Chapter. 04

토픽 모델링

잠재 디리클레 할당 (LDA)

FASTCAMPUS

ONLINE

금융공학/퀀트 I

강사. 장순용

키포인트

- 토픽 모델링.
- 연역추론과 귀납추론.
- 잠재 디리클레 할당 (LDA).

I 잠재 디리클레 할당 LDA (Latent Dirichlet Allocation)

- Topic 모델링의 대표적인 방법.
- 텍스트 데이터 바탕의 일종의 “군집분석”이다.

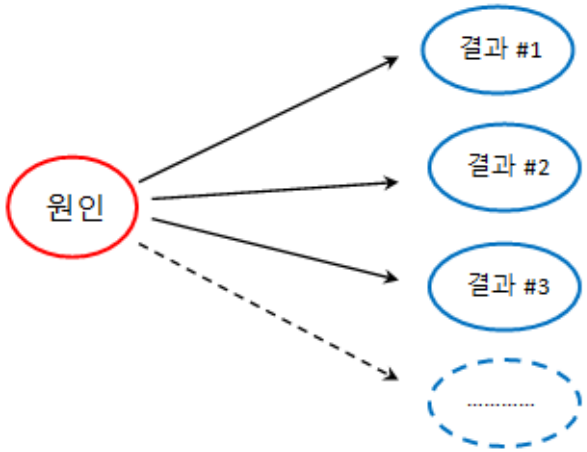
  ⇒ LDA의 topic ≈ 기존 군집분석의 centroid.

  ⇒ Topic을 중심점으로 삼아서 유사한 문서들을 **군집화** 할 수 있다.

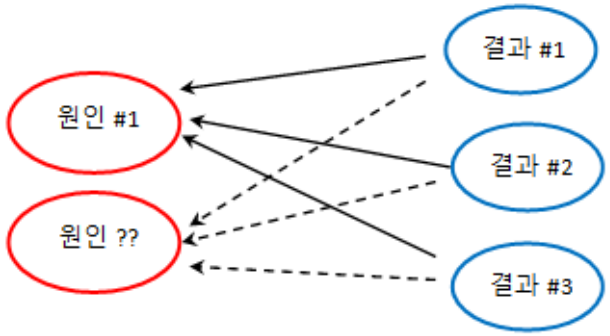
- 새로운 문서가 있을 때에도 유사도를 사용해서 가장 가까운 topic으로 분류할 수 있다.

I 연역 추론과 귀납 추론

- 다음과 같이 “연역 추론”과 “귀납 추론”에 대해서 생각해 본다.



연역 추론



귀납 추론

I 잠재 디리클레 할당 (LDA)의 원리

- 디리클레 확률분포는 베이즈 통계법에서 다항분포의 사전켤레 (prior conjugate)이다.
- LDA에서는 디리클레 분포를 따르고 있는 잠재 토픽 모델을 전제하고 단어는 다항분포를 따르게 됨.
- 단어를 관찰 했을 때, 베이즈 통계법을 적용해서 연역과 귀납 추론을 연결시켜서 토픽을 알아 낼 수 있음.

I 잠재 디리클레 할당 (LDA)의 원리

- 먼저 “연역 추론” 과정에 대해서 생각해 본다.

⇒ 문서가 생성되는 과정 (모형)을 알고 있다는 가정을 해본다.

⇒ 다음과 같은 순서로 단어의 집합인 문서가 생성된다고 생각할 수 있다.

K = 토픽의 개수, M = 문서의 개수.

Dirichlet(β)확률 분포에 의해서 특정 토픽의 확률이 정해짐.

φ_k , $k = 1, \ldots, K$

→

Dirichlet(α) 확률 분포에 의해서 특정 문서의 확률이 정해짐.

θ_i , $i = 1, \ldots, M$

→

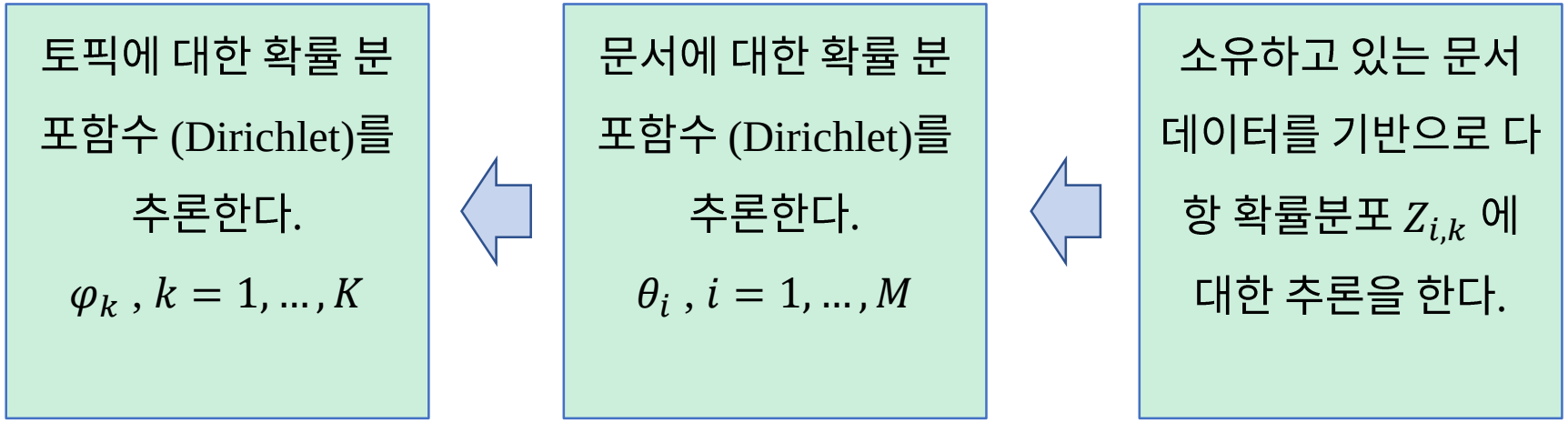
선택된 문서에서 선택된 토픽의 다항 확률 분포 $Z_{i,k}$ 에 의해서 단어가 추가된다.

I 잠재 디리클레 할당 (LDA)의 원리

- 이제는 “귀납 추론” 과정에 대해서 알아본다.

⇒ 문서 데이터가 있고 모형은 모르는 상태이다.

⇒ 연역 추론과는 반대 방향으로 감. K = 토픽의 개수, M = 문서의 개수.



끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.