☼ 토픽 모델링의 개요

- 구조화되지 않는 대량의 텍스트로부터 숨겨져 있는 주제구조를 발견하기 위한 통계적 추론 알고리즘

☼ LDA의 개요

- 문서 같은 데이터의 집합에 대한 Generative Probabilistic Model (생성적 확률모델)
- 1. Choose N ~ Poisson(ξ).
  2. Choose θ ~ Dir(α).
  3. For each of the N words W_n
     (a) Choose a topic Z_n~Multinomial(θ)
     (b) Choose a word W_n from ($W_n \mid Z_n, \beta$),
     a multinomial probability conditioned on the topic Z_n.

☼ 토픽 모델링의 주요변수

- β_k: 단어 사전에서 i번째 단어가 k번째 주제에 해당할 확률
- w_k: i번째 단어이면서 k번째 주제에 해당하는 단어
- z_k: i번째 단어의 k번째의 주제
- θ: 디리클레 분포에서 추출되는 차원 k를 갖는 주체벡터
- k: 주체(토픽)의 개수
- N: 문서의 길이