동국대학교
dongguk university




Lecture 6
토픽모델링

텍스트마이닝

웹크롤링

텍스트 전처리

토픽분석

Tokeniz ation

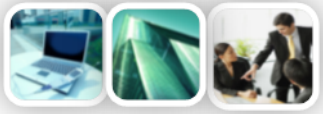
Normali zation

Stemmi ng

TF-IDF

LSA

LDA



토픽 모델링

- TF-IDF
- LSA
- pLSA
- LDA

+1 로그 취하는 이유

확률분포 추정

- 각 문서는 토픽의 분포로 구성되고 각 토픽은 단어의 분포로 구성된다.

d1 : Chicago Chocolate. Retro candies made with love.
d2 : Chocolate sweets and candies. Collection with mini love hearts.
d3 : Retro sweets from Chicago for chocolate lovers.

$$X = \begin{array}{c} \\ \text{chicago} \\ \text{chocolate} \\ \text{retro} \\ \text{candy} \\ \text{made} \\ \text{love} \\ \text{sweet} \\ \text{collection} \\ \text{mini} \\ \text{heart} \end{array} \begin{array}{c} \begin{array}{ccc} d_1 & d_2 & d_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{array}$$

$$U_2 = \begin{array}{c} \\ \text{chicago} \\ \text{chocolate} \\ \text{retro} \\ \text{candy} \\ \text{made} \\ \text{love} \\ \text{sweet} \\ \text{collection} \\ \text{mini} \\ \text{heart} \end{array} \begin{array}{c} \begin{array}{cc} \text{concept 1} & \text{concept 2} \end{array} \\ \begin{bmatrix} -0.318 & \mathbf{0.424} \\ \mathbf{-0.486} & 0.018 \\ -0.318 & \mathbf{0.424} \\ -0.333 & -0.148 \\ -0.166 & 0.257 \\ \mathbf{-0.488} & 0.018 \\ -0.320 & -0.239 \\ -0.168 & -0.406 \\ -0.168 & -0.406 \\ -0.168 & -0.406 \end{bmatrix} \end{array}$$

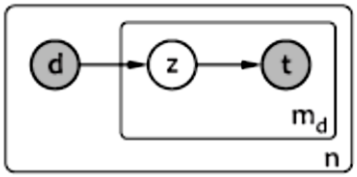
$$\Sigma_2 = \begin{bmatrix} 3.562 & 0 \\ 0 & 1.966 \end{bmatrix}$$

$$V_2 = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array} \begin{array}{c} \begin{array}{cc} \text{concept 1} & \text{concept 2} \end{array} \\ \begin{bmatrix} -0.592 & 0.505 \\ -0.598 & -0.798 \\ -0.541 & 0.329 \end{bmatrix} \end{array}$$

$q_{chicago} = [\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]$
$q_{candy} = [\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ ]$

Query	d_1	d_2	d_3
Chicago	0.891	-0.510	0.806
Candy	0.183	0.969	0.338

- 문서는 어느 정도의 확률적 프로세스를 따라 단계적인 용어의 조합으로 생성된다.



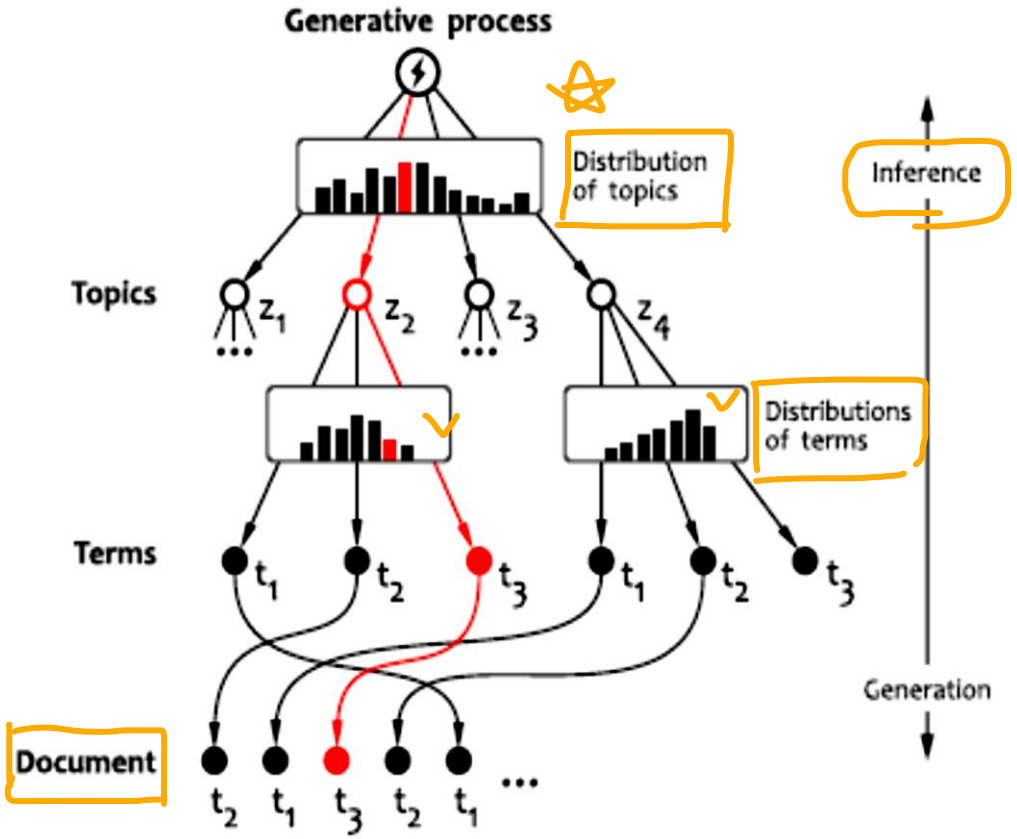
Katsov, 2017

LDA

- **Latent Dirichlet Allocation** (Blei, Ng, Jordan, 2003)

"워드 임베딩"

- TF-IDF
- TM
- Deep 러닝



생성 과정 반대 ⇒ 추론

Katsov, 2017

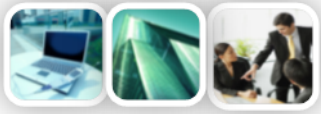
Intuition

문서	단어1	단어2
문서1	사과	포도
문서2	사과	바나나
문서3	바나나	사과
문서4	고양이	강아지
문서5	강아지	병아리
문서6	강아지	병아리

문서7	사과	강아지

단어	토픽1	토픽2
사과	50%	0%
포도	17%	0%
바나나	33%	0%
고양이	0%	17%
강아지	0%	50%
병아리	0%	33%
문서	토픽1	토픽2
문서1	100%	0%
문서2	100%	0%
문서3	100%	0%
문서4	0%	100%
문서5	0%	100%
문서6	0%	100%

문서	토픽1	토픽2
문서7	50%	50%



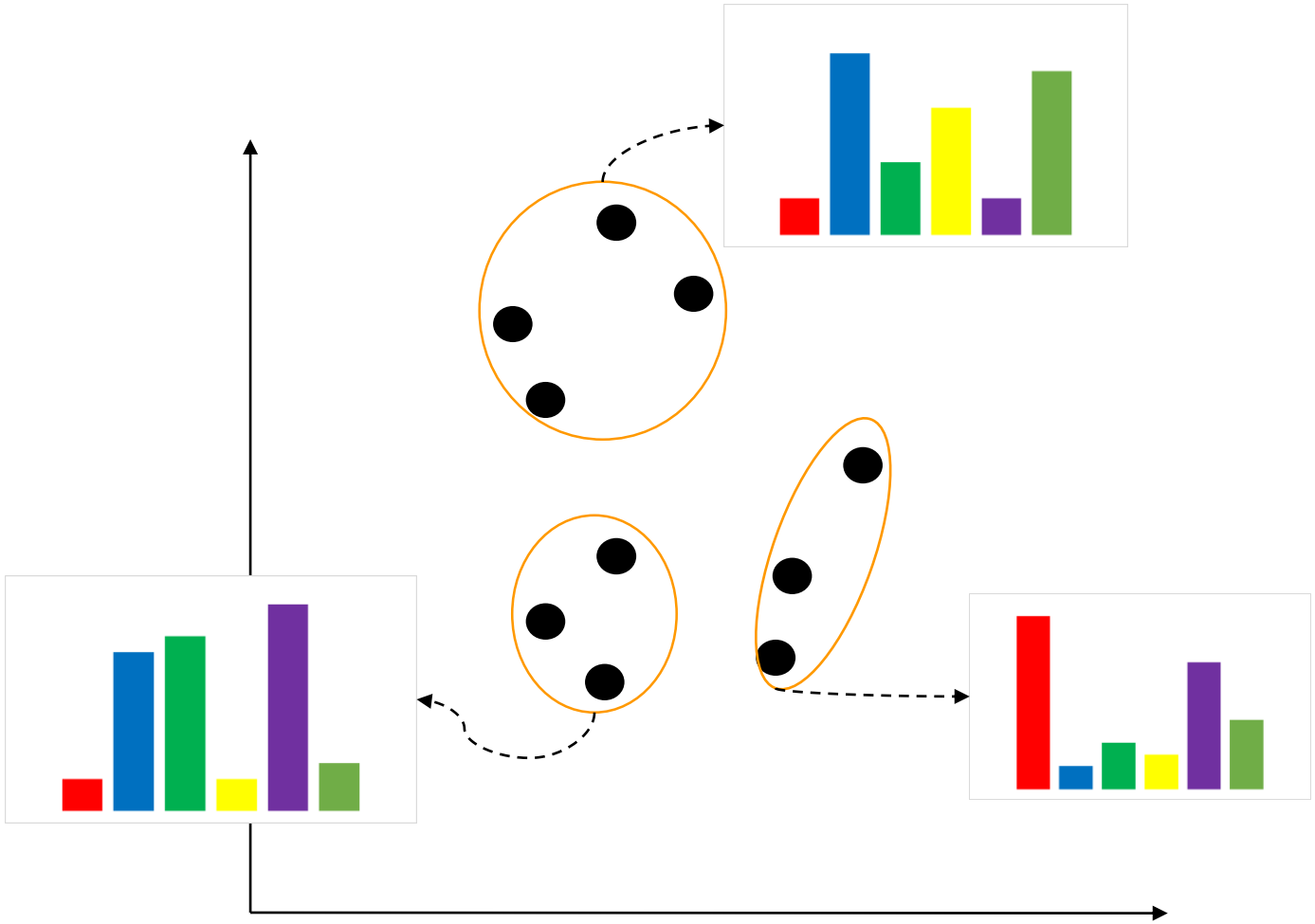
Example

"美 마이크론, 화웨이에 메모리 반도체 공급중단...삼성·SK 하이닉스 반사이익"

미국의 메모리 반도체 회사 '마이크론'이 도널드 트럼프 행정부의 수출 제한 조치에 따라 중국 화웨이에 D램 등 부품 공급을 중단한 것으로 전해졌다. 마이크론은 D램과 낸드플래시 분야에서 각각 세계 3위와 4위를 차지하고 있다. 미국 매사추세츠공과대학(MIT) 소유 정보기술(IT) 전문지 'MIT 테크놀로지 리뷰'의 중문판인 '딥 테크'는 30일 "마이크론이 전날(29일) 우리에게 화웨이에 대한 부품 공급을 잠정적으로 중단한다고 밝혔다"며 "이 회사가 화웨이 사태와 관련해 입장을 밝힌 것은 이번이 처음"이라고 보도했다. 마이크론의 전체 매출에서 대(對)화웨이 매출이 차지하는 비중은 약 13%에 이르는 것으로 알려졌다. 마이크론이 빠져나가면서 화웨이는 스마트폰과 서버 등에 필수적인 메모리 반도체 공급에 차질을 빚을 것으로 예상된다. 중국 정부는 미국의 화웨이 봉쇄로 각국 기업들이 등을 돌리자 자국 기업들에 자력갱생을 강조하고 있다. 중국 공업신식화부의 왕즈쥔(王志軍) 부부장(차관)은 최근 관영 매체와의 인터뷰에서 "올 하반기 64단 3D 낸드플래시 메모리 양산이 예정돼 있다"고 밝히기도 했다. 중국에서 유일하게 낸드플래시 양산에 나서고 있는 국유기업 '창장메모리(YMTC)' 소속 최고기술담당임원(CTO)이 올 3월 밝힌 내용을 중국 정부 고위관계자가 처음 확인한 것이다. YMTC는 중국 국유기업 '칭화유니'가 우한의 국유기업 'XMC'를 인수해 2016년 설립한 메모리 반도체 회사다. 시진핑 중국 국가주석은 지난해 4월 XMC 공장을 시찰해 "반도체 기술에서 중대 돌파구를 서둘러 마련해 세계 메모리 반도체 기술의 높은 봉우리에 올라야 한다"고 강조한 바 있다. 한편 한국의 삼성전자와 SK하이닉스는 이에 따른 반사이익을 볼 것으로 전망된다. 시장조사기관 '트렌드포스'에 따르면, 세계 3대 반도체 메모리 회사는 마이크론과 삼성전자, SK하이닉스다. 이 가운데 삼성전자와 SK하이닉스의 시장 점유율은 각각 42.7%와 29.9%다. 마이크론은 23%다.

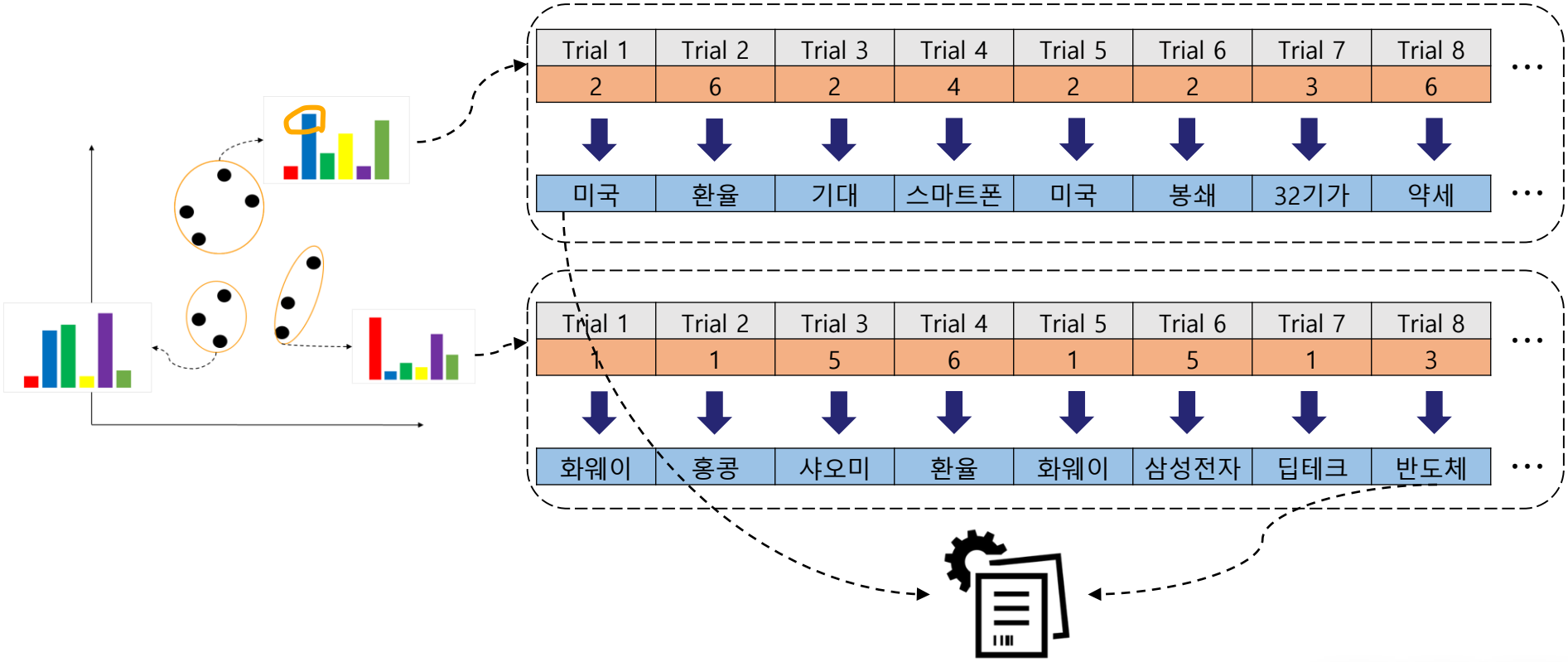
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
중국	수출	메모리	매출	반도체기업	외환
딥테크	봉쇄	D램	D램	마이크론	환율
화웨이	시장점유율	낸드플래시	스마트폰	모토로라	환차익
시진핑	기대	64단	노트북	샤오미	약세
홍콩	미국	32기가	스마트워치	삼성전자	달러
우한	공급	반도체	반사이익	SK하이닉스	유로화

문서 군집의 중심토픽 분포 도출

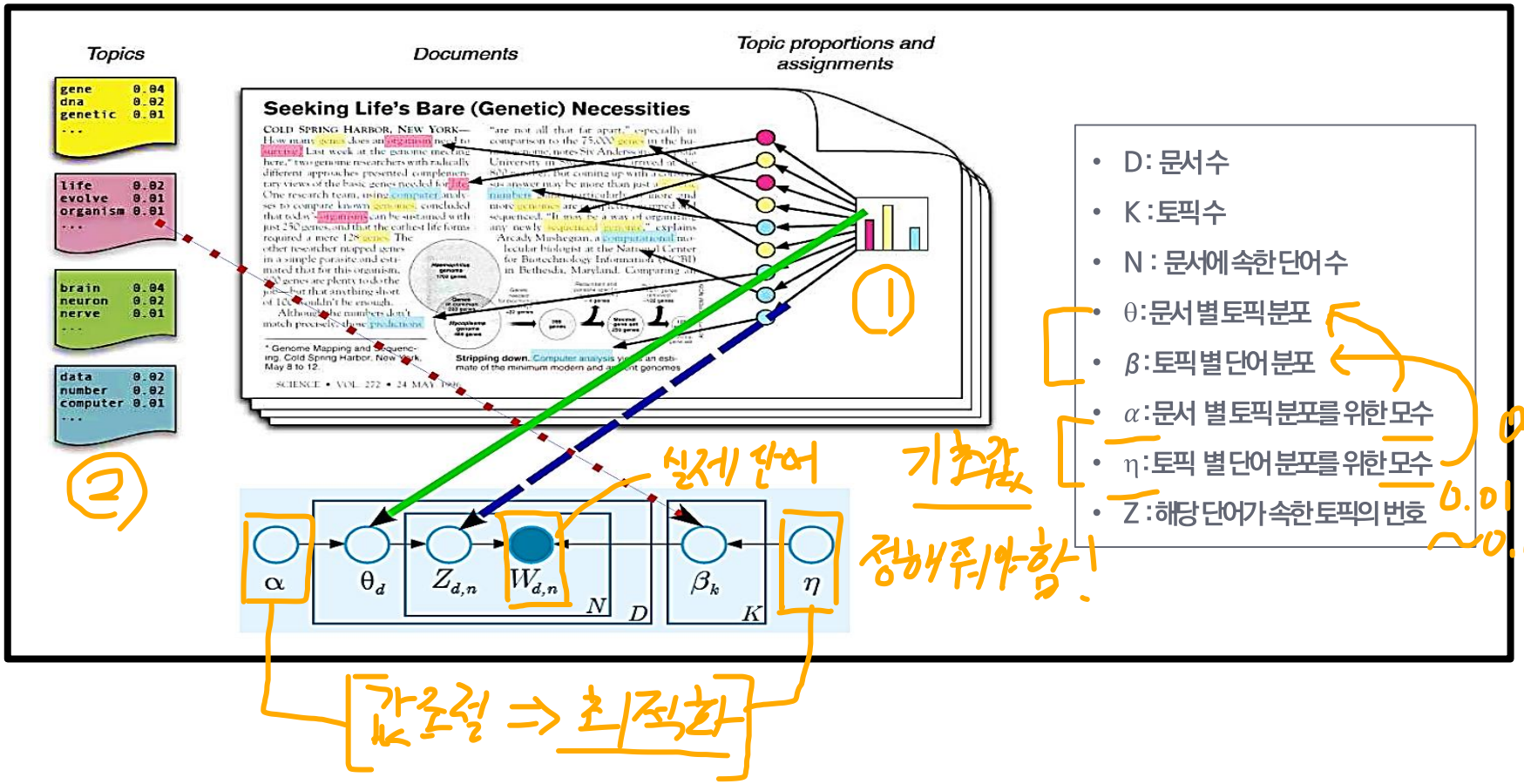


문서생성과정

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
중국	수출	메모리	매출	반도체기업	외환
딥테크	봉쇄	D램	D램	마이크론	환율
화웨이	시장점유율	낸드플래시	스마트폰	모토로라	환차익
시진핑	기대	64단	노트북	샤오미	약세
홍콩	미국	32기가	스마트워치	삼성전자	달러
우한	공급	반도체	반사이익	SK하이닉스	유로화



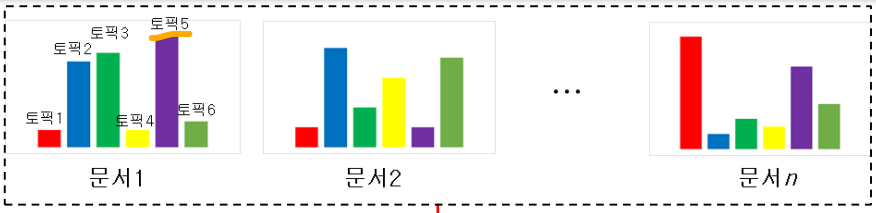
LDA 모델 (Blei, 2012)

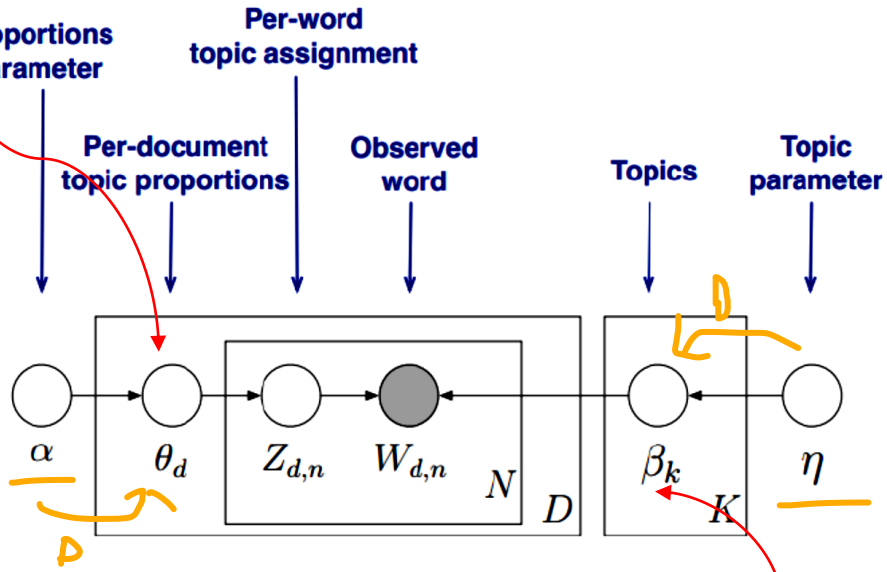


- D: 문서수
- K: 토픽수
- N: 문서에 속한 단어수
- θ: 문서별 토픽분포
- β: 토픽별 단어분포
- α: 문서 별 토픽 분포를 위한 모수
- η: 토픽 별 단어분포를 위한 모수
- Z: 해당 단어가 속한 토픽의 번호

0.1
0.01 ~ 0.001

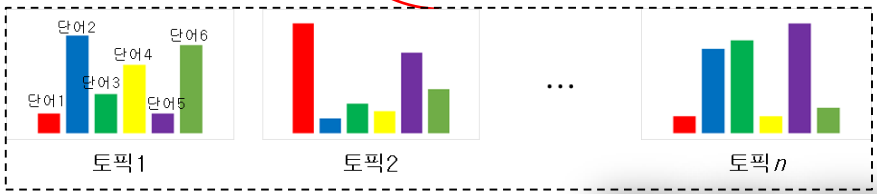
LDA Process – Document Generation





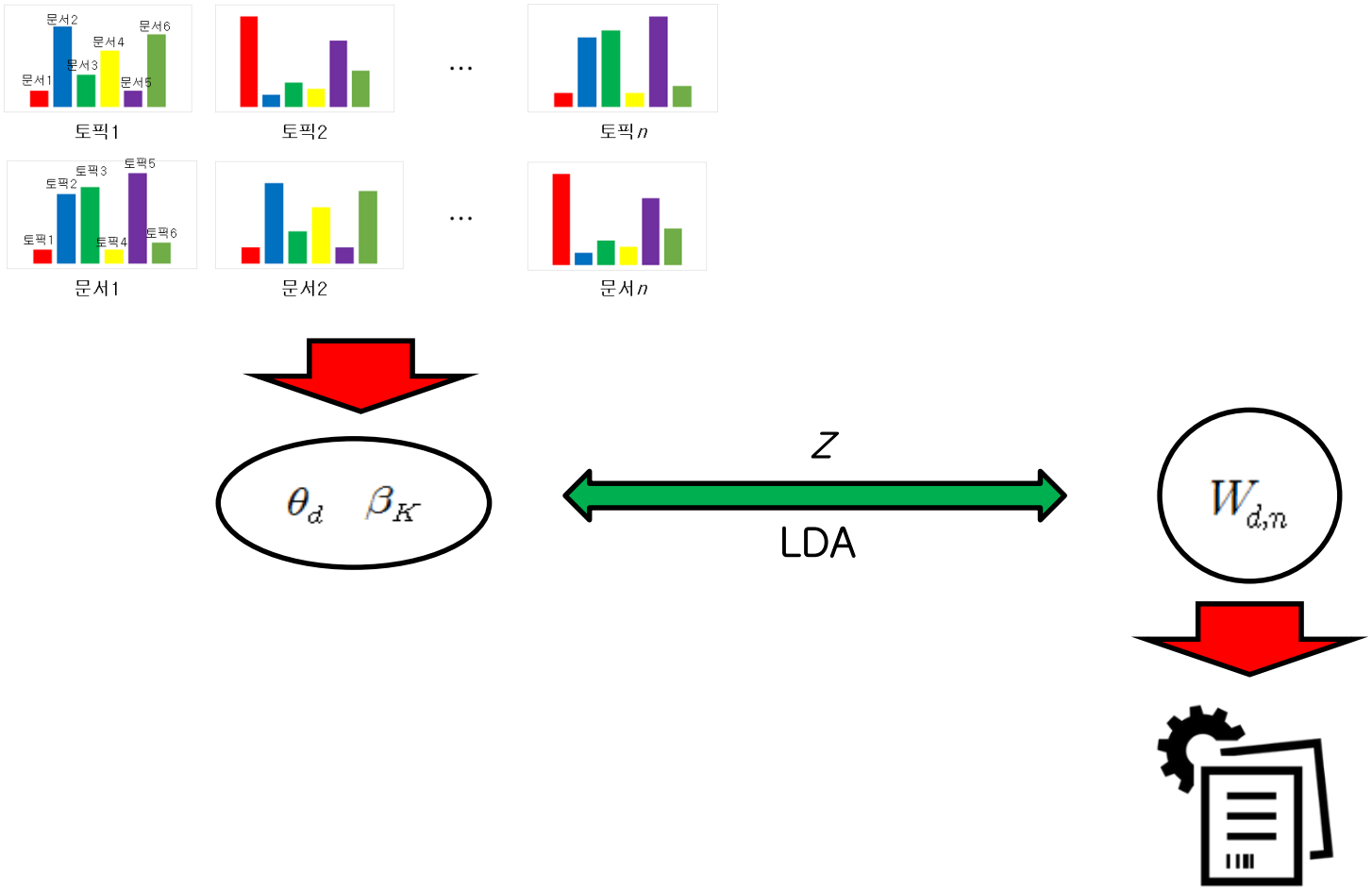
LDA 수식

$$p(\beta,\theta,\mathbf{z},\mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i|\eta)\right)\left(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K},z_{d,n})\right)$$



https://user.eng.umd.edu/~smiran/LDA.pdf

LDA Process – Inference



LDA Process – Inference - Example

- 토픽의 개수를 정한다
- 모든 문서의 모든 단어를 각 토픽이 임의 할당한다.
- 모든 단어의 토픽은 정확하게 할당되었지만 특정 문서 내 특정 단어 하나는 잘못 할당되었다고 가정한 후 그 문서 내 단어들의 토픽 할당비율과 그 단어가 모든 문서들에서 할당된 토픽의 비율을 고려하여 토픽을 재할당한다.

  - 문서 1: 탕수육을 여의도 지하철역 근처에서 먹고 커피와 케이크도 먹었다.
  - 문서 2: 충무로에는 짜장면과 탕수육을 잘 하는 식당이 있고 탕수육을 잘 하는 식당들은 지하철역 근처에 많다.

문서1

단어	탕수육	여의도	지하철역	커피	케이크
토픽	?	1	1	2	2

문서2

단어	충무로	짜장면	탕수육	탕수육	지하철역
토픽	1	2	2	2	2

- 모든 단어에 대해 수렴할 때까지 반복한다.