오렌지 30%
케일 30%,
아침밥 30%
…

토픽 B

고양이 30%
강아지 20%
개구리 10%
물고기 10%
…

토픽 모델링의 이해

나는 바나나와 오렌지를 좋아한다.

개구리와 물고기는 연못에 산다

나는 시금치와 사과 주스를 먹는다

고양이와 강아지는 푹신하다

내 고양이는 케일을 좋아한다.

토픽 추출

나는 바나나와 오렌지를 좋아한다.
음식

개구리와 물고기는 연못에 산다
동물

나는 시금치와 사과 주스를 먹는다
음식

고양이와 강아지는 푹신하다
동물

내 고양이는 케일을 좋아한다.
음식 + 동물

음식

바나나 40%
오렌지 30%
케일 30%,
아침밥 30%
…

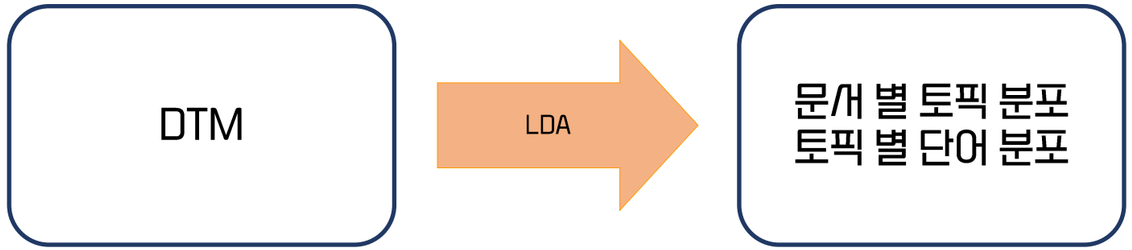
동물

고양이 30%
강아지 20%
개구리 10%
물고기 10%
…

LDA(Latent Dirichlet Allocation)

LDA(Latent Dirichlet Allocation)의 이해

- Document Term Matrix에서 문서 별 토픽 분포와 토픽 별 단어 분포를 찾아가는 기법입니다.



- LDA는 관찰된 문서 내 단어들을 이용하여 베이즈 추론을 통해 잠재된 문서 내 토픽 분포와 토픽 별 단어 분포를 추론하는 방식입니다.
- 이 때 LDA의 베이즈 추론의 사전 확률 분포로 사용되는 것이 디리클레 분포 입니다.
  - 사전 확률을 초기에 계산한 후 다시 사후 확률이 계산이 되면, 그 사후 확률을 다시 사전 확률로 업데이트 하는 방식입니다.

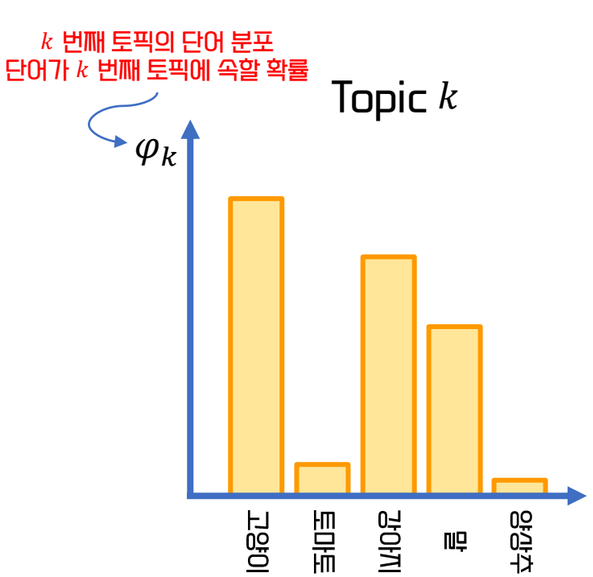
베이즈 추론 켤레 사전 분포

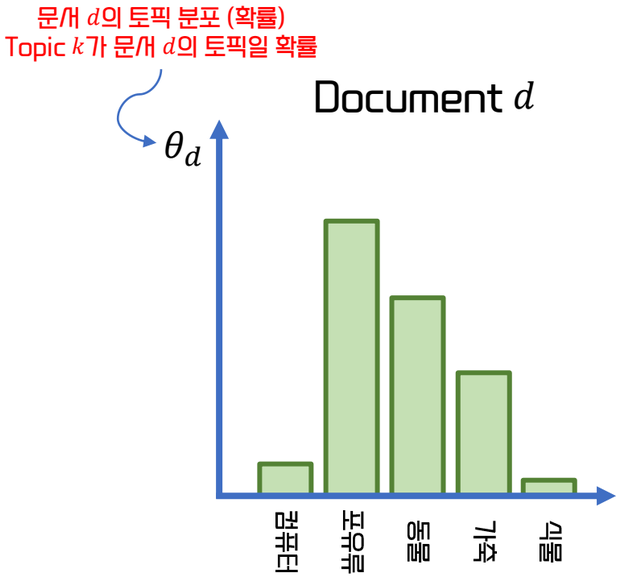
- 이항 분포
  - 두 가지 중에 하나만 발생하는 사건이 여러 번 반복되는 확률 분포 입니다. 예를 들면 동전을 10번 던져 2번이 앞이 되는 경우에 대한 확률 분포라 할 수 있습니다.
  - 켤레 사전 분포는 베타 분포 입니다.

- 다항 분포
  - 여러 개(보통은 3개) 중에 하나가 발생하는 사건이 여러 번 반복되는 확률 분포 입니다. 예를 들면 로또는 100번 사서 1등에서 5등까지가 몇 번씩 나왔는지에 대한 확률 분포라 하겠습니다.
  - 켤레 사전 분포는 디리클레 분포 입니다.

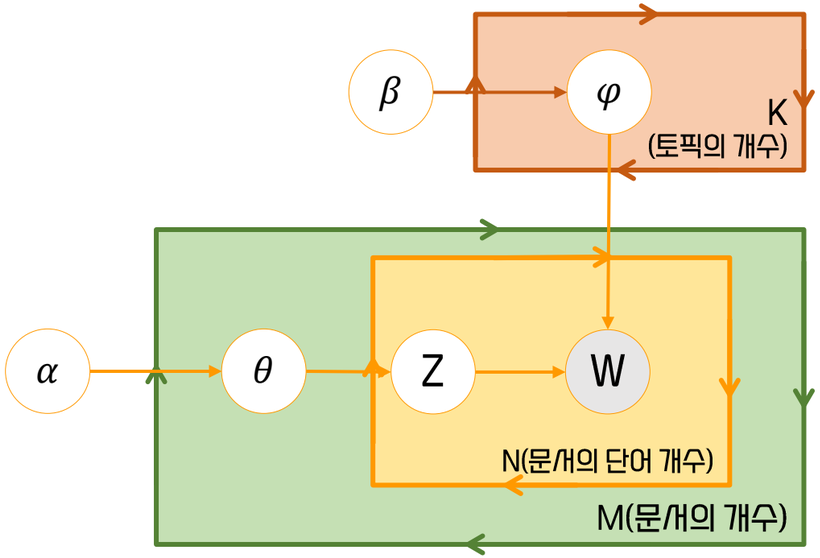
- 관측되는 단어분포와 디리클레 사전 확률 분포를 결합하여 지속적으로 문서의 주제 분포와 주제 단어 분포들의 사후 확률 분포를 업데이터 합니다.

토픽 모델링의 분포

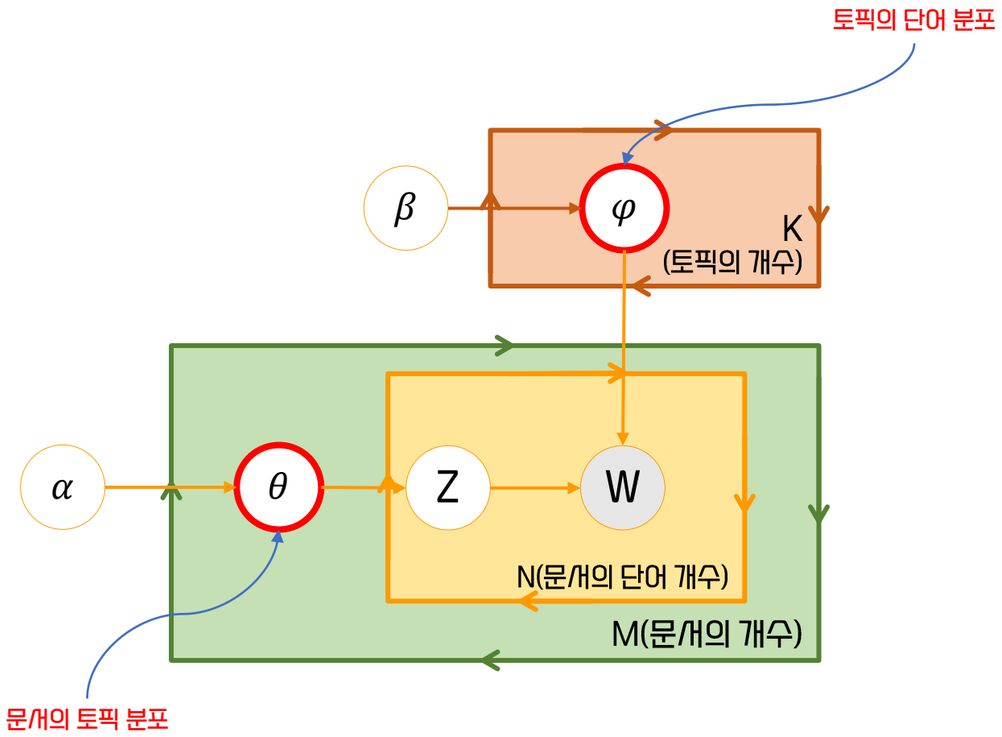

k 번째 토픽의 단어 분포
단어가 k 번째 토픽에 속할 확률
Topic k
φ_k
고양이
토마토
강아지
말
양갱수


문서 d의 토픽 분포 (확률)
Topic k가 문서 d의 토픽일 확률
Document d
θ_d
컴퓨터
포유류
동물
가축
식물

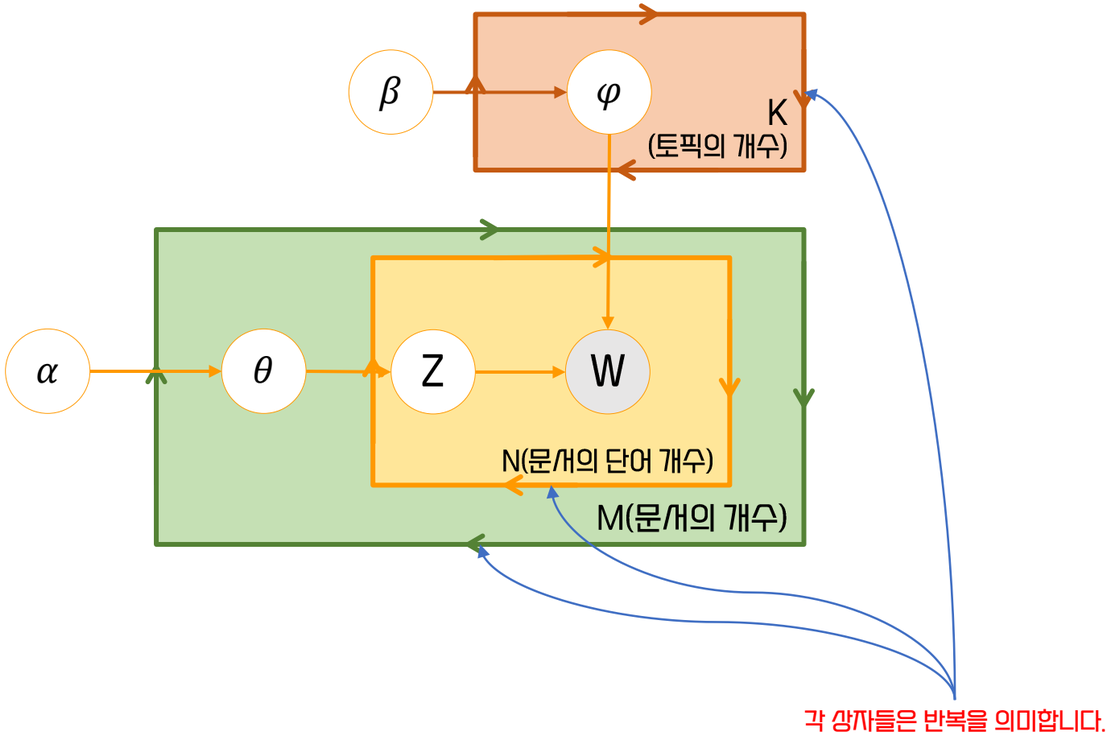
LDA 구성 요소와 작동 방식



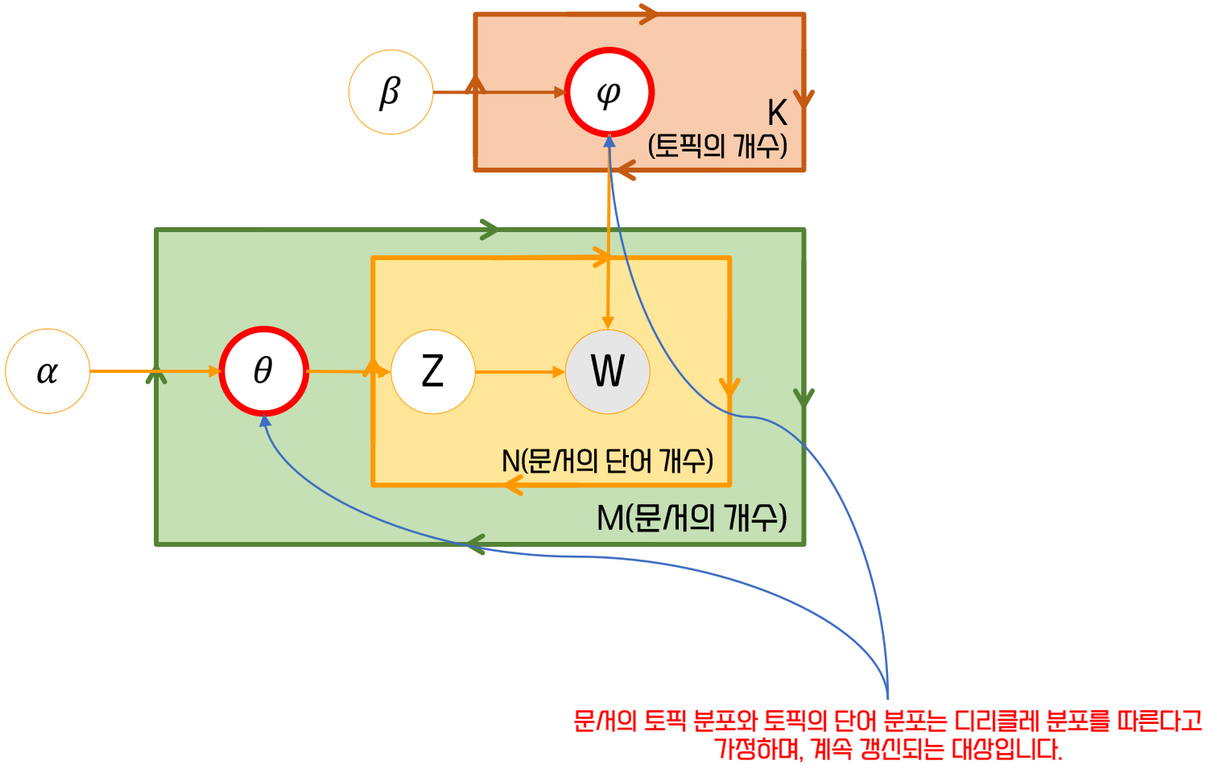
LDA 구성 요소와 작동 방식



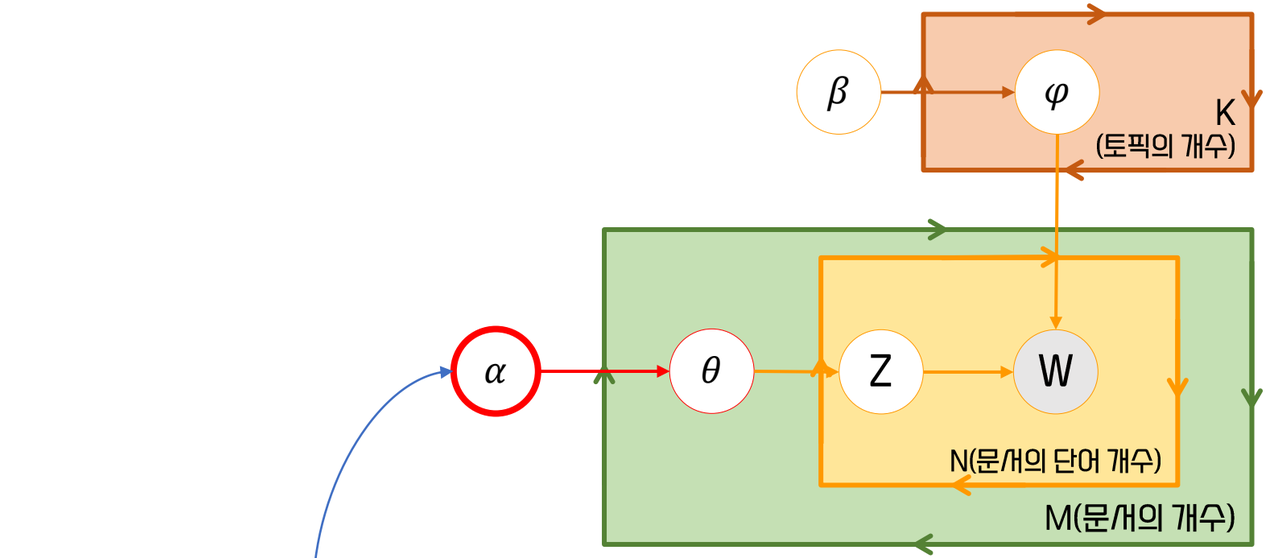
LDA 구성 요소와 작동 방식



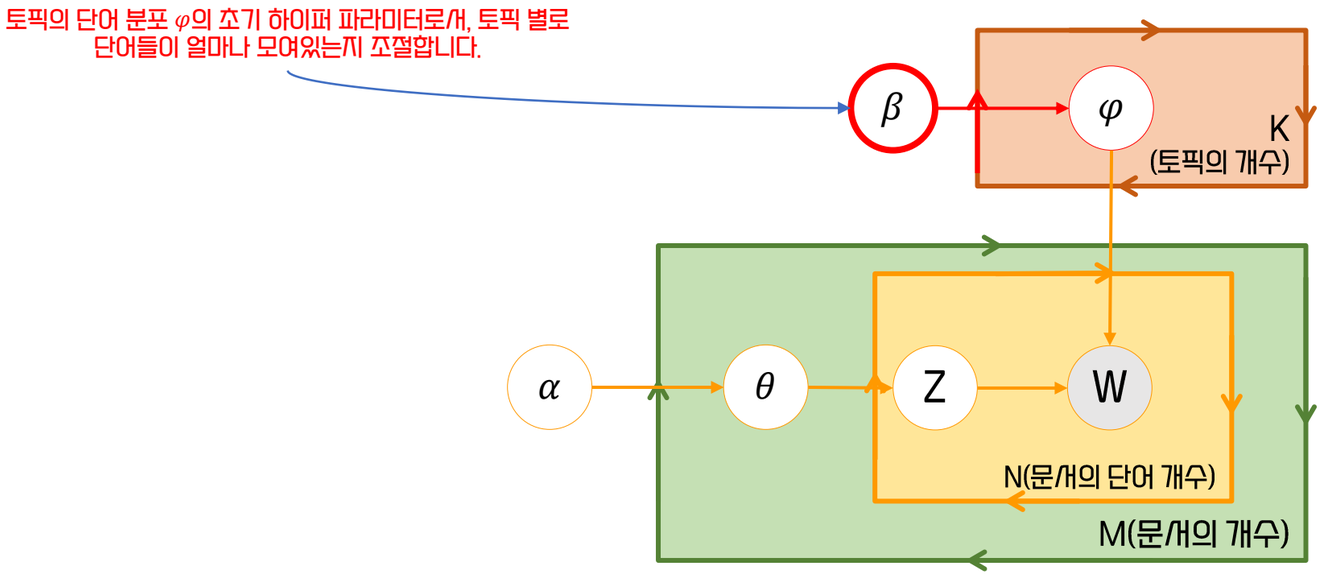
LDA 구성 요소와 작동 방식



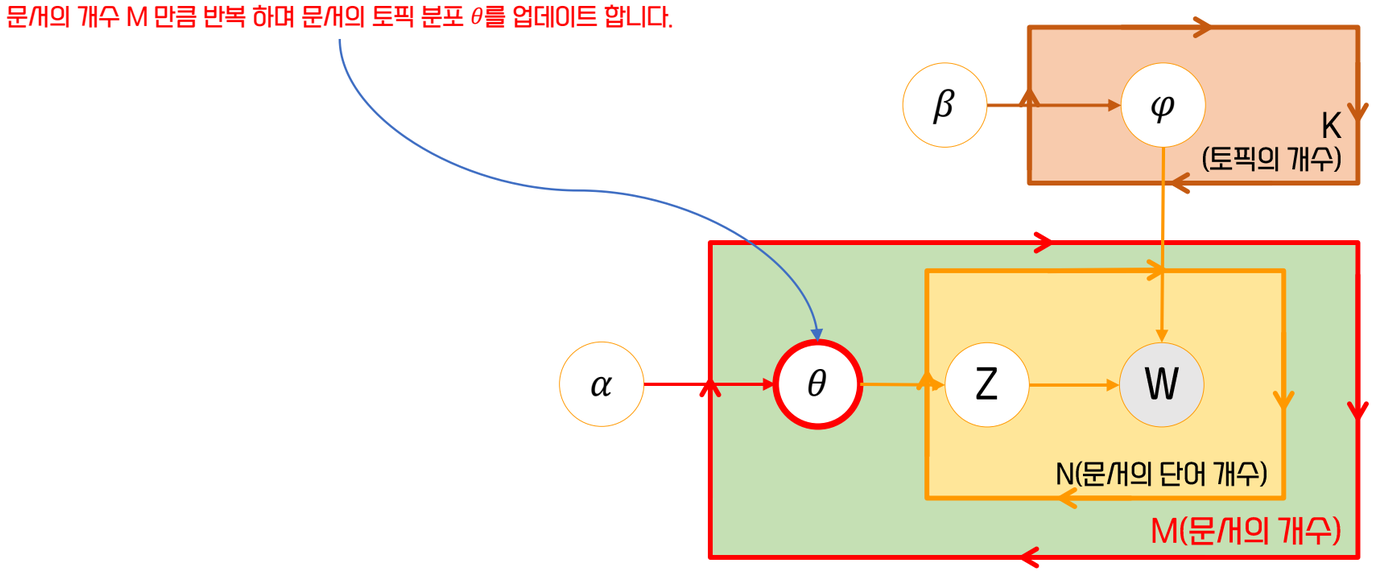
LDA 구성 요소와 작동 방식



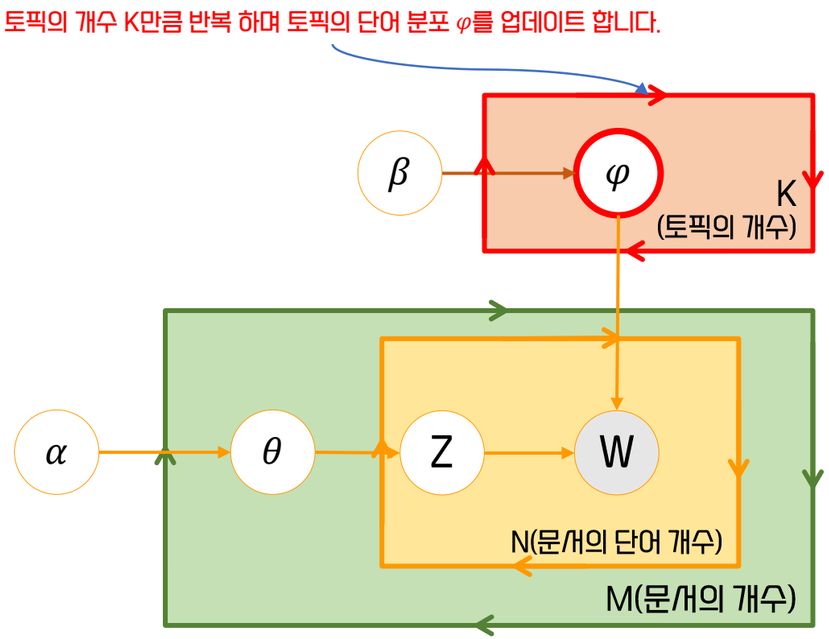
LDA 구성 요소와 작동 방식



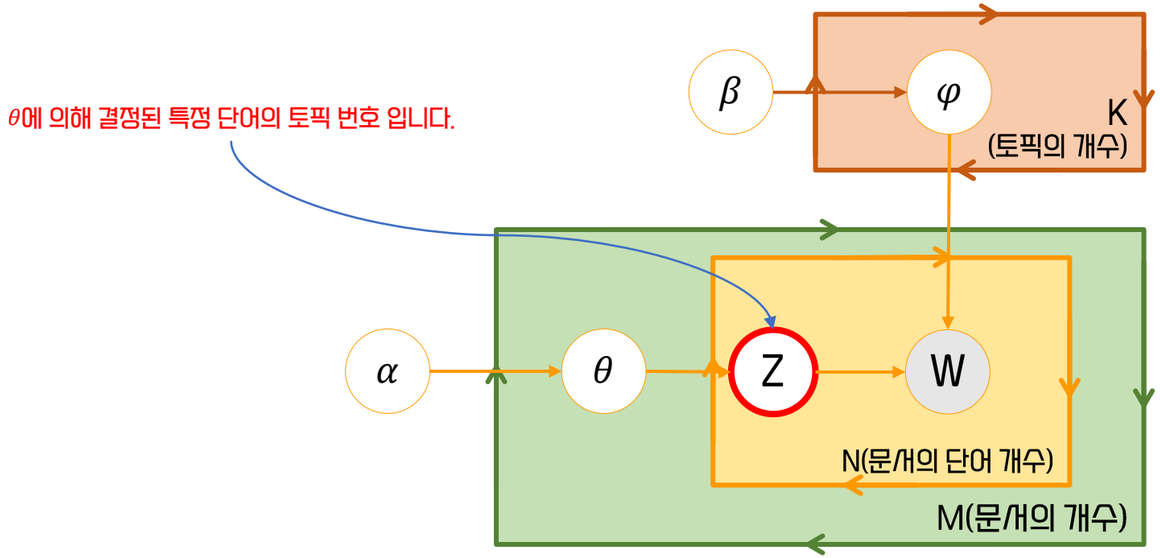
LDA 구성 요소와 작동 방식



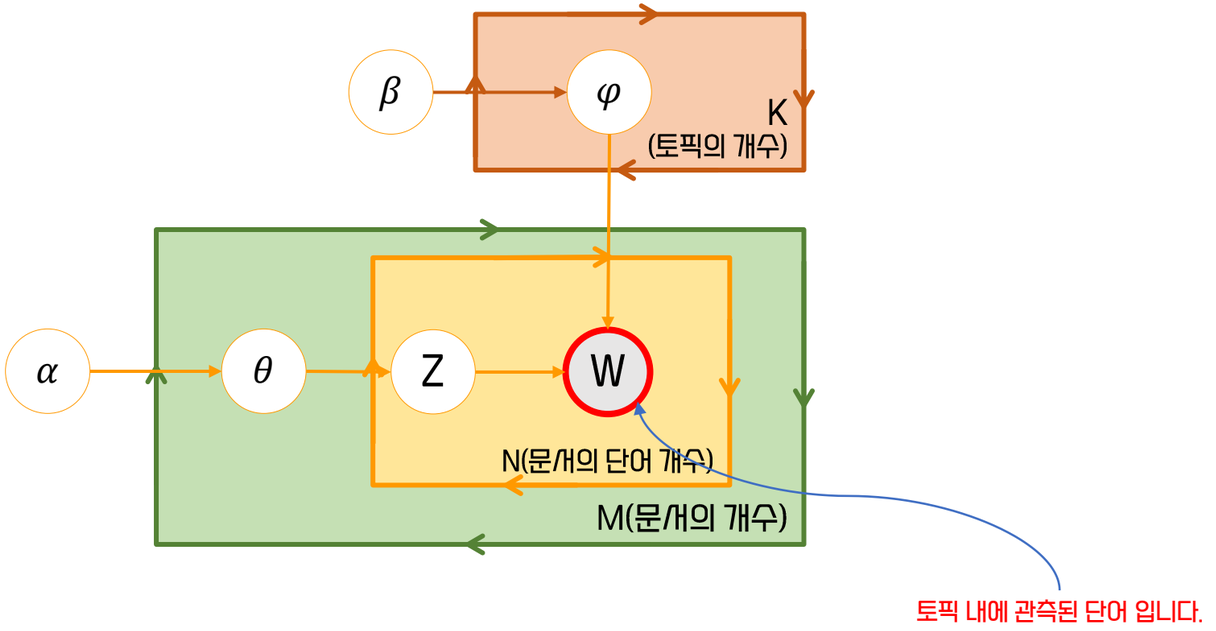
LDA 구성 요소와 작동 방식



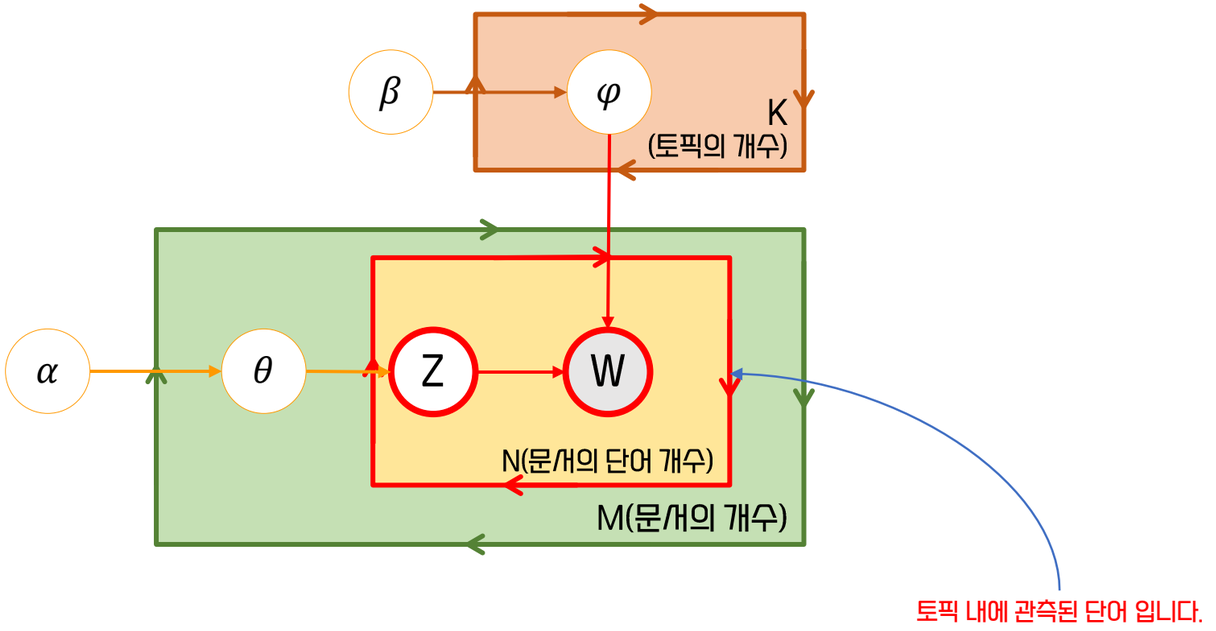
LDA 구성 요소와 작동 방식



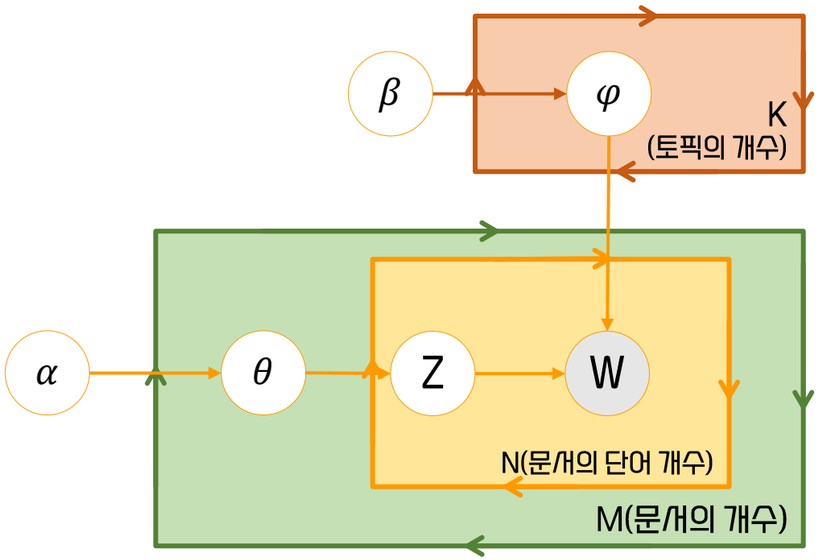
LDA 구성 요소와 작동 방식



LDA 구성 요소와 작동 방식



LDA 구성 요소와 작동 방식



문서 1	이 집은 사장님이 너무 친절하세요. 최고!
문서 2	이 집은 감자탕이 너무 맛있어요~! 고기가 살살 녹아요 ㅎㅎ
문서 3	사장님과 알바생 분들도 친절하시고 감자탕도 진짜 맛있어요

	문서 1		문서 2		문서 3			
단어	사장님	친절	감자탕	고기	사장님	알바생	친절	감자탕
토픽	토픽 1	토픽 1	토픽 2	토픽 2	토픽 1	토픽 1	토픽 1	토픽 2

θ 문서 내 토픽 분포

	문서 1	문서 2	문서 3
토픽 1	1	0	0.75
토픽 2	0	1	0.25

φ 문서 내 토픽 분포

	사장님	친절	알바생	감자탕	고기
토픽 1	0.4	0.4	0.2	0	0
토픽 2	0	0	0	0.66	0.33