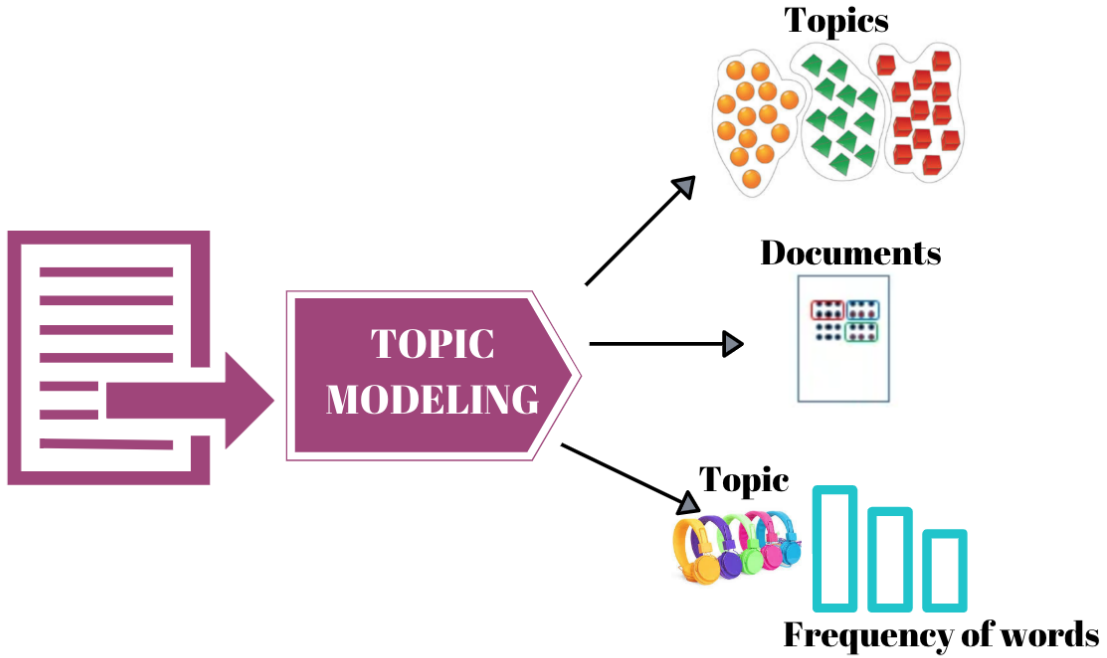
KOREAN EMBEDDING STUDY

- Latent Dirichlet Allocation -

20200310

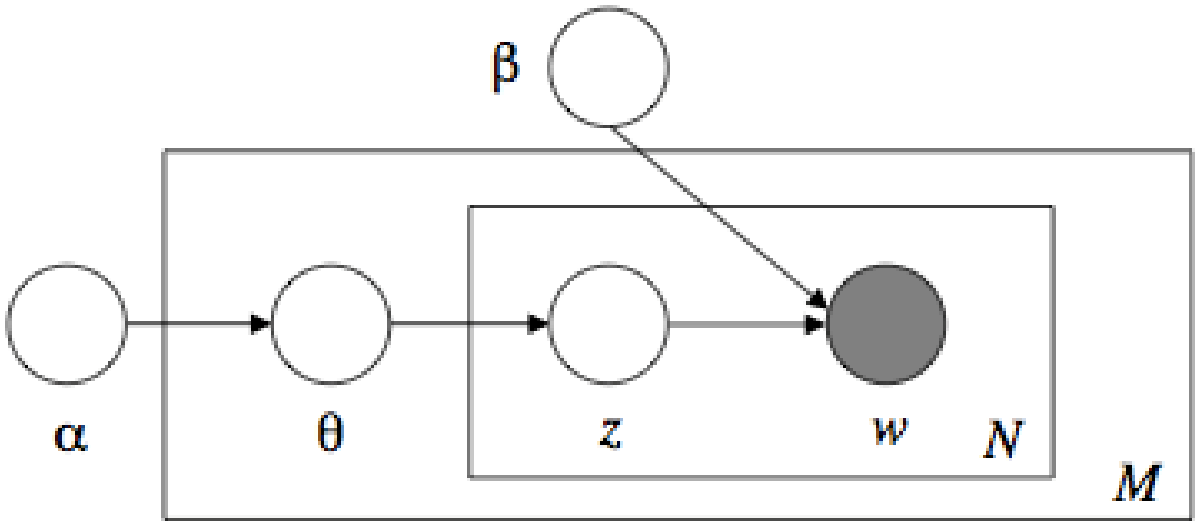
Kwang-June Choi



토픽 모델링(Topic Modeling)

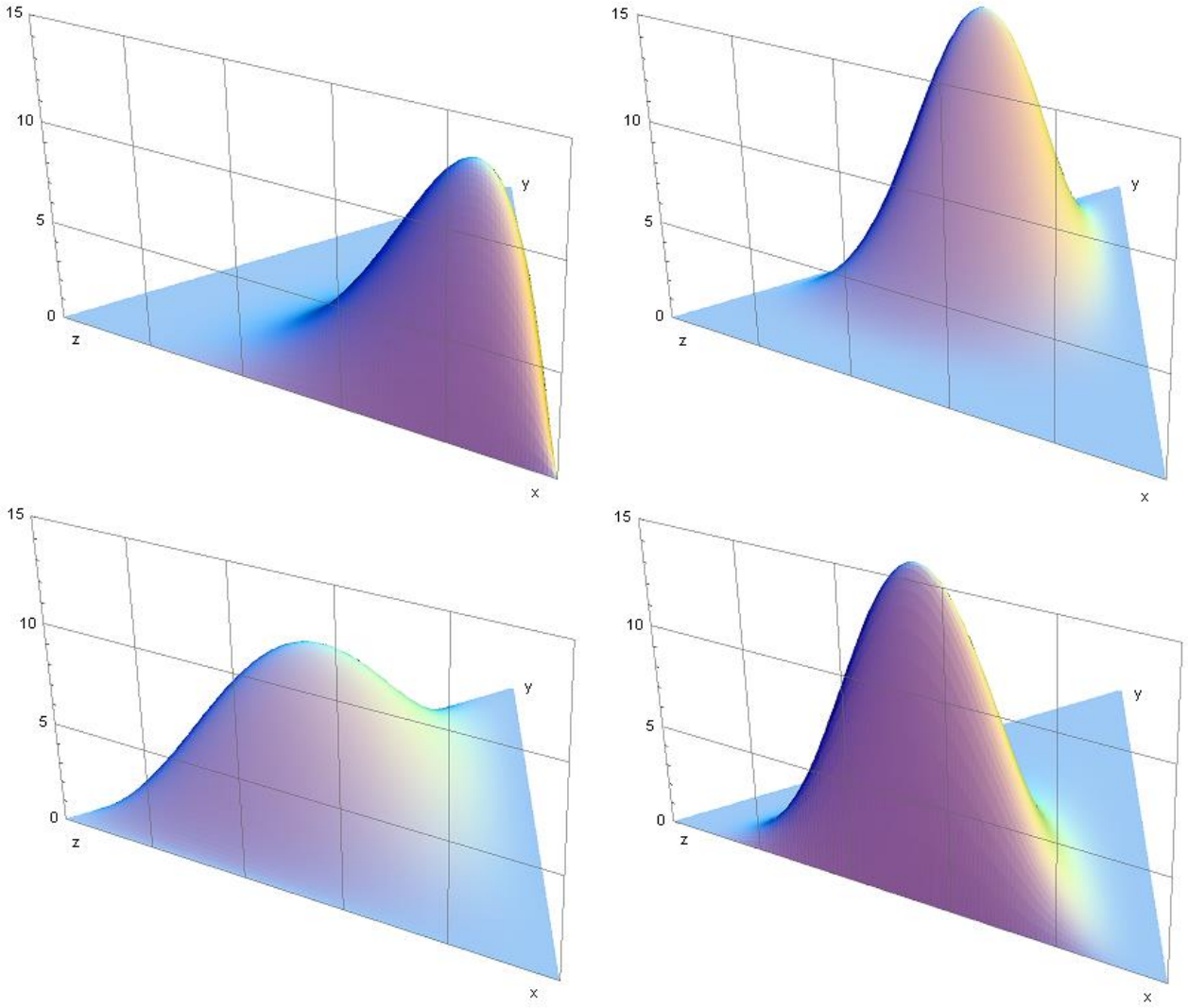
기계 학습 및 자연어 처리 분야에서 토픽이라는 문서 집합의 추상적인 주제를 발견하기 위한 통계적 모델

텍스트 본문의 숨겨진 의미 구조를 발견하기 위해 사용되는 텍스트 마이닝 기법



잠재 디리클레 할당 LDA, Latent Dirichlet Allocation

- 주어진 문서에 대하여 각 문서에 어떤 토픽(또는 주제)들이 존재하는지에 대한 확률 모형



디리클레 분포

- K차원의 실수 벡터 중 벡터의 요소가 양수이며 모든 요소를 더한 값이 1인 경우에 확률이 정의되는 연속확률분포

- 2이상의 자연수 k와 상수 α1,…. ,αk 에 대하여 디리클레분포의 확률밀도함수는

$x_1, \ldots, x_k$가 모두 양의 실수이며 $\sum_{i=1}^{k} x_i = 1$을 만족할 때,

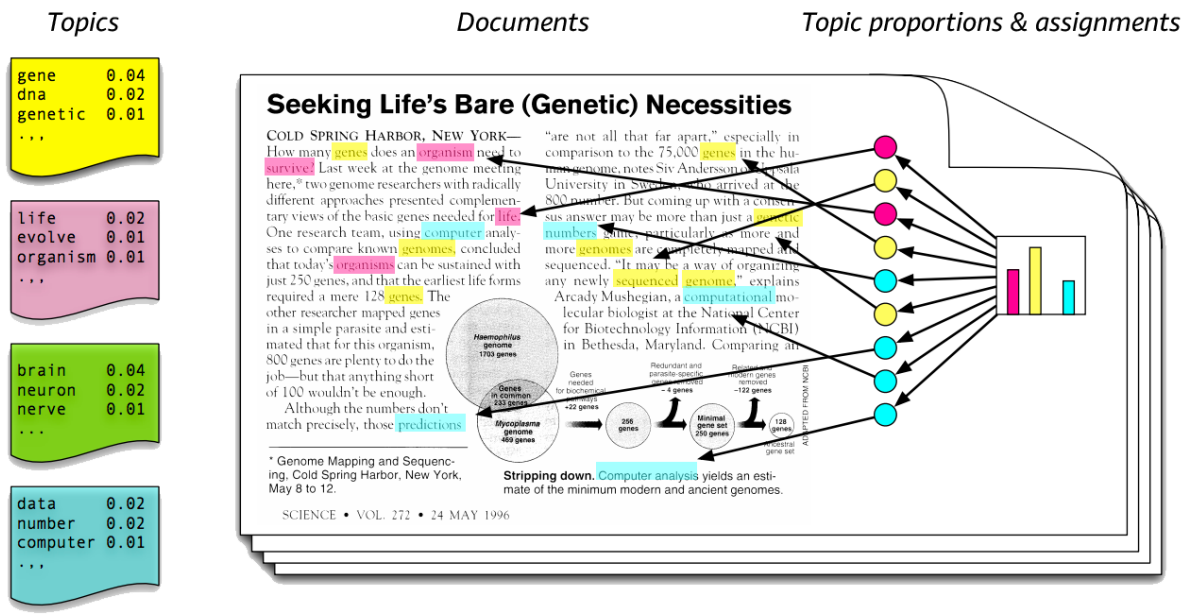
$$f(x_1, \ldots x_k; \alpha_1, \ldots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} {x_i}^{\alpha_i - 1}$$

그 외의 경우는 0이다.

로 정의 됨

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}$$

잠재 디리클레 할당 LDA, Latent Dirichlet Allocation

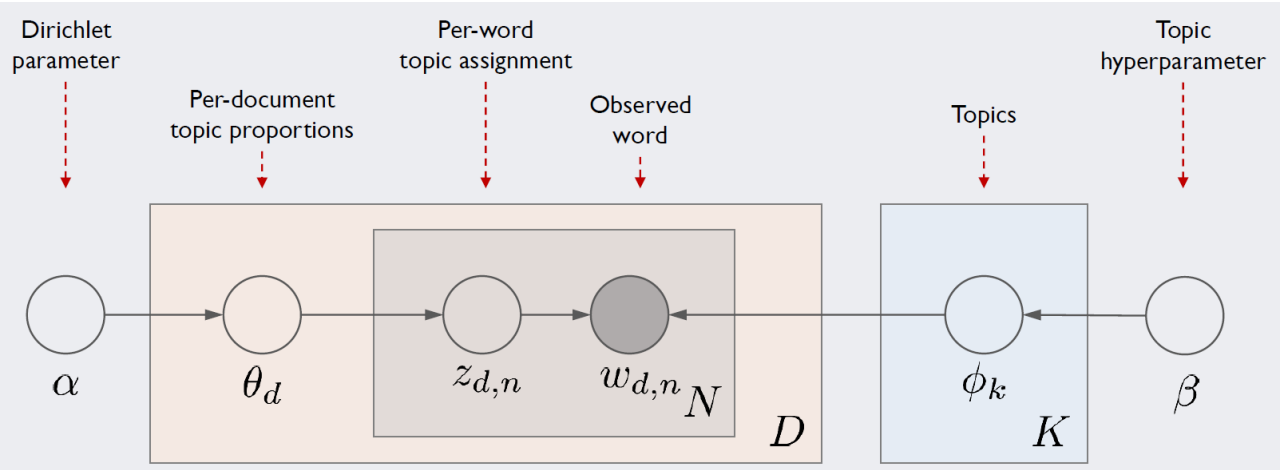


- LDA는 토픽 별 단어의 분포, 문서 별 토픽의 분포를 모두 추정함
- 토픽에 해당하는 단어가 뽑힌다는 것이 LDA가 가정하는 문서 생성 과정
- 학습을 통해 말뭉치 이면에 존재하는 잠재정보를 파악함

https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/

ϕ_k

LDA Architecture



Terms	Topic 1	Topic 2	Topic 3
Baseball	0.000	0.000	0.200
Basketball	0.000	0.000	0.267
Boxing	0.000	0.000	0.133
Money	0.231	0.313	0.400
Interest	0.000	0.312	0.000
Rate	0.000	0.312	0.000
Democrat	0.269	0.000	0.000
Republican	0.115	0.000	0.000
Cocus	0.192	0.000	0.000
President	0.192	0.063	0.000

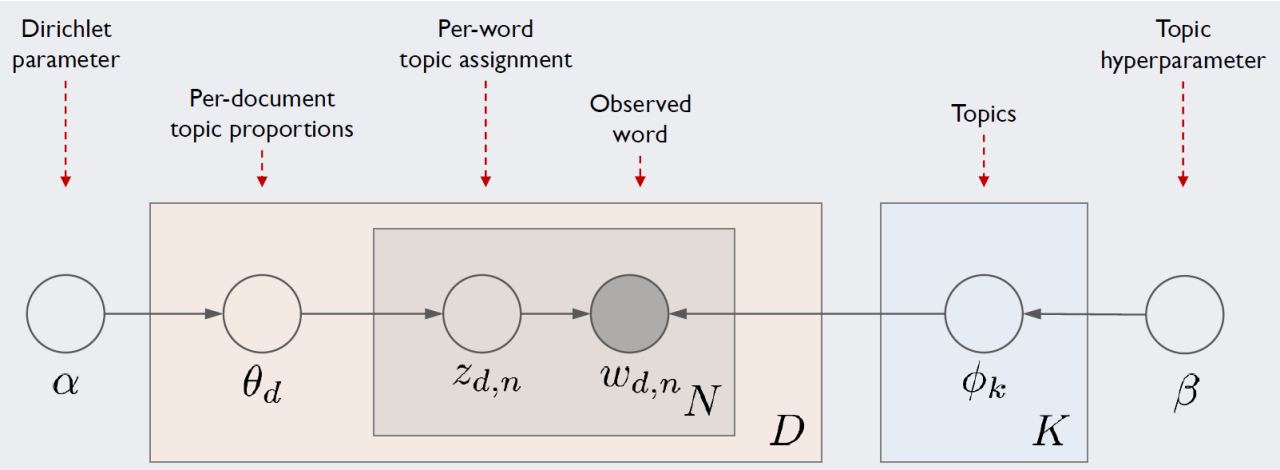
표기	값
D	문서 총 갯수
K	토픽의 총 갯수
N	d번째 문서의 총 단어 갯수
Θ	문서당 토픽 분포, $\Theta_d \sim Dir(\alpha)$ for $d \in \{1, 2, \ldots, D\}$
Φ	토픽당 단어 분포, $\Phi_k \sim Dir(\beta)$ for $k \in \{1, 2, \ldots, K\}$
$z_{d,n}$	해당 단어의 토픽 분포, $z_{d,n} \sim Multi(\Theta_d)$
$w_{d,n}$	d번째 문서의 n번째 단어, $w_{d,n} \sim Multi(\Phi_{z_{d,n},n})$

https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/

ϕ_k

LDA Architecture

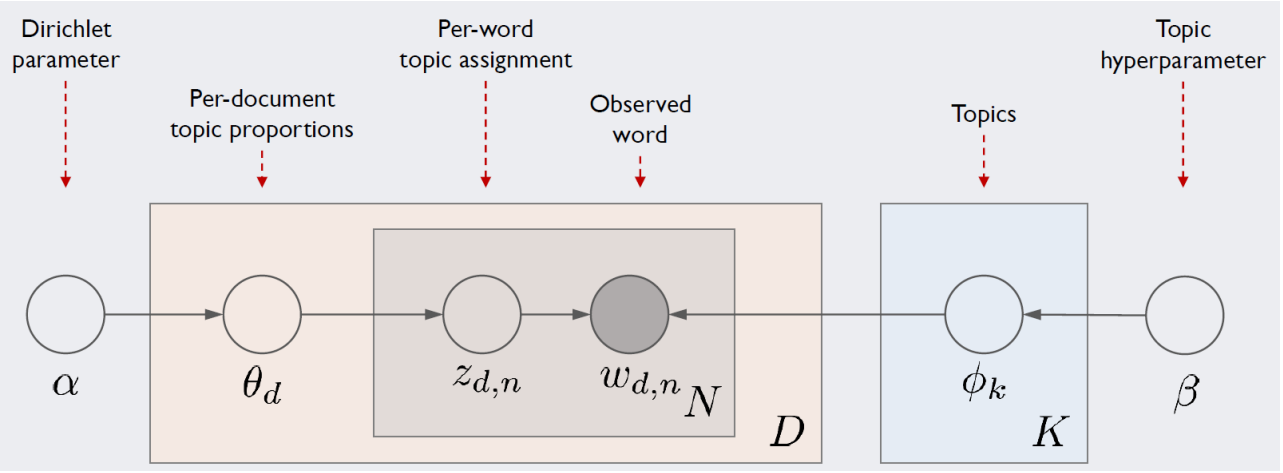
Docs	Topic 1	Topic 2	Topic 3
Doc 1	0.400	0.000	0.600
Doc 2	0.000	0.600	0.400
Doc 3	0.375	0.625	0.000
Doc 4	0.000	0.375	0.625
Doc 5	0.500	0.000	0.500
Doc 6	0.500	0.500	0.000



표기	값
D	문서 총 갯수
K	토픽의 총 갯수
N	d번째 문서의 총 단어 갯수
Θ	문서당 토픽 분포, $\Theta_d \sim Dir(\alpha)$ for $d \in \{1, 2, \ldots, D\}$
Φ	토픽당 단어 분포, $\Phi_k \sim Dir(\beta)$ for $k \in \{1, 2, \ldots, K\}$
$z_{d,n}$	해당 단어의 토픽 분포, $z_{d,n} \sim Multi(\Theta_d)$
$w_{d,n}$	d번째 문서의 n번째 단어, $w_{d,n} \sim Multi(\Phi_{z_{d,n},n})$

https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/

LDA inference



표기	값
D	문서 총 갯수
K	토픽의 총 갯수
N	d번째 문서의 총 단어 갯수
Θ	문서당 토픽 분포, $\Theta_d \sim Dir(\alpha)$ for $d \in \{1, 2, \ldots, D\}$
Φ	토픽당 단어 분포, $\Phi_k \sim Dir(\beta)$ for $k \in \{1, 2, \ldots, K\}$
$z_{d,n}$	해당 단어의 토픽 분포, $z_{d,n} \sim Multi(\Theta_d)$
$w_{d,n}$	d번째 문서의 n번째 단어, $w_{d,n} \sim Multi(\Phi_{z_{d,n},n})$

$$p(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\phi_i|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left\{ \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\phi_{1:K}, z_{d,n}) \right\}$$

- 어떤 문서에 대한 파라미터 θ
- 앞에서부터 단어를 하나씩 채울 때 마다 θ로부터 하나의 토픽 선택
- 다시 그 토픽으로부터 단어를 선택 후 반복하는 방식으로 문서 생성 과정을 모델링

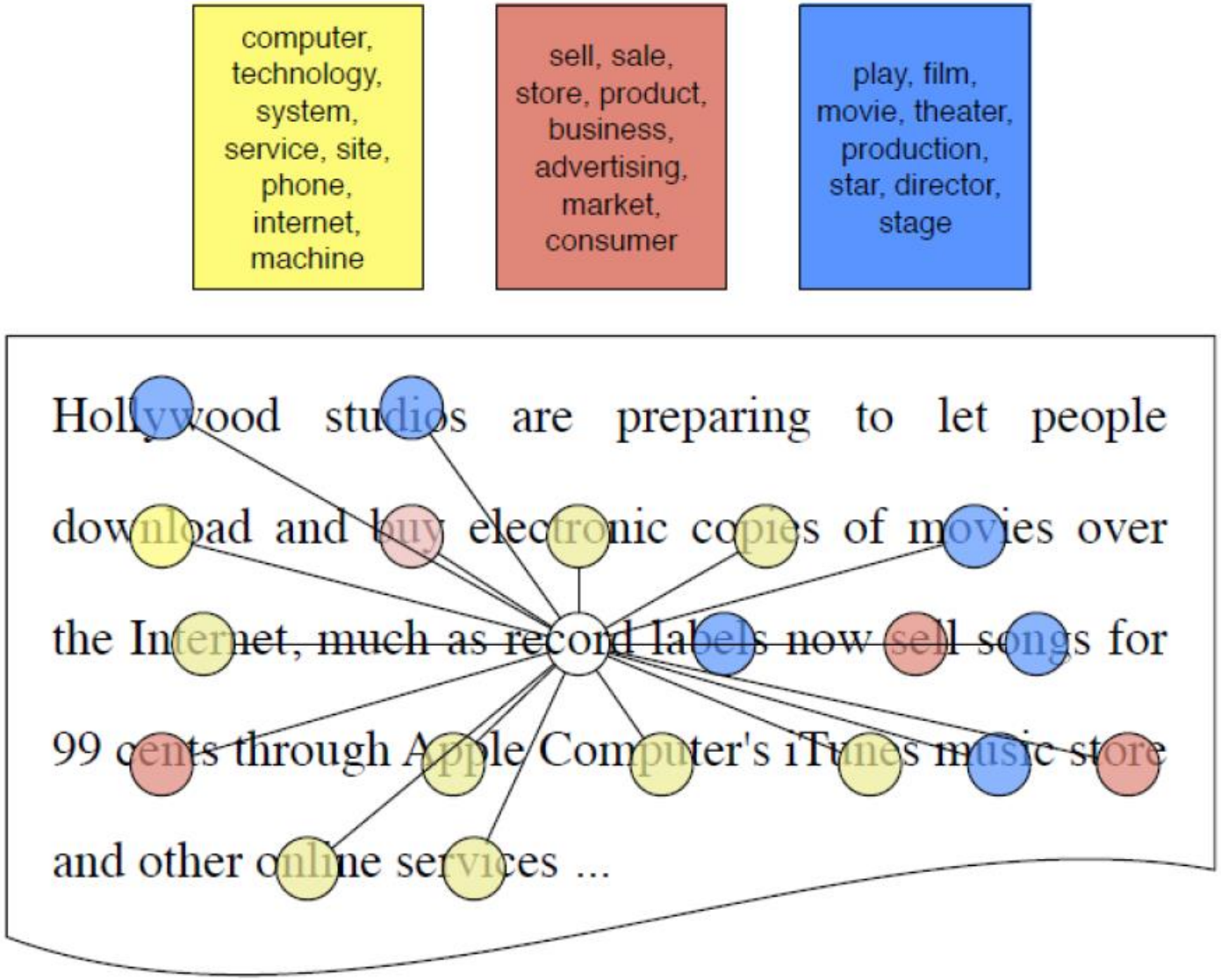
$$\underbrace{p(\theta|\text{data})}_{\text{posterior}} \propto \underbrace{\ell(\text{data}|\theta)}_{\text{likelihood}} \underbrace{\tilde{p}(\theta)}_{\text{prior}}$$

LDA gibbs sampling

$$p(z_i = j | z_{-i}, w)$$

(gibbs sampling 수식)

- w ~ 이미 알고 있는 값
- z ~ 각 단어가 어떤 토픽에 할당되어 있는지를 나타내는 변수



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

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

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Thank you