Text Analytics

Ch7 : Topic Modeling

서수원

Business Intelligence Lab.

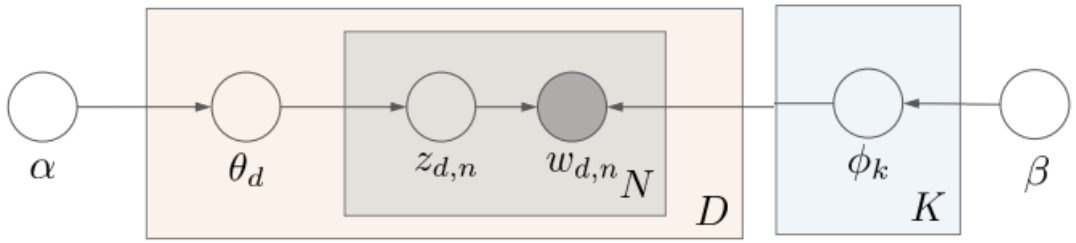
산업경영공학과, 명지대학교

01

LDA inference

LDA Inference

- LDA structure



<각 문서의 토픽 분포>

문서1 : 토픽 A 100%

문서2 : 토픽 B 100%

문서3 : 토픽 B 60%, 토픽 A 40%

<각 토픽의 단어 분포>

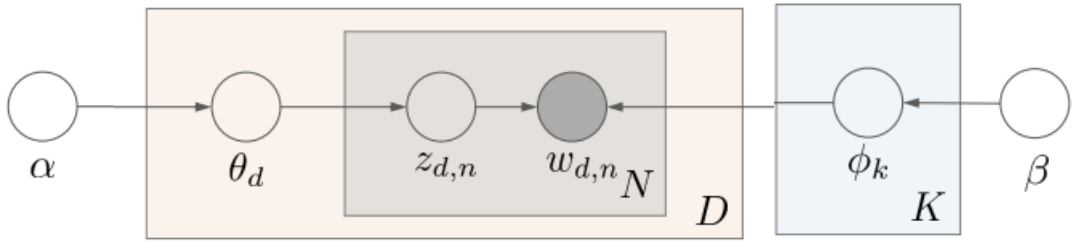
토픽A : **사과 20%, 바나나 40%, 먹어요 40%**, 귀여운 0%, 강아지 0%, 깜찍하고 0%, 좋아요 0%

토픽B : 사과 0%, 바나나 0%, 먹어요 0%, **귀여운 33%, 강아지 33%, 깜찍하고 16%, 좋아요 16%**

- 알파와 베타는 하이퍼 파라미터이다.
- 세타는 코퍼스 안에 문서이고 각 문서안에 몇 개의 토픽이 있는가를 의미한다.
  - ✓ 세타d가 정해졌다는 것은 문서안에 토픽의 분포가 정해졌다는 의미이고, 해당 문서안에 토픽의 분포를 통해 z를 구하게 된다.
- 파이는 코퍼스 안에 토픽 중 토픽별로 각 단어들이 얼마만큼 발생 하는지를 의미한다.
  - ✓ 파이k가 정해졌다는 것은 코퍼스 안에서 토픽이 정해졌다는 의미이고, 해당 토픽안에 단어의 분포를 통해 w를 구하는데 가중치를 준다.
- Z는 d번째 문서에서 n번째 단어는 어떤 토픽에서 오는지를 의미한다.
  - ✓ 이 값을 통해 문서안에서 w를 구하는데 가중치를 주게 된다.
- W는 우리가 보는 값이다.(문서안의 단어)
- D는 문서별, N은 문서안 단어별, K는 토픽별로 계속 값이 달라진다.

LDA Inference

- LDA structure

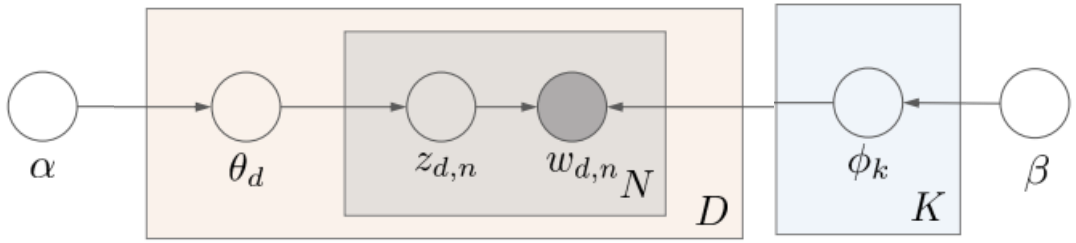


$$p(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\phi_i|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \Big(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\phi_{1:K}, z_{d,n}) \Big)$$

  - P안의 값들을 최대로 해야 한다.
    - ✓ 수식의 뒤쪽부터 의미를 풀어본다.
      - ❖ 세타가 주어졌을때 단어들이 갖는 토픽 할당 확률이 주어지고, 토픽에 대한 코퍼스 안에 단어의 분포가 주어졌을 때 단어의 등장 확률을 의미한다.
      - ❖ 알파가 주어졌을 때 문서들의 토픽 비중을 의미한다.
      - ❖ 베타가 주어졌을 때 코퍼스 안에서 토픽별로 단어들이 가진 빈도수를 의미한다.

LDA Inference

- LDA structure



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

  - 추론해야 하는 값
    - ✓ 단어가 어디 토픽에서 오는지를 추론해야 한다.(z)
    - ✓ 문서안에 토픽 비중을 추론해야 한다.(세타)
    - ✓ 코퍼스별로 토픽 안의 단어 분포를 추론해야 한다.(파이)

LDA Inference

- Inference

$$p(\phi, \theta, \mathbf{z}|\mathbf{w}) = \frac{p(\phi, \theta, \mathbf{z}, \mathbf{w})}{\int_{\phi}\int_{\theta}\sum_{\mathbf{z}} p(\phi, \theta, \mathbf{z}, \mathbf{w})}$$

  - 문서안에 단어들이 관측 되었을 때(w) 파이,세타,z에 대한 확률은 우변과 같이 나타낼 수 있다.
  - 우변의 분모는 계산이 안된다.
    - ✓ 근사를 해야 한다.
- 근사의 방법론
  - Mean filed variational methods
  - Expectation propagation
  - Collapsed Gibbs sampling
  - Collapsed variational inference
  - Online variational inference

LDA Inference

- Dirichlet Distribution
  - 이항분포와 다항분포, 베타분포에 대해 알아야 Dirichlet Distribution을 이해 할 수 있다.
- 이항분포

$$p(X = x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

  - 일반적인 주사위를 10회 던져서 숫자 6이 나오는 횟수를 센다. 이 분포는 n = 10이고 p = 1/6인 이항분포이다.
- 다항분포

$$p(x_1, ..., x_k|n, p_1, ..., p_k) = \frac{N!}{\prod_{i=1}^{k} x_i!} p_i^{x_i}, \quad \sum_i x_i = N, \quad x_i \geq 0$$

  - 주사위를 던져서 1이5번,2가3번,3이1번,4가1번,5와 6이 0번 나올 확률은

$$\frac{10!}{5!3!1!0!0!} \left(\frac{1}{6}\right)^5 \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^0$$

LDA Inference

- Dirichlet Distribution
  - 이항분포와 다항분포, 베타분포에 대해 알아야 Dirichlet Distribution을 이해 할 수 있다.
- 베타분포

$$B(\alpha,\beta)=\int_0^1 t_{\alpha-1}(1-t)^{\beta-1}dt \qquad B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$p(p|\alpha,\beta)=\frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

$$\Gamma(x)=\int_0^\infty \frac{t^{x-1}dt}{exp(t)},\quad (x>0) \qquad \Gamma(n)=(n-1)!$$

  - 분포 자체를 추정 가능하다.
    - ✓ 일반적인 동전은 앞면 뒷면이 나올 확률은 ½이다. 만약 10번 던졌을 때 앞면이 8번, 뒷면이 2번 나오면 실제 앞면 뒷면이 나올 확률은?
- 디리클레분포
  - 베타분포를 다항분포로 확장 한 것을 의미한다.

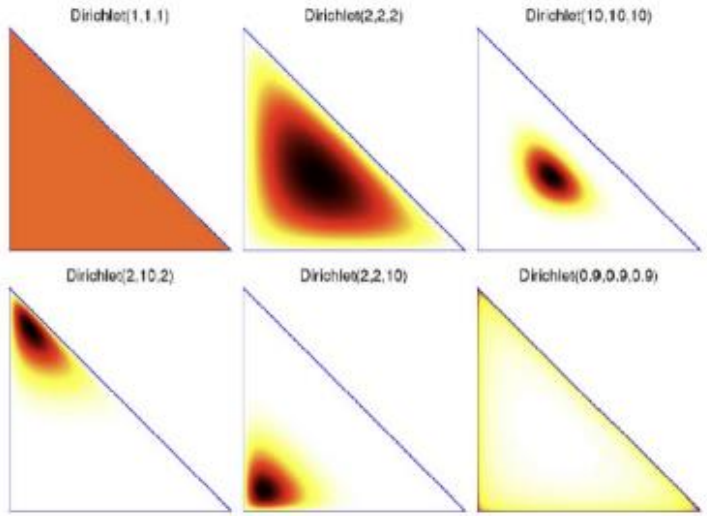
$$p(P=\{p_i\}|\alpha_i)=\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\prod_i p_i^{\alpha_i-1} \qquad \checkmark\ \textstyle\sum_i p_i=1,\ \ p_i\geq 0$$

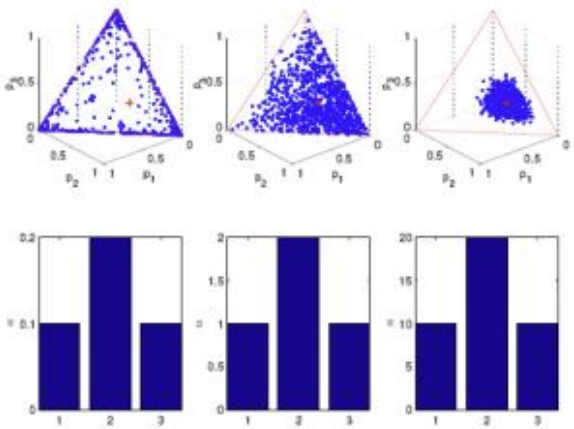
    - ✓ 베타 분포의 일반화 된 확장판이다.
  - $X_1 = P_1, X_k=P_n$ 즉 P_i는 1~k 까지 확률 값이 주어졌을 때 확률값을 의미한다.
    - ✓ 알파는 파라미터 이다.
    - ✓ N은 X_i의 모든 값을 더한 것 이다.

$$p(\{p_i\}|x_1,...,x_k)=\frac{\Gamma(N+\sum_i \alpha_i)}{\prod_i^k \Gamma(\alpha_i+x_i)}\prod_i p_i^{\alpha_i+x_i-1}$$

LDA Inference

- Dirichlet Distribution
  - 알파값은 평균에 대한 모양과 세타의 희소성을 조정한다.
    - ✓ 알파가 1이라면, 균일한 uniform분포의 형태를 띈다.
    - ✓ 알파가 2라면 값들이 평균으로 더 몰린다.
    - ✓ 알파가 10이라면 평균으로 더 모아진다.
      - ❖ 삼각주사위를 30번 던졌을 때 각 값이 10번씩 나오는 것을 예로 들 수 있다.
        - » 100번 던지면?
    - ✓ 알파가 0 보다 작으면 면에 값들이 몰리는 것을 볼 수 있다.

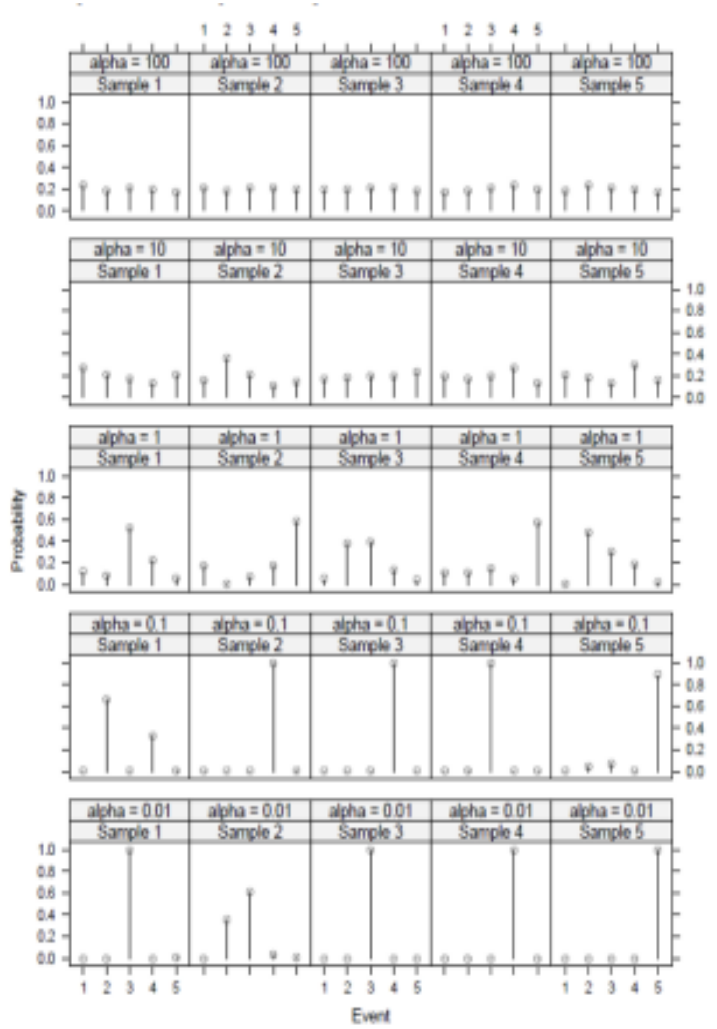




LDA Inference

- Dirichlet Distribution
  - 알파값은 평균에 대한 모양과 세타의 희소성을 조정한다.
    - ✓ 알파가 1이라면, 균일한 uniform분포의 형태를 띈다.
    - ✓ 알파가 2라면 값들이 평균으로 더 몰린다.
    - ✓ 알파가 10이라면 평균으로 더 모아진다.
      - ❖ 삼각주사위를 30번 던졌을 때 각 값이 10번씩 나오는 것을 예로 들 수 있다.
        - » 100번 던지면?
    - ✓ 알파가 0 보다 작으면 면에 값들이

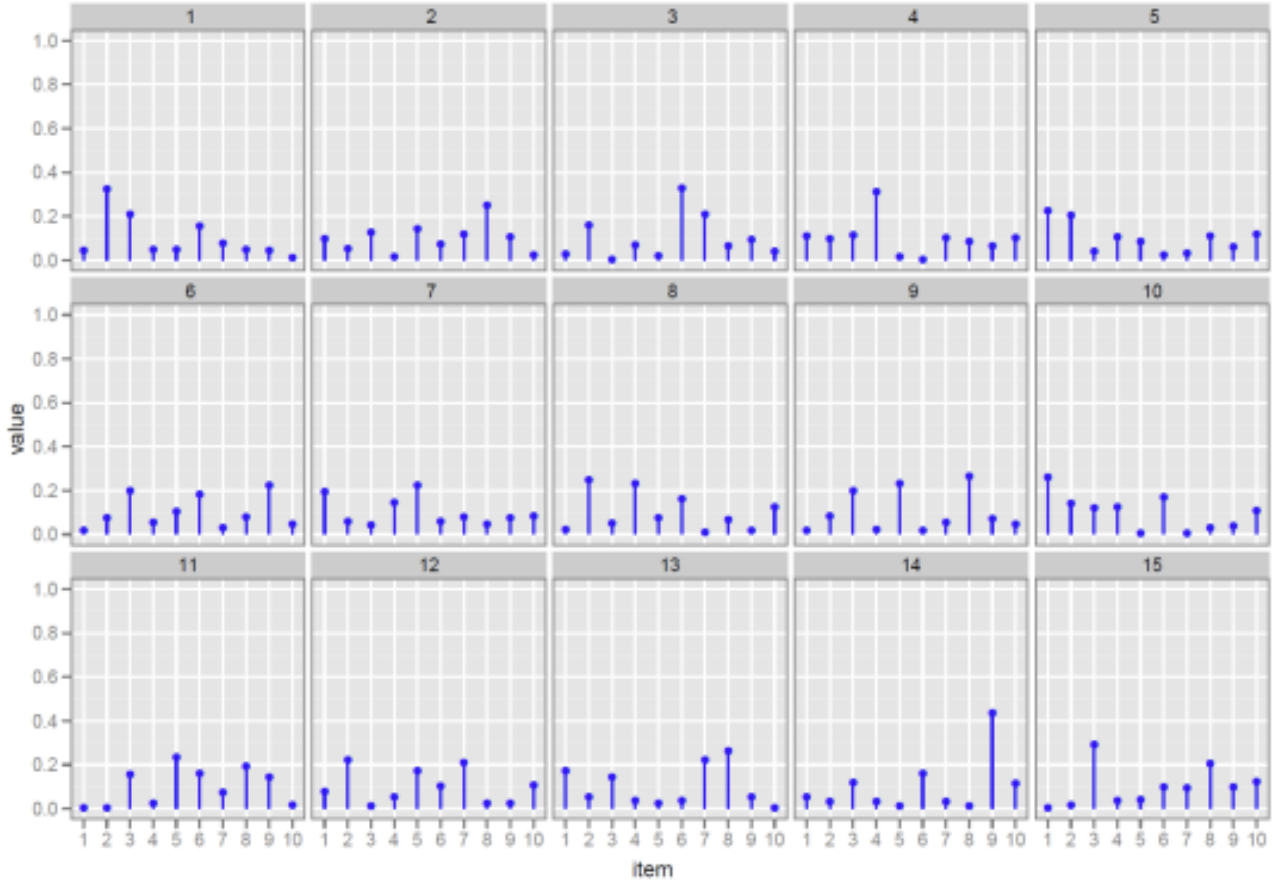
      몰리는 것을 볼 수 있다.



LDA Inference

- Dirichlet Distribution

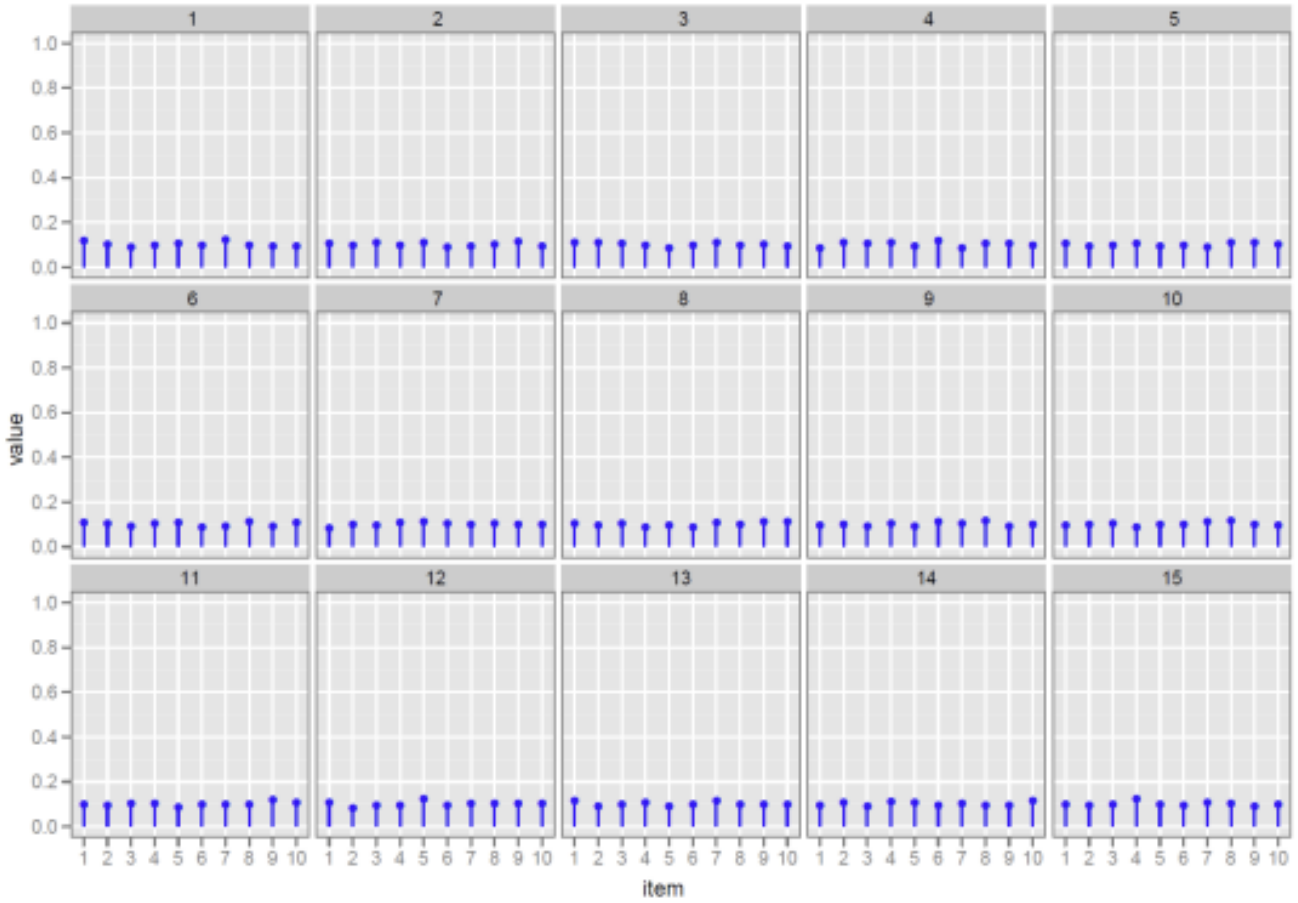
✓ $\alpha = 1$



LDA Inference

- Dirichlet Distribution

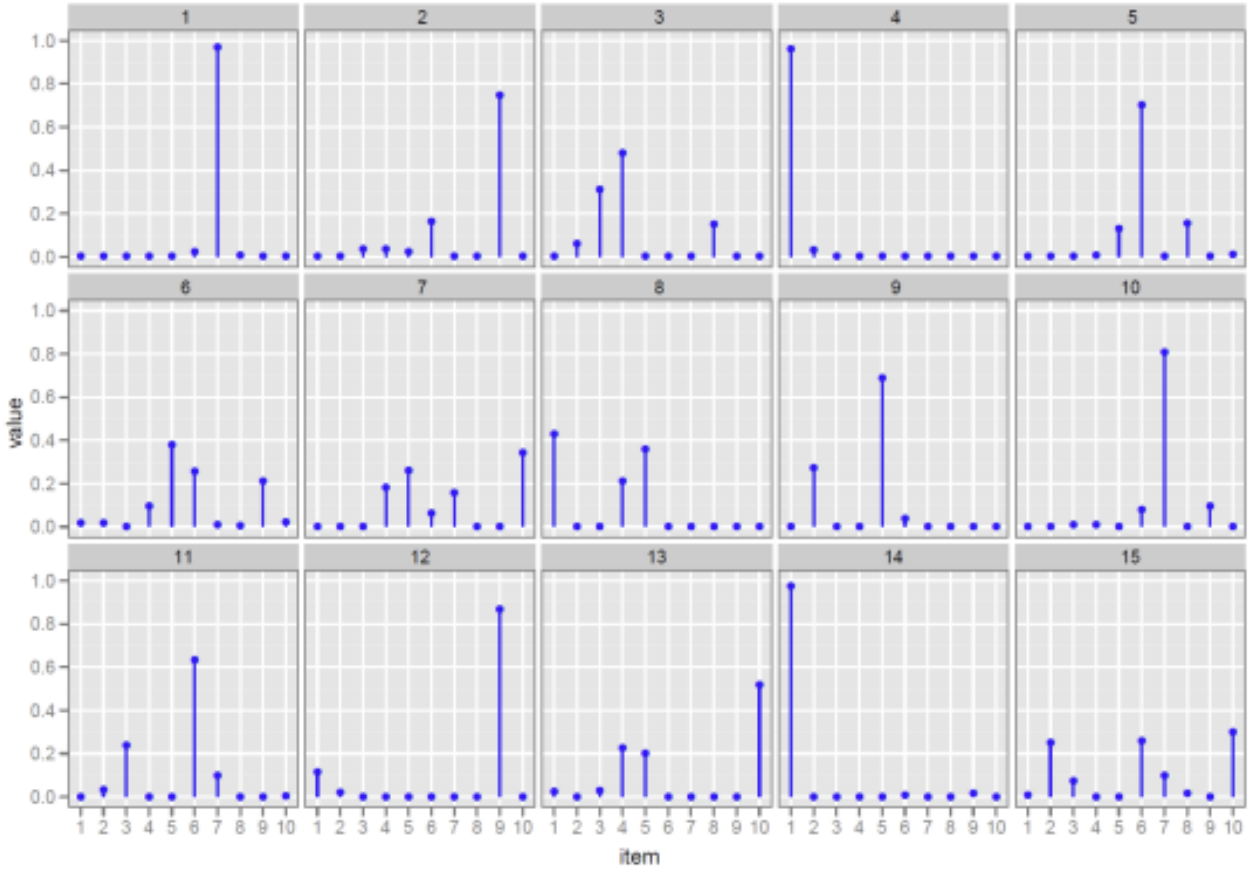
✓ $\alpha = 100$



LDA Inference

- Dirichlet Distribution

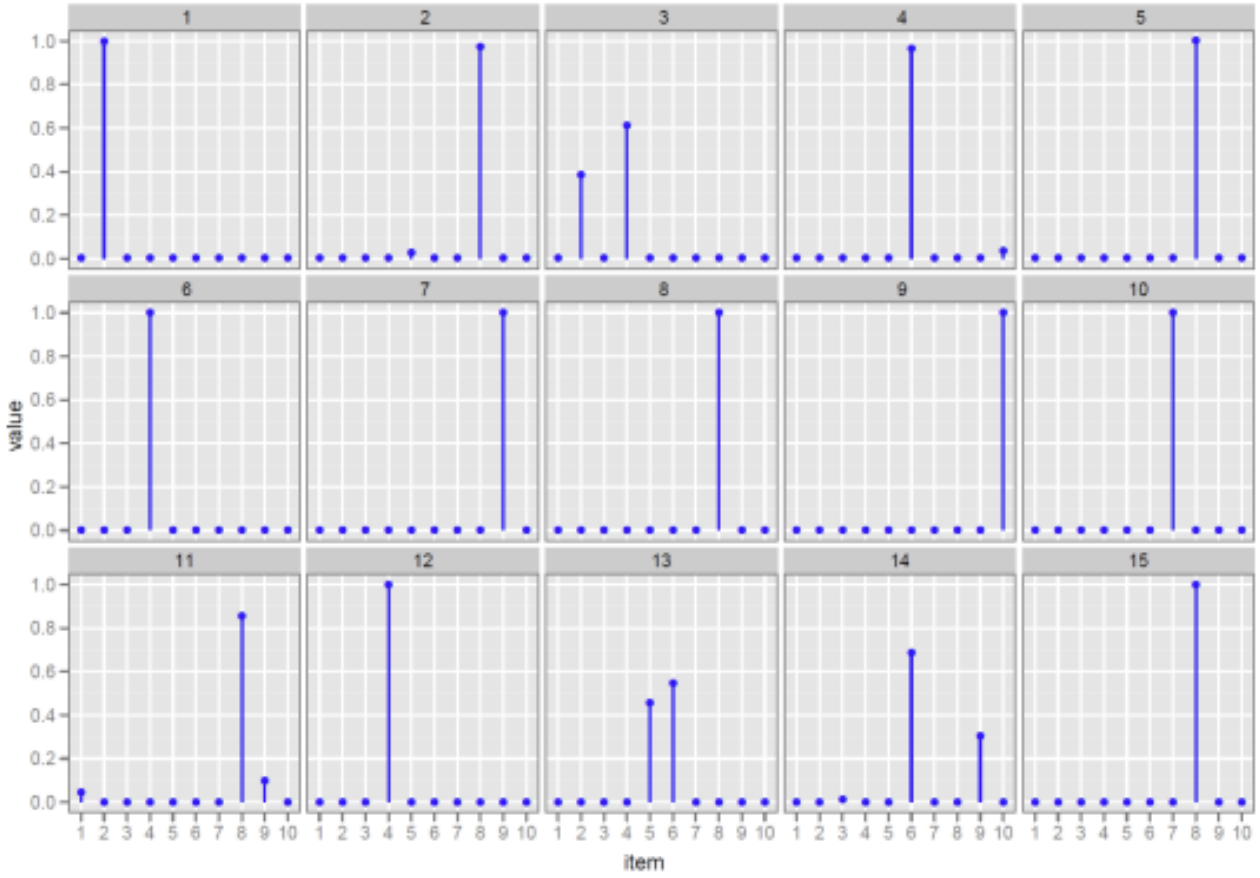
✓ $\alpha = 0.1$



LDA Inference

- Dirichlet Distribution

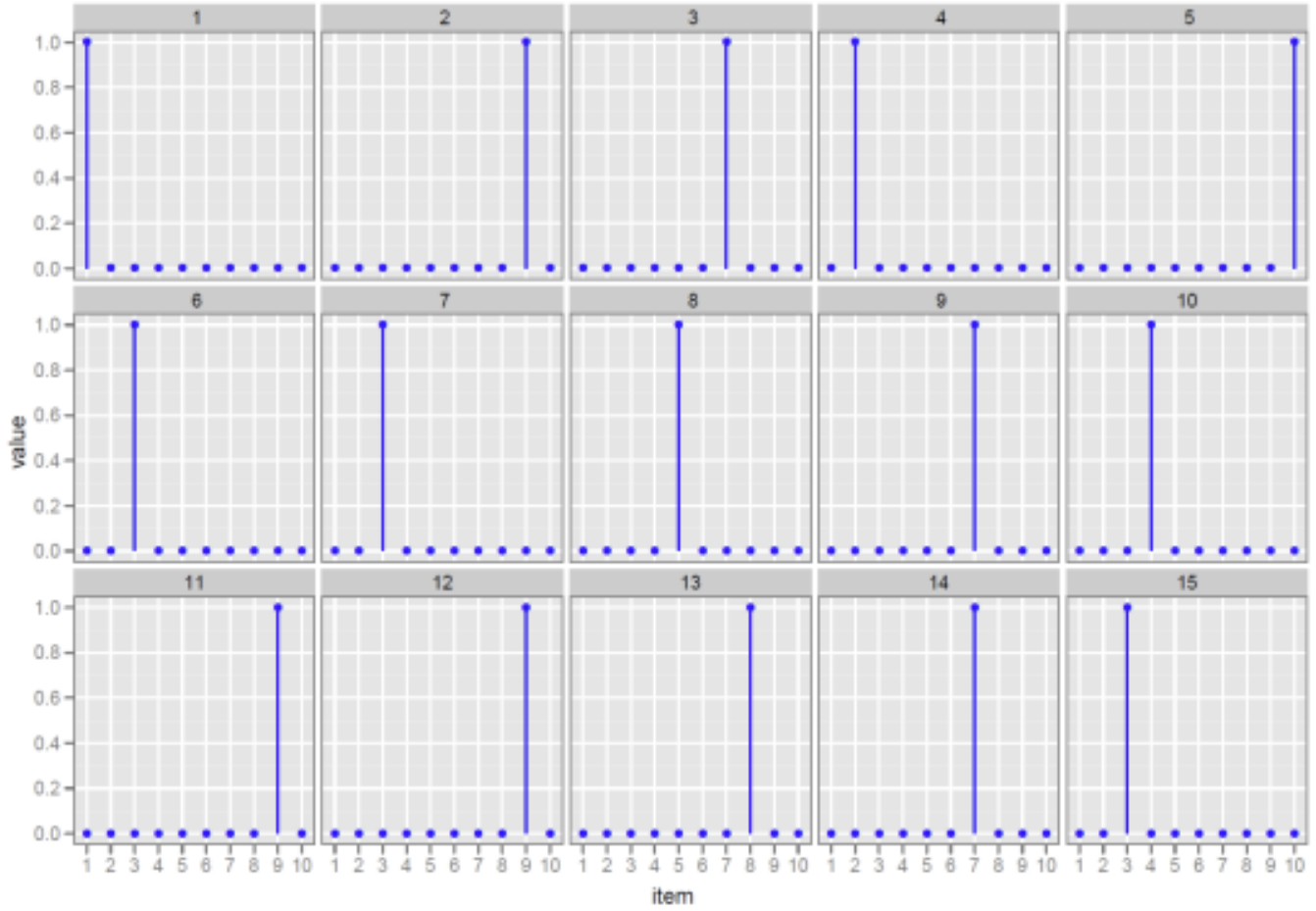
✓ $\alpha = 0.01$



LDA Inference

- Dirichlet Distribution

✓ $\alpha = 0.001$



LDA Inference

- Dirichlet Distribution
  - 알파와 베타는 정해줘야 하는 값이다.
    - ✓ 그 이후로 세타 파이 z값이 찾아진다.
  - 알파는 0보다 작게, 베타는 1이 관례적이다.

LDA Inference

- LDA Inference

$$p(\mathbf{z}, \phi, \theta | \mathbf{w}, \alpha, \beta)$$

$$p(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$$

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

  - 하이퍼 파라미터와 문서에 대한 관측치가 정해져 있다면, 단어단위 토픽(z),토픽별 단어(파이),문서별 토픽(세타)를 구할 수 있다.
  - 우변을 보면 세타와 파이와 z에 대한 값을 최대로 하는게 목표이다.

LDA Inference

- Gibbs Sampling
  - Markov Chain Monte Carlo의 하나의 형태이다.
  - Z_k제외 나머지는 다 주어졌다고 생각한다.
    - ✓ Gibbs Sampling
      - ❖ B,C고정 A그리기
      - ❖ A,C고정 B그리기
      - ❖ A,B고정 C그리기
    - ✓ Collapsed Gibbs Sampling
      - ❖ C고정 A그리기
      - ❖ A고정 C그리기
      - ❖ B는 신경 안써도 된다.

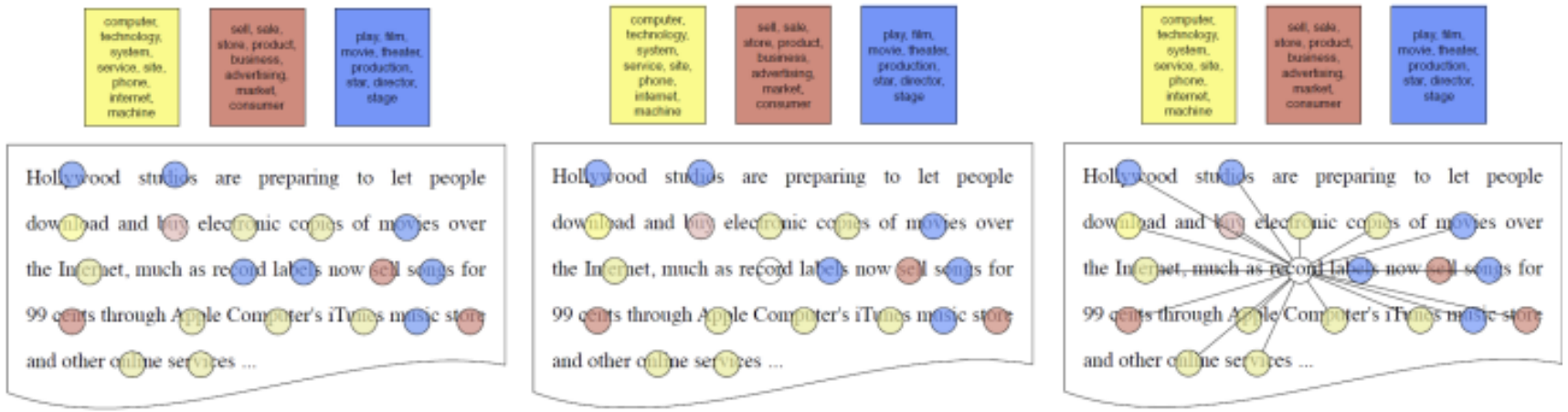
LDA Inference

- LDA Inference : Gibbs Sampling

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w})$$

  - 모든 단어에 대해서 알고, i번째 단어 빼고 모든 단어에 대해 토픽도 알 때, i가 j토픽일 확률을 구한다.
    - ✓ 예시는 토픽이 총 3개가 있다.

✓ θ and ϕ are integrated out

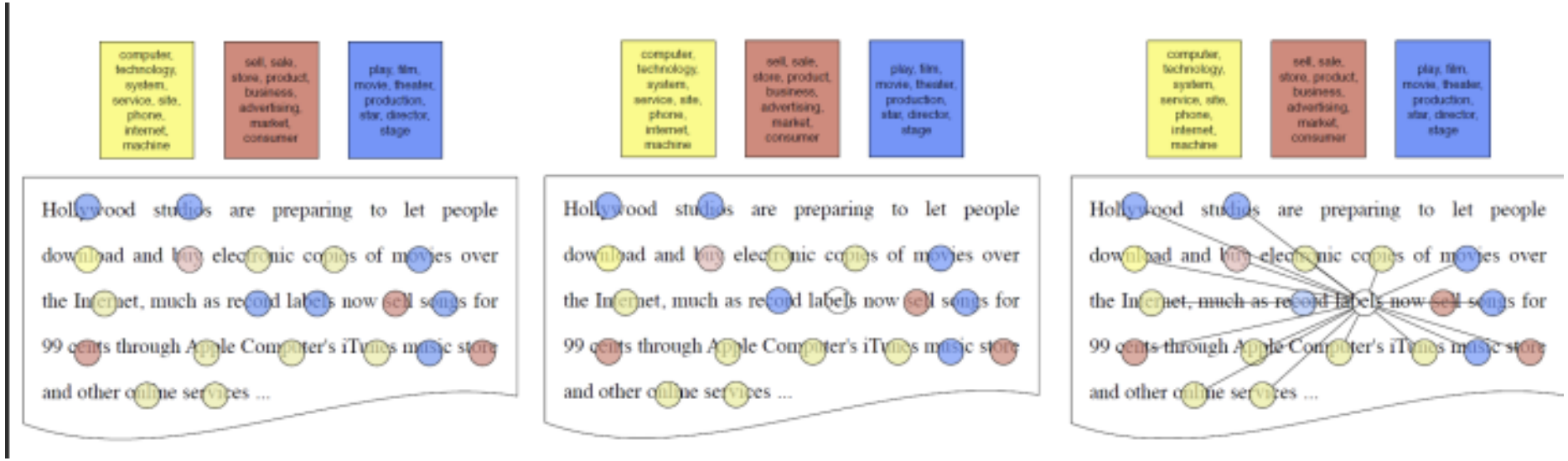


LDA Inference

- LDA Inference : Gibbs Sampling

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w})$$

  - 모든 단어에 대해서 알고, i번째 단어 빼고 모든 단어에 대해 토픽도 알 때, i가 j토픽일 확률을 구한다.
    - ✓ 예시는 토픽이 총 3개가 있다.
    - ✓ 모든 단어에 대해 토픽이 할당되면, 1iter 라고 한다.(많으면 10000번 이상 이터레티어가 돌 수 있다.)



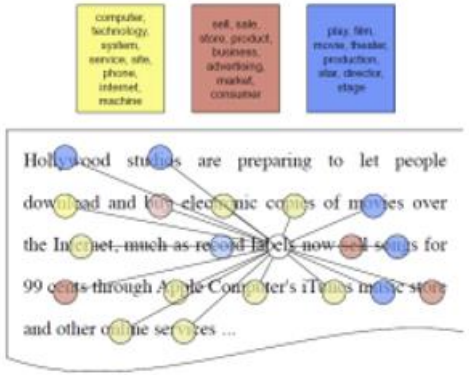
LDA Inference

- LDA Inference : Gibbs Sampling

$$p(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto p(z_i = j, \mathbf{z}_{-i}, \mathbf{w})$$

$$= p(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i})p(z_i = j|\mathbf{z}_{-i}, \mathbf{w}_{-i})$$

$$= p(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i})p(z_i = j|\mathbf{z}_{-i})$$



  - 맨 위 식은 조건부 확률에서 분자만 따왔을 때 비례 한다는 의미이다.
    - ✓ 그 식은, i빼고 모든 단어에 대해 알고, i빼고 모든 단어에 대해 토픽을 알 때, i번째 단어의 토픽이 j에서 나왔다 치고, i빼고 모든 단어에 대해 알고, i빼고 모든 단어에 대해 토픽을 알 때 i가 j번째 토픽에서 나왔을 때의 곱으로 분해가 가능 하다.
    - ✓ I빼고 모든 단어를 의미하는 w_{-i}는 상수 취급이 가능하다.
  - 파란색은 likelihood이고, 빨간색은 prior이다.

LDA Inference

- LDA Inference : Gibbs Sampling

$$p(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \int p(w_i|z_i = j, \phi^{(j)}) p(\phi^{(j)}|\mathbf{z}_{-i}, \mathbf{w}_{-i}) d\phi^{(j)}$$
$$= \int \phi_{w_i}^{(j)} p(\phi^{(j)}|\mathbf{z}_{-i}, \mathbf{w}_{-i}) d\phi^{(j)}$$

$$p(\phi^{(j)}|\mathbf{z}_{-i}, \mathbf{w}_{-i}) \propto p(\mathbf{w}_{-i}|\phi^{(j)}, \mathbf{z}_{-i}) p(\phi^{j}) \sim Dirichlet(\beta + n_{-i,j}^{(w)})$$

  - 파이j는 j토픽이 가질 수 있는 모든 단어의 분포를 의미한다.

$$p(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta}$$

  - $n_{-i,j}^{(\cdot)}$ 는 j번째 토픽에 몇 개의 단어가 할당 되어 있는지를 의미한다.(w_i 제외)

  - 베타는 스무딩을 위한 파라미터 이고, V는 vocab size를 의미한다.

LDA Inference

- LDA Inference : Gibbs Sampling

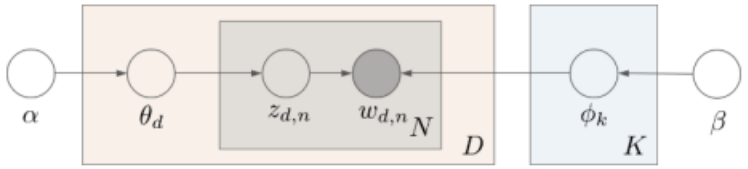
$$p(z_i = j|\mathbf{z}_{-i}) = \int p(z_i = j|\theta^{(d)})p(\theta^{(d)}|\mathbf{z}_{-i})d\theta^{(d)}$$

$$p(\theta^{(d)}|\mathbf{z}_{-i}) \propto p(\mathbf{z}_{-i}|\theta^{(d)})p(\theta^{(d)}) \sim Dirichlet(n^{(d)}_{-i,j} + \alpha)$$

$$p(z_i = j|\mathbf{z}_{-i}) = \frac{n^{(d)}_{-i,j} + \alpha}{n^{(d)}_{-i,\cdot} + K\alpha}$$

LDA Inference

- LDA Inference : Gibbs Sampling



$$p(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha}$$

- $n_{-i,j}^{(d)}$ 문서에서 wi 제외 토픽j에 할당된 단어 수를 의미한다.
- $n_{-i,\cdot}^{(d)}$ 문서에서 wi 제외 모든 토픽에 대한 모든 단어 수를 의미한다.
- $n_{-i,j}^{(w_i)}$ 코퍼스에서 wi 제외 토픽j에 할당된 단어의 수를 의미한다.
- $n_{-i,j}^{(\cdot)}$ 코퍼스에서 wi 제외 토픽j에 할당된 전체 단어의 수를 의미한다.

LDA Inference

- Parameter Estimation

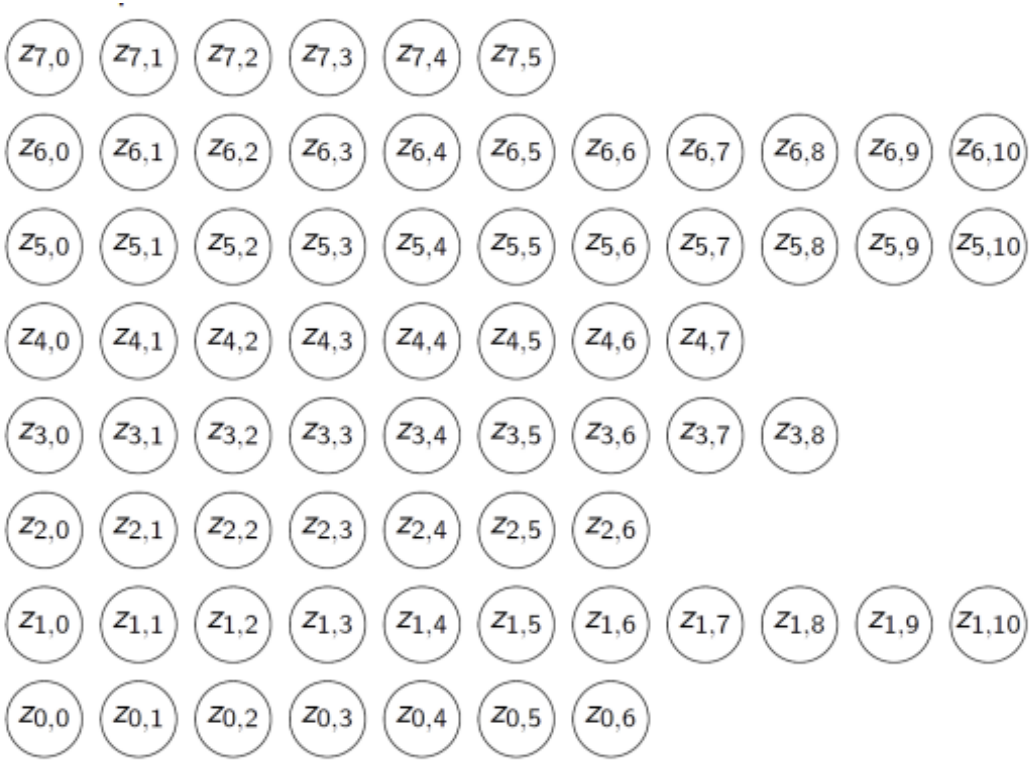
$$\phi_{j,w} = \frac{n_w^{(j)} + \beta}{\sum_{w=1}^{V} n_w^{(j)} + V\beta} \qquad \theta_j^{(d)} = \frac{n_j^{(d)} + \alpha}{\sum_{z=1}^{K} n_z^{(d)} + K\alpha}$$

  - J토픽에 대한 단어의 비중은 w가 j토픽에 할당 된 횟수 + 베타 / j번째 토픽에 할당된 전체 단어 수 +베타로 나타낼 수 있다.
  - 문서안에 j토픽에 대한 비중은, 문서별로 j토픽의 비중+알파/ 문서별로 전체 단어 수 + 알파로 나타낼 수 있다.

$$p(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha} \qquad \frac{n_{d,k} + \alpha_k}{\sum_{i=1}^{K}(n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j} + \beta_j)}$$

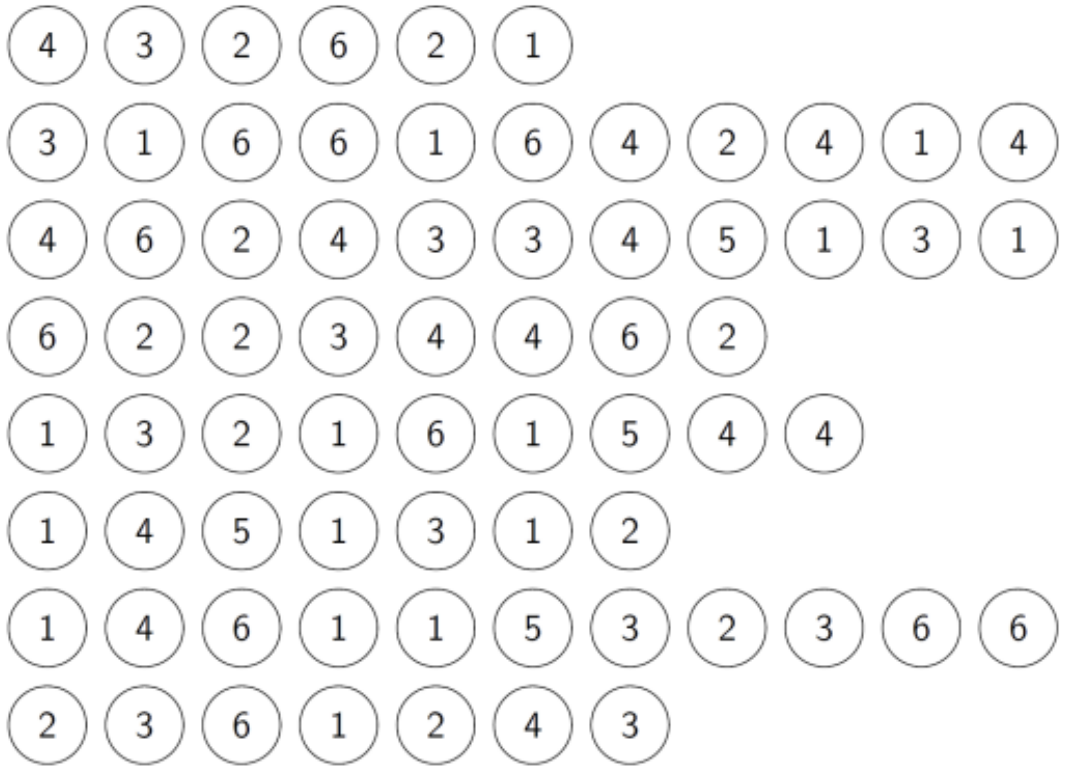
LDA Inference

- Example
  - 가로축은 문서별로 단어의 개수를 의미한다.
  - 세로축은 문서를 의미한다.



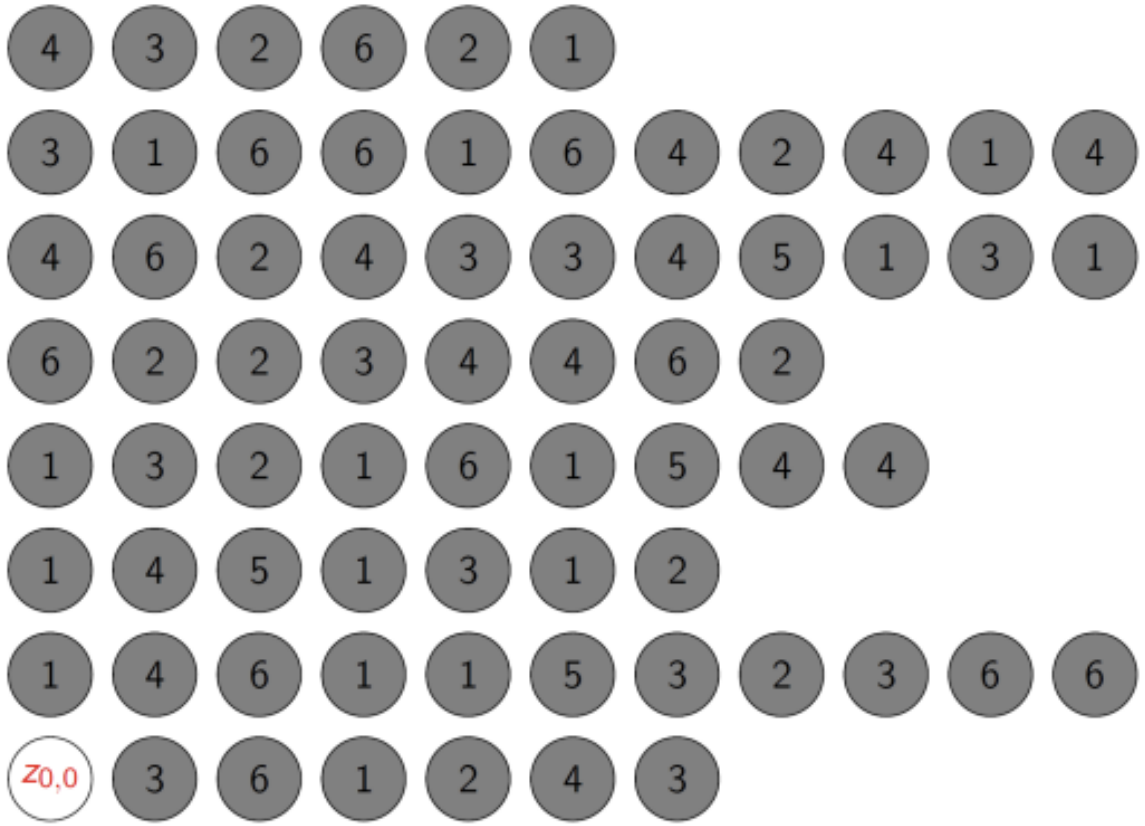
LDA Inference

- Example
  - 각 단어에 대해 랜덤으로 토픽을 할당한다.



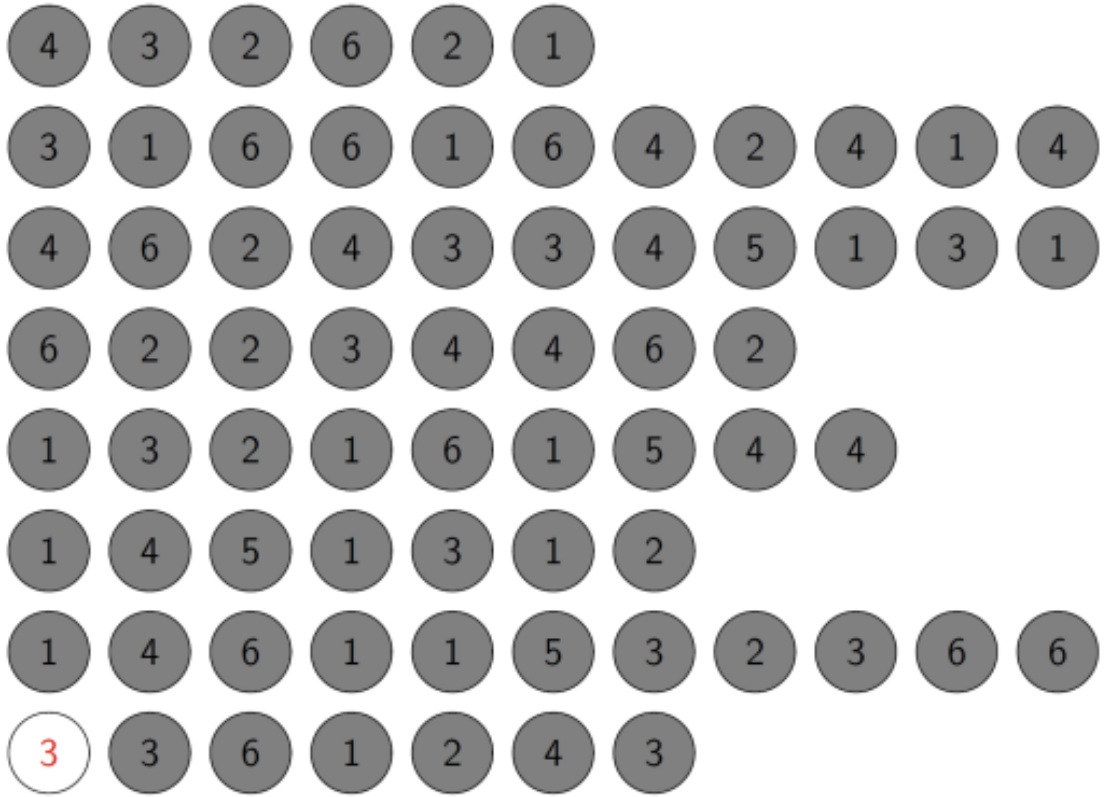
LDA Inference

- Example
  - z_{-i} 빼고 토픽을 고정시킨다.
  - w에 대해서 알고 있다.



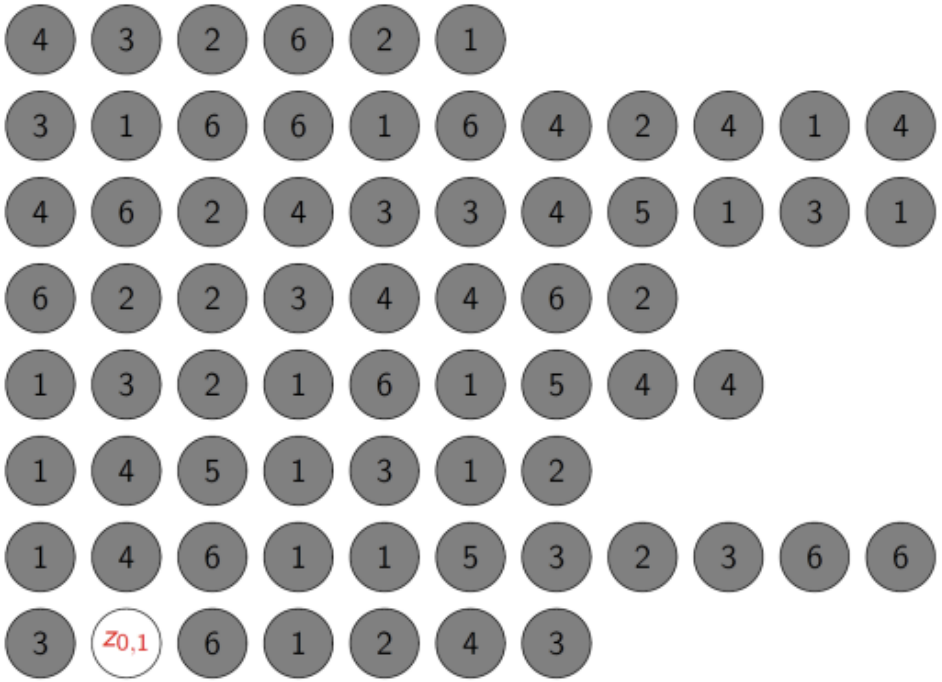
LDA Inference

- Example
  - z_{-i} 빼고 토픽을 고정시킨다.
  - w에 대해서 알고 있다.



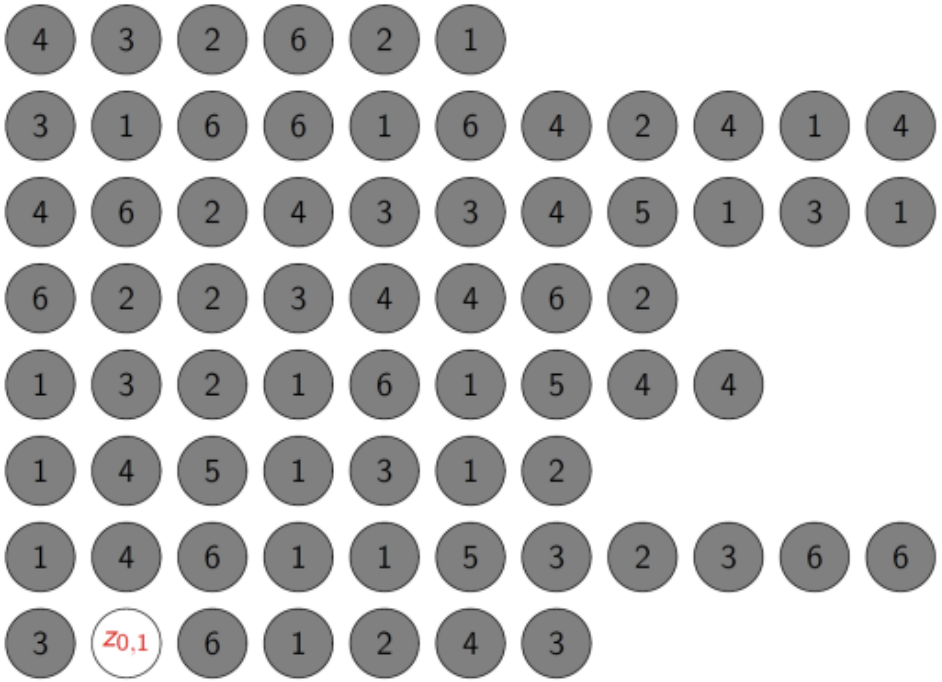
LDA Inference

- Example
  - z_{-i} 빼고 토픽을 고정시킨다.
  - w에 대해서 알고 있다.
  - 모든 단어에 대해 한바퀴 돌면 1iter라고 한다.



LDA Inference

- Example
  - z_{-i} 빼고 토픽을 고정시킨다.
  - w에 대해서 알고 있다.
  - 모든 단어에 대해 한바퀴 돌면 1iter라고 한다.



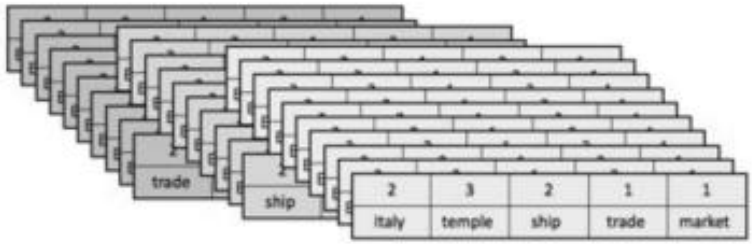
LDA Inference

- Example
  - 단어들에 대한 토픽을 랜덤으로 배정한다.
    - ✓ 모든 문서에 대해 반복한다.

Etruscan	trade	price	temple	market

z → w

3	2	1	3	1
Etruscan	trade	price	temple	market



3	2	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

Total counts from all docs

$$\frac{n_{d,k}+\alpha_k}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)} \times \frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j}+\beta_j)}$$

LDA Inference

- Example
  - Trade의 토픽을 마스킹 한다.
    - ✓ 문서안에서도 코퍼스 안에서도 수가 바뀐 것을 확인할 수 있다.
  - 토픽1에 단어가 2개, 토픽3에 단어가 2개 할당 되어서 파란 막대가 긴 것을 볼 수 있다.
    - ✓ Topic2에 파란 막대가 있는 이유는?
  - 코퍼스 안에 단어의 빈도를 가지고 빨간 막대를 그릴 수 있다.

3	2	1	3	1
Etruscan	trade	price	temple	market

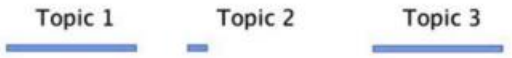
	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	7	1
...			

$$\frac{n_{d,k}+\alpha_k}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)} \times \frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j}+\beta_j)}$$

3	?	1	3	1
Etruscan	trade	price	temple	market

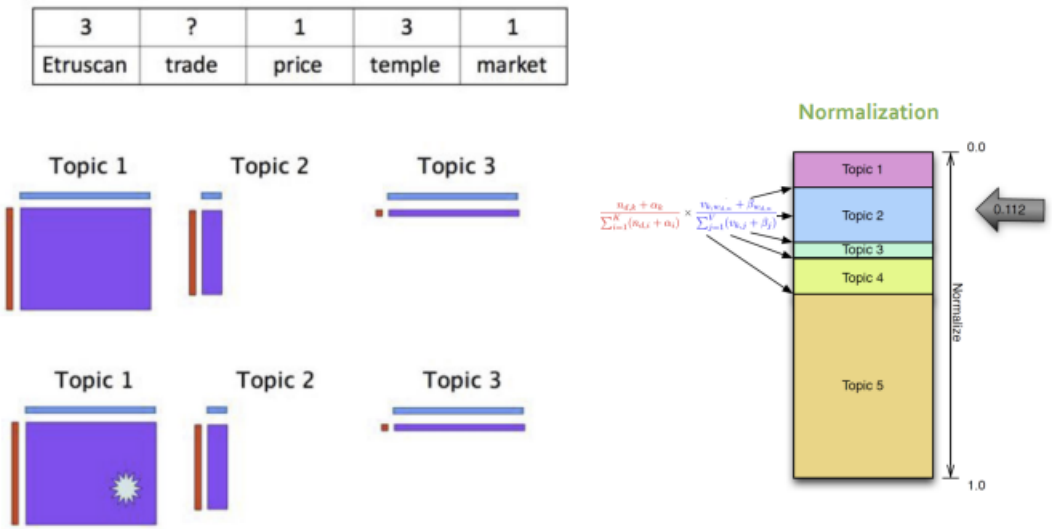


3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
trade	10	7	1

LDA Inference

- Example
  - 가로축
    - ✓ 현재 문서가 해당 토픽을 선호 하는 정도를 의미한다.
  - 세로축
    - ✓ 토픽의 단어 선호도를 의미한다.



LDA Inference

- Example
  - Update count

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**10**	7	1
...			

3	**1**	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**11**	7	1
...			

02

LDA Evaluation

LDA Evaluation

- LDA Evaluation
  - 퍼플렉시티를 통해 수치적으로 구할 수 있다.
    - ✓ 하지만 실제로는 정성적인 방법을 많이 사용한다.

$$\text{Perplexity}(w) = \exp\left\{-\frac{\log(p(w))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right\} \qquad \log(p(w)) = \sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)} \log\left[\sum_{K=1}^{k} \theta_K^{(d)} \beta_K^{(j)}\right]$$