

Probabilistic Topic Models

Jeonghun Yoon

Natural Language Processing

Natural language(자연어) : 일상 생활에서 사용하는 언어

Natural language processing(자연어 처리) : 자연어의 의미를 분석하여 컴퓨터가 여러가지 일들을(tasks) 처리할 수 있도록 하는 것

Easy

- Spell checking, Keyword search, Finding synonyms

Medium

- Parsing information from websites, documents, etc

Hard

- Machine translation
- ***Semantic analysis***
- ***Coherence***
- Question answering

Semantic analysis

언어학에서의 의미 분석

- 자연어를 이해하는 기법 중 하나로, 문장의 의미(meaning, semantic)에 근거하여 문장을 해석하는 것을 의미
- Syntax analysis의 반대(lexicon, syntax analysis)

머신러닝에서의 의미 분석

- Corpus에 있는 많은 documents 집합에 내제되어 있는(latent) meanings, concepts, subjects, topics 등을 추정할 수 있는 구조를 생성하는 것을 의미
- 대표적인 의미 분석 기법
 - Latent Semantic Analysis(LSA or LSI)
 - PLSI
 - Latent Dirichlet Allocation(LDA)
 - Hierarchical Dirichlet Processing(HDP)

Semantic analysis

Representation of documents



Axes of a spatial

- Euclidean space에서 정의 가능
- Hard to interpret

Probabilistic topics

- 단어상에 정의된 probability distribution
- Interpretable

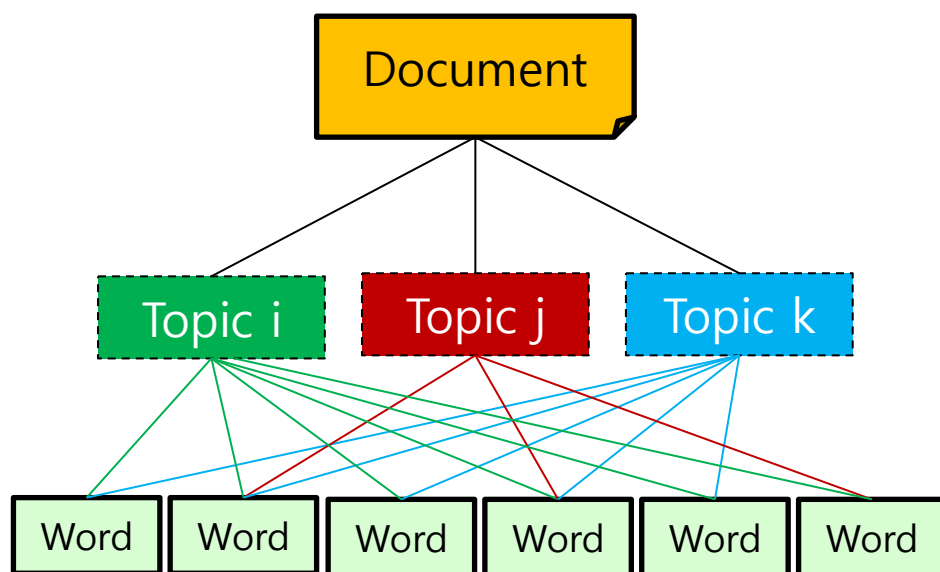
1. Probabilistic topics
- LDA

2. Axes of a spatial
- LSA

Topic models

Topic models의 기본 아이디어

- 문서는 토픽들의 혼합 모델이며 각 토픽은 단어상에 정의된 확률분포



Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic models

예제)

Doc 1 : I like to eat broccoli and bananas.

Doc 2 : I ate a banana and tomato smoothie for breakfast.

Doc 3 : Dogs and cats are cute.

Doc 4 : My sister adopted a cats yesterday.

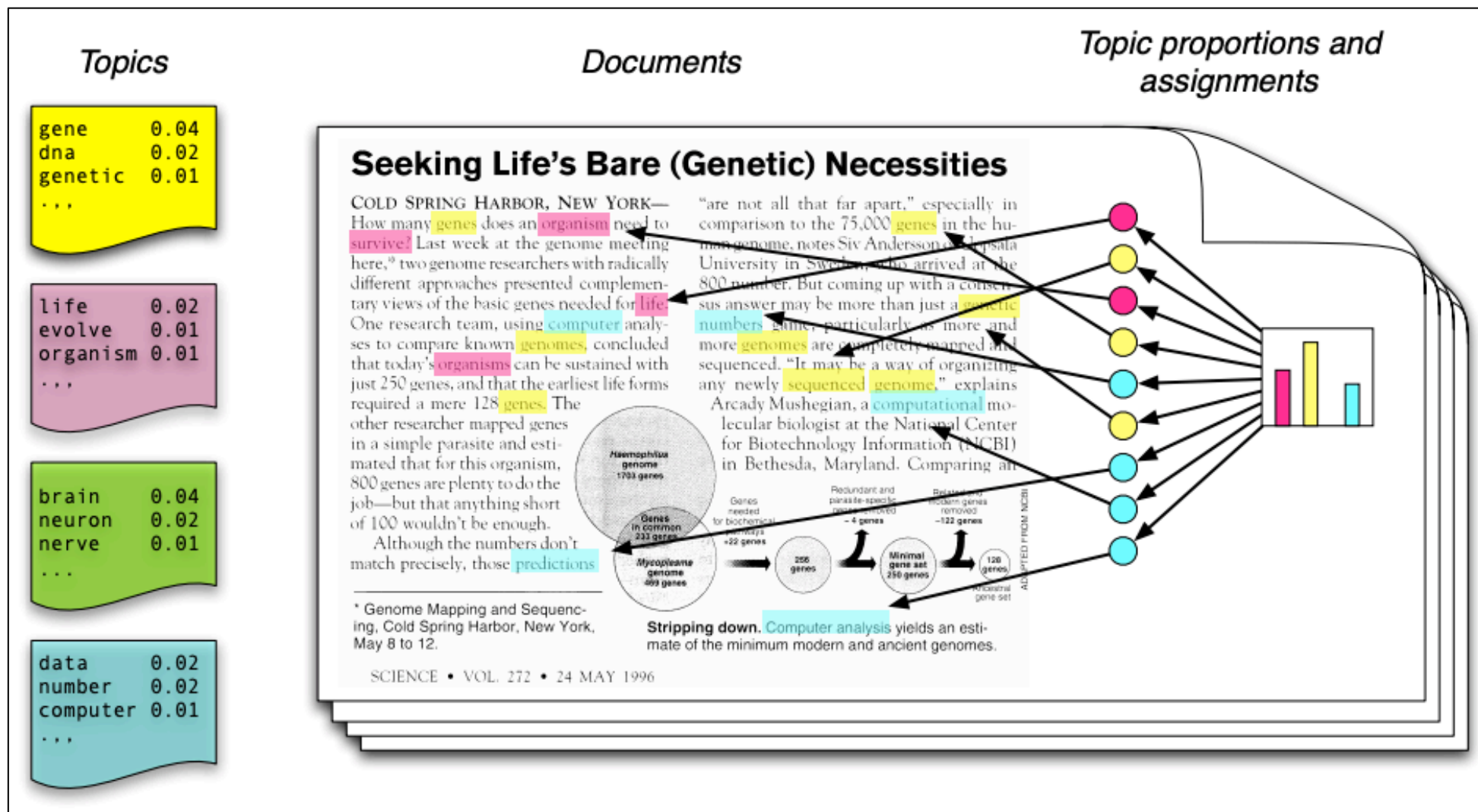
Doc 5 : Look at this cute hamster munching on a piece of broccoli.



- Doc 1 and 2 : 100% topic A
- Doc 3 and 4 : 100% topic B
- Doc 5 : 60% topic A, 40% topic B

- Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, ...
- Topic B: 20% cats, 20% cute, 15% dogs, 15% hamster, ...

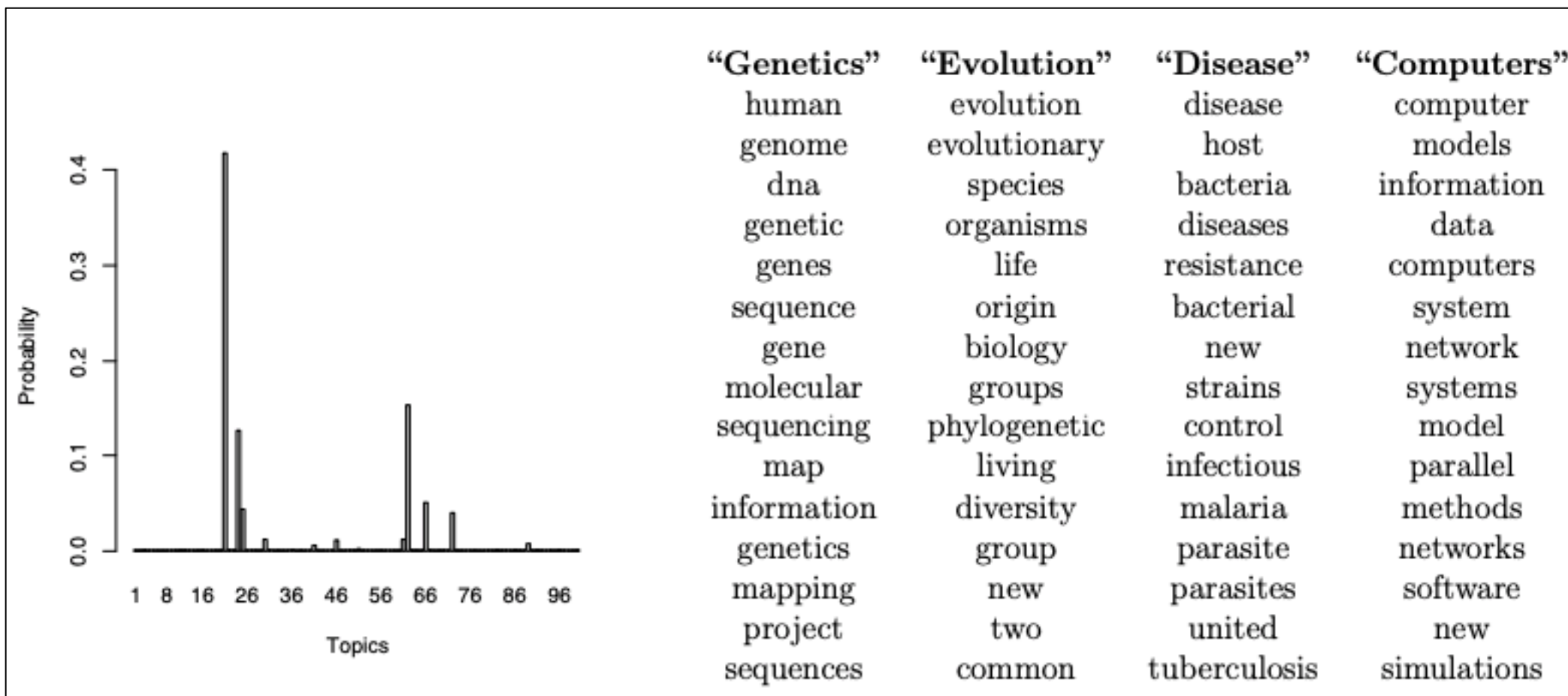
Topic models



Topic models

(Left) 문서에서의 topic proportion

(Right) 문서에서 비중이 높았던 토픽에 대하여,
토픽별 문서내 빈도수가 가장 높은 단어



Probabilistic Topic Models의 구조

모델의 정의에 앞서, 필요한 단어들의 수학적 표기

- Word : $\{1, \dots, V\}$ 를 인덱스로 가지는 vocabulary 상의 items
- Document : N word의 sequence
 - $\mathbb{w} = (w_1, w_2, \dots, w_N)$, w_n : word의 sequence내에서 n 번째에 있는 word
- Corpus : D documents의 collection
 - $\mathcal{C} = \{\mathbb{w}_1, \mathbb{w}_2, \dots, \mathbb{w}_D\}$

Probabilistic Topic Models의 구조

문서 d 의 단어 w_i 대한 분포 :

$$P(w_i) = \sum_{k=1}^K P(w_i|z_i = k)P(z_i = k)$$

- $P(w_i|z_i = k)$: 토픽 k 에서, 단어 w_i 의 probability
 - 각 토픽에서 어떤 단어들이 중요할까?
- $P(z_i = k)$: i 번째 단어에 토픽 k 가 할당되는 probability (즉, 토픽 j 가 i 번째 단어를 위해 샘플링 될 확률)

$\beta_k = P(w|z = k)$: 토픽 k 에서, 단어들의 multinomial distribution

$\theta_d = P(z)$: 문서 d 에서, 토픽들의 multinomial distribution

Latent Dirichlet Allocation의 등장

문서 d 의 단어 w_i 대한 분포 :

$$P(w_i) = \sum_{k=1}^K P(w_i | z_i = k) P(z_i = k)$$

LDA는 Dirichlet distribution을 θ 의 prior로 사용
(Blei et. Al, 2003)

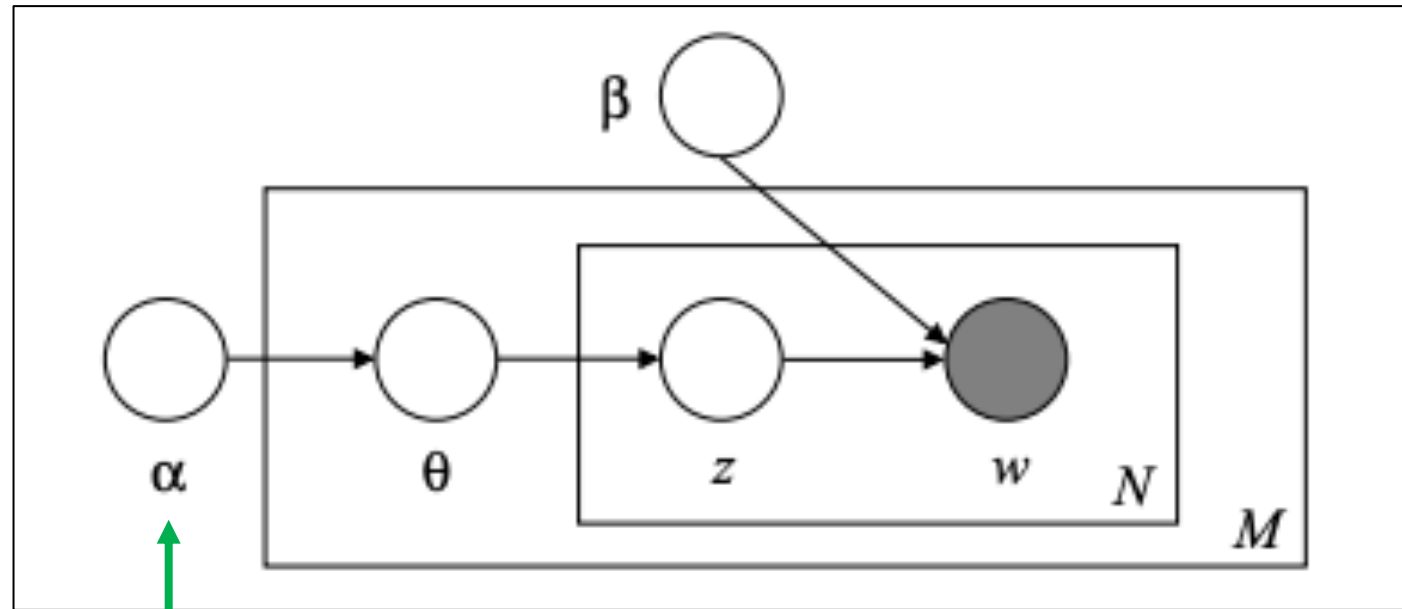
디리클레 분포(Dirichlet distribution)은 multinomial distribution의 켤레 사전 분포로(conjugate prior) 사용

다항 분포(Multinomial distribution) $p = (p_1, \dots, p_K)$ 에 대한 Dirichlet distribution :

$$Dir(\alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1}$$

- Hyperparameter α_j : 문서 d 에서 토픽 j 가 샘플링 된 횟수에 대한 사전 관찰 count (문서로부터 단어가 실제로 관찰되기 이전의 값)


Latent Dirichlet Allocation의 등장



LDA :
Dirichlet parameter

Variant LDA의 등장

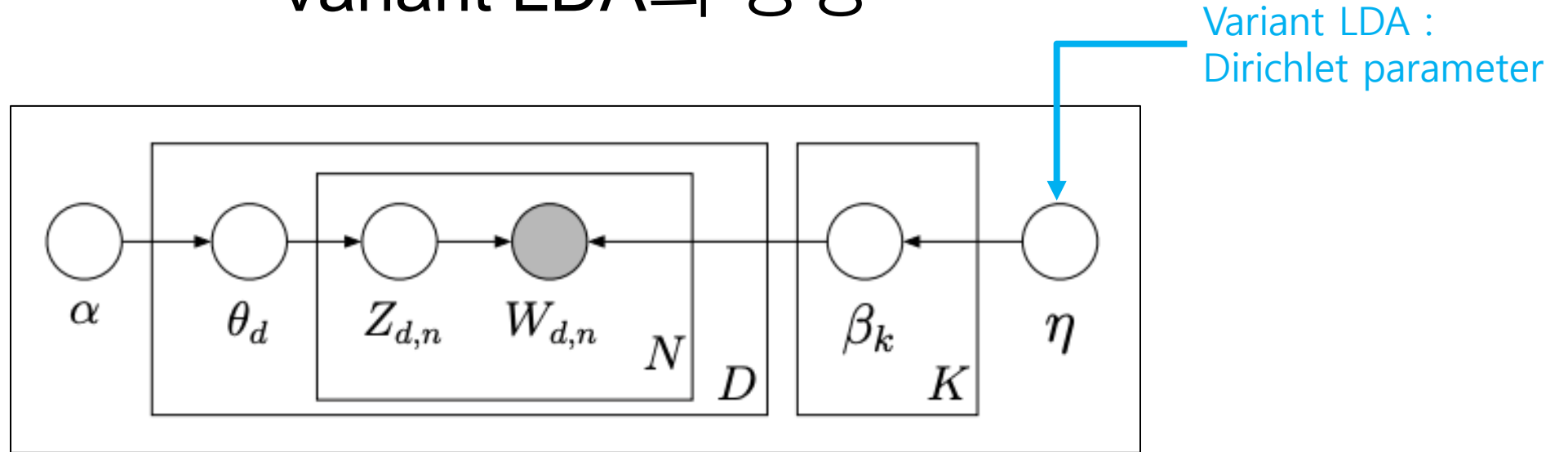
문서 d 의 단어 w_i 대한 분포 :

$$P(w_i) = \sum_{k=1}^K P(w_i | z_i = k) P(z_i = k)$$


Variant LDA는 symmetric Dirichlet distribution(η)을 β 의 prior로 사용
(Griffiths and Steyvers, 2004)

Hyperparameter η : Corpus의 단어가 관찰되기 이전에, 토픽에서 단어가 샘플링 된 횟수에 대한 사전 관찰 count

Variant LDA의 등장



α	Dirichlet parameter		
θ_d	문서 d 에서 토픽 비율(proportion)	$\theta_{d,k}$	문서 d 에서 특정 토픽 k 의 proportion
Z_d	문서 d 에서 토픽 할당(assignment)	$Z_{d,n}$	문서 d 에서 n -th 단어에 대한 토픽 할당
W_d	문서 d 에서 관찰된 단어들	$W_{d,n}$	문서 d 에서 n -th 단어
β_k	토픽 k 의 vocabulary에서의 분포 (단어 전체 셋에서 정의된 토픽 k 의 분포)	η	Dirichlet parameter
The plate surrounding θ_d		각 문서 d 에 대하여, 토픽 분포의 sampling (총 D 개의 문서)	
The plate surrounding β_k		각 topic k 에 대하여, 단어 분포의 sampling (총 K 개의 토픽)	

LDA 모델의 변수

θ'_d s :

Document	Topic 1	Topic 2	Topic 3	...	Topic K
Document 1 (θ_1)	0.2	0.4	0.0	...	0.1
Document 2 (θ_2)	0.8	0.1	0.0	...	0.0
...
Document M (θ_M)	0.5	0.4	0.1	...	0.0

→ 합 : 1

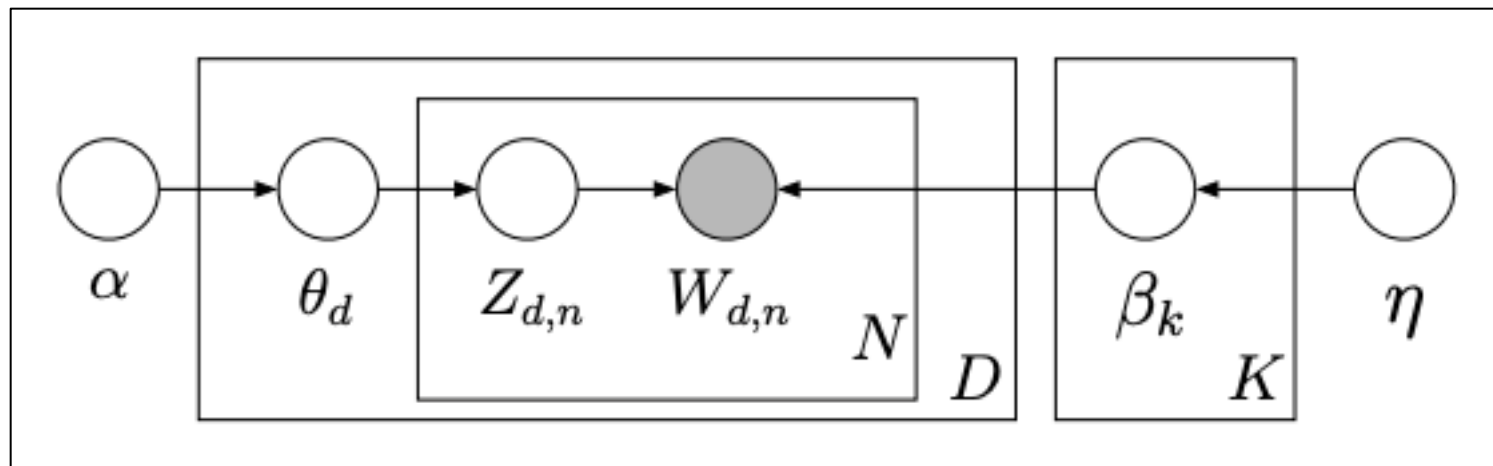
β'_k s :

Terms	Topic 1 (β_1)	Topic 2 (β_2)	Topic 3 (β_3)	...	Topic K (β_K)
Word 1	0.02	0.09	0.00	...	0.00
Word 2	0.08	0.52	0.37	...	0.03
...
Wordt V	0.05	0.12	0.01	...	0.45

↓
합 : 1

- Variant LDA version을 사용하고 있으므로, 이 version을 LDA 라고 지칭 하겠음

LDA의 Generative process



LDA는 **generative model**이다.

1. 문서의 단어의 갯수 N 을 Poisson 분포를 이용하여 선택한다. $N \sim \text{Poisson}(\xi)$
2. 문서의 토픽 분포(proportion) θ_d 를 Dirichlet(α) 분포를 이용하여 선택한다. $\theta_d \sim \text{Dirichlet}(\alpha)$
3. 문서의 단어 각각에 대하여
 - a. 토픽 분포 θ_d 를 이용하여, 단어에 토픽을 할당한다. $Z_{d,n} \sim \text{Multinomial}(\theta)$
 - b. $p(W_{d,n}|Z_{d,n}, \beta)$ 를 이용하여 단어를 선택한다. 이 확률분포는 다항분포이다.

LDA의 Generative process

예제)

1. 새로운 문서 D 의 길이를 5로 선택한다. 즉, $D = (w_1, w_2, w_3, w_4, w_5)$
2. 문서 D 의 토픽 분포를 50%는 음식(food), 50%는 동물(animal)로 선택한다.
3. 각 단어에 대하여,
 1. 첫번째 단어 w_1 에 food topic을 할당한다. Food topic에서 broccoli를 w_1 으로 선택한다.
 2. 두번째 단어 w_2 에 animal topic을 할당한다. Animal topic에서 panda를 w_2 으로 선택한다.
 3. 세번째 단어 w_3 에 animal topic을 할당한다. Animal topic에서 adorable 를 w_3 으로 선택한다.
 4. 네번째 단어 w_4 에 food topic을 할당한다. Food topic에서 cherries 를 w_4 으로 선택한다.
 5. 다섯번째 단어 w_5 에 food topic을 할당한다. Food topic에서 eating 를 w_5 으로 선택한다.



D : broccoli panda adorable cherries eating

LDA 모델의 inference

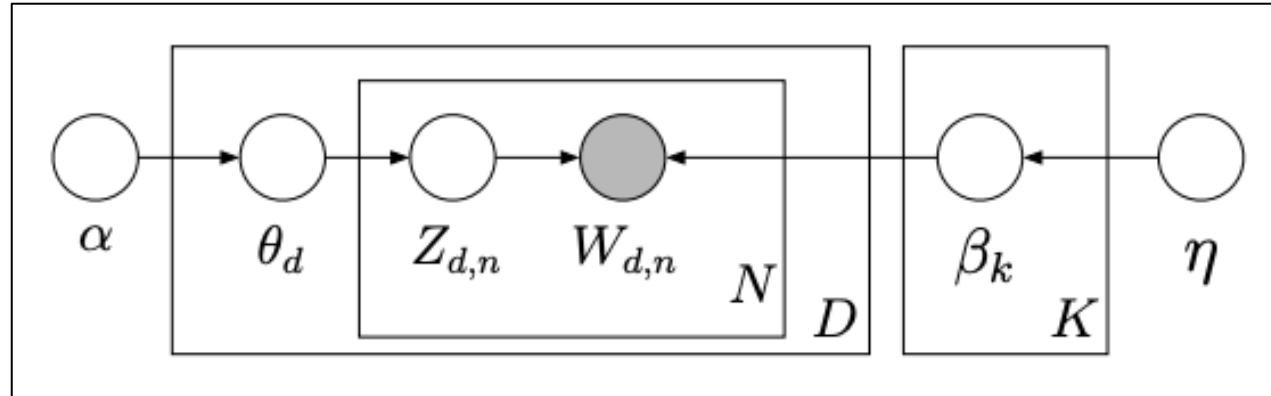
관찰 가능한 문서 내 단어 $w_{d,n}$ 를 이용하여, LDA 모델의 잠재 변수(hidden variable)인 문서의 토픽분포 θ_d 와 토픽의 단어분포 β_k 를 추정하는 과정이 inference이다.

Generative probabilistic modeling에서는, data는 잠재 변수(hidden variable)를 포함하는 generative process에서부터 발생하는것으로 다룬다. 따라서, generative process는 observed random variable과 hidden random variable의 결합 확률밀도(joint probability distribution)를 정의한다.

- Observed variables : 문서내의 단어들
- Hidden variables : 문서의 토픽 분포, 토픽의 단어 분포 (topic structure)

결합 확률밀도함수를 이용하여 observed variable이 주어졌을 때 hidden variable의 조건부 분포를 구한다. 이 분포는 사후 확률분포(posterior distribution)이다.

LDA 모델의 inference



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

관찰 가능 데이터 $w_{1:D}$ 를 통해서 inference해야 할 변수 : $\beta_{1:D}, \theta_{1:D}, z_{1:D}$

Posterior dist.

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

$\beta_{1:K}$	토픽 1~K의 vocabulary에서의 분포
$\theta_{1:D}$	문서 1~D에서의 토픽 비율
$z_{1:D}$	문서 1~D에서의 토픽 할당
$w_{1:D}$	문서 1~D에서 관찰된 단어들

LDA 모델의 inference

Posterior distribution을 구하는 것은 쉬운것인가?

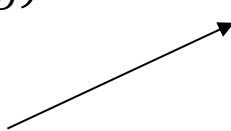
분자의 경우를 먼저 살펴보자.

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

모든 random variable의 결합 확률밀도 함수는,
hidden variable이 임의로 셋팅된다면 쉽게 계산 가능

LDA 모델의 inference

분모의 경우를 살펴보자.

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



Observed variable의 주변 밀도함수(marginal probability)

- 임의의 topic model에서, observed corpus를 볼 수 있을 확률을 구하는 것
- 모든 hidden topic structure의 가능한 경우(instantiation)를 구하고, 결합 확률밀도 함수를 summation

가능한 hidden topic structure는 지수적으로 많다. 따라서 해당 밀도함수를 구하는 것은 매우 어렵다.

Modern probabilistic models, Bayesian statistics에서는 분모의 분포 때문에 posterior를 계산하는 것이 어렵다. 따라서 posterior를 효과적으로 추정하는 기법에 대한 연구가 많이 이루어지고 있다.

따라서, topic modeling algorithms에서도 **posterior distribution을 추정하기 위한 기법**을 활용한다.

- sampling based method
 - variational method
- 

Difficulty of deriving marginal probability $p(w_{1:D})$

- Topic mixture $\theta \ni$ joint distribution (parameter : α, β)

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta)$$

- Document \ni marginal distribution

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta$$

- Corpus \ni probability

$$p(D | \alpha, \beta) = \prod_{d=1}^M \int p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d$$

Algorithm for extracting topics

Gibbs sampling

$$P(\mathbf{z}_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \left(\frac{C_{w_{ij}}^{VK} + \eta}{\sum_{w=1}^V C_{wj}^{VK} + V\eta} \right) \left(\frac{C_{d_{ij}}^{DK} + \alpha}{\sum_{t=1}^K C_{d_{it}}^{DK} + T\alpha} \right)$$

Smoothing

문서에서 i 번째에 나오는 단어 w 의 토픽이 j 일 확률에 영향을 미치는 요소

- 요소 1 : 토픽 j 에 할당된 전체 단어 중에서 해당 단어의 점유율이 높을수록 j 일 확률이 크다.
- 요소 2 : w_i 가 속한 문서 내 다른 단어가 토픽 j 에 많이 할당되었을수록 j 일 확률이 크다.

$z_i = j$	문서에서 i 번째에 나오는 단어 w 에 토픽 j 가 할당
\mathbf{z}_{-i}	i 번째 단어를 제외한 다른 단어들에 대한 토픽 할당
w_i	단어 index
d_i	문서 index
\cdot	다른 정보 및 observed information
C_{wj}^{WK}	단어 w 가 토픽 j 에 할당된 횟수 (현재 i 는 제외)
C_{dj}^{DK}	문서 d 의 단어들 중에서 토픽 j 에 할당된 횟수 (현재 i 는 제외)
η	토픽의 단어 분포 생성에 사용되는 Dirichlet parameter
α	문서의 토픽 분포 생성에 사용되는 Dirichlet parameter

Algorithm for extracting topics

예제)

Doc 0 : ($z_{0,0}, z_{0,1}, z_{0,2}, z_{0,3}$)

Doc 1 : ($z_{1,0}, z_{1,1}, z_{1,2}$)

Doc 2 : ($z_{2,0}, z_{2,1}, z_{2,2}, z_{2,3}$)

Doc 3 : ($z_{3,0}, z_{3,1}, z_{3,2}, z_{3,3}, z_{5,4}$)

Doc 4 : ($z_{4,0}, z_{4,1}, z_{4,2}, z_{4,3}, z_{4,4}$)

Doc 5 : ($z_{5,0}, z_{5,1}, z_{5,2}, z_{5,3}, z_{5,4}, z_{5,5}$)

$z_{i,j}$: i 번째 문서에 j 토픽이 할당된 것을 나타내는 확률변수

1. 확률변수에 랜덤하게 토픽을 할당
2. $z_{0,0}$ 을 제외한 값들을 토대로 $z_{0,0}$ 의 값을 업데이트
3. $z_{0,1}$ 을 제외한 값들을 토대로 $z_{0,1}$ 의 값을 업데이트
-
4. $z_{5,5}$ 을 제외한 값들을 토대로 $z_{5,5}$ 의 값을 업데이트
5. 확률변수가 수렴할 때까지 반복

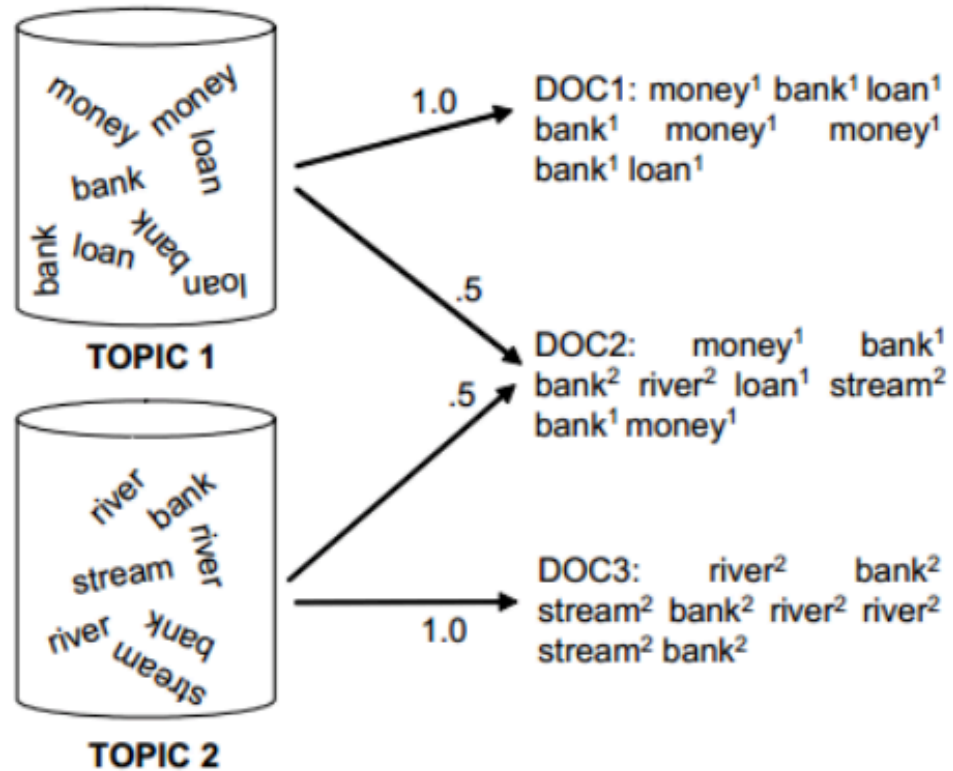
Algorithm for extracting topics

$$C^{VK} = \begin{pmatrix} C_{11} & C_{12} & \dots & C_{1k} & \dots & C_{1K} \\ C_{21} & C_{22} & \dots & C_{2k} & \dots & C_{2K} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ C_{v3} & C_{v3} & \dots & C_{vk} & \dots & C_{vK} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ C_{V1} & C_{V2} & \dots & C_{Vk} & \dots & C_{VK} \end{pmatrix}$$

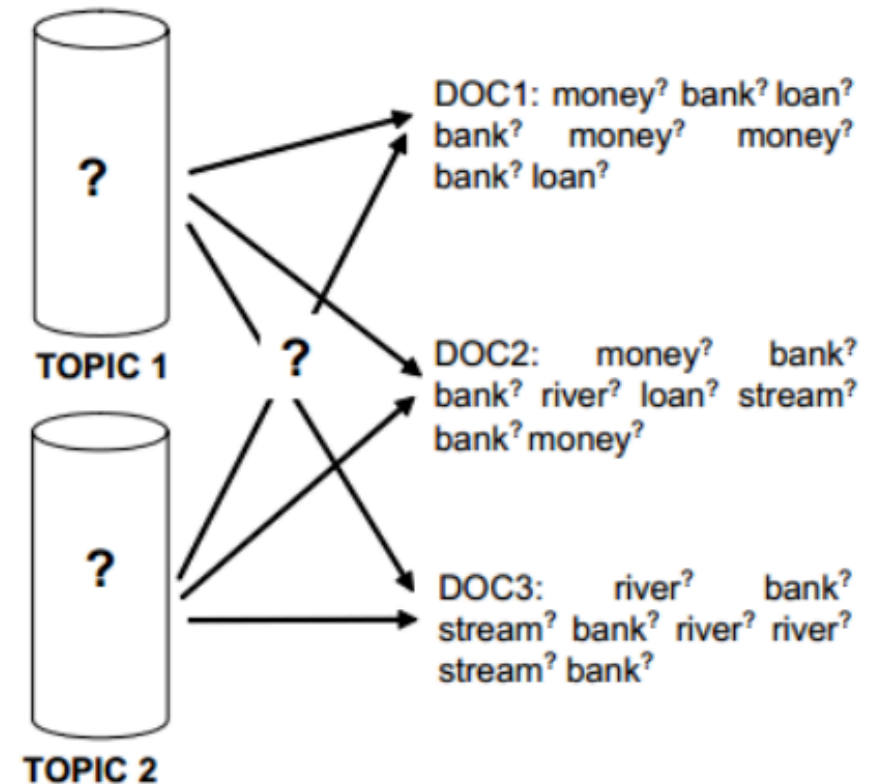
$$C^{DK} = \begin{pmatrix} C_{11} & C_{12} & \dots & C_{1k} & \dots & C_{1K} \\ C_{21} & C_{22} & \dots & C_{2k} & \dots & C_{2K} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ C_{d3} & C_{d3} & \dots & C_{dk} & \dots & C_{dK} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ C_{D1} & C_{D2} & \dots & C_{Dk} & \dots & C_{DK} \end{pmatrix}$$

Generative model vs. Statistical inference

PROBABILISTIC GENERATIVE PROCESS



STATISTICAL INFERENCE



최적의 토픽 수

Perplexity

- Language modeling에서 주로 컨벤션으로 사용한다.
- LDA에서 추정된 토픽 정보를 이용하여 단어의 발생 확률을 계산하였을 때, 확률값이 높을수록 generative process를 제대로 설명한다고 본다.

$$Perplexity(C) = \exp\left(-\frac{\sum_{d=1}^D \log p(\mathbf{w}_d)}{\sum_{d=1}^D N_d}\right)$$

- $p(\mathbf{w}_d)$: 토픽의 단어분포 정보와 문서내 토픽의 비중 정보의 곱을 이용하여 계산
- $p(\mathbf{w}_d)$ 는 클수록 좋으므로, perplexity는 작을수록 좋다.

최적의 토픽 수

Topic coherence