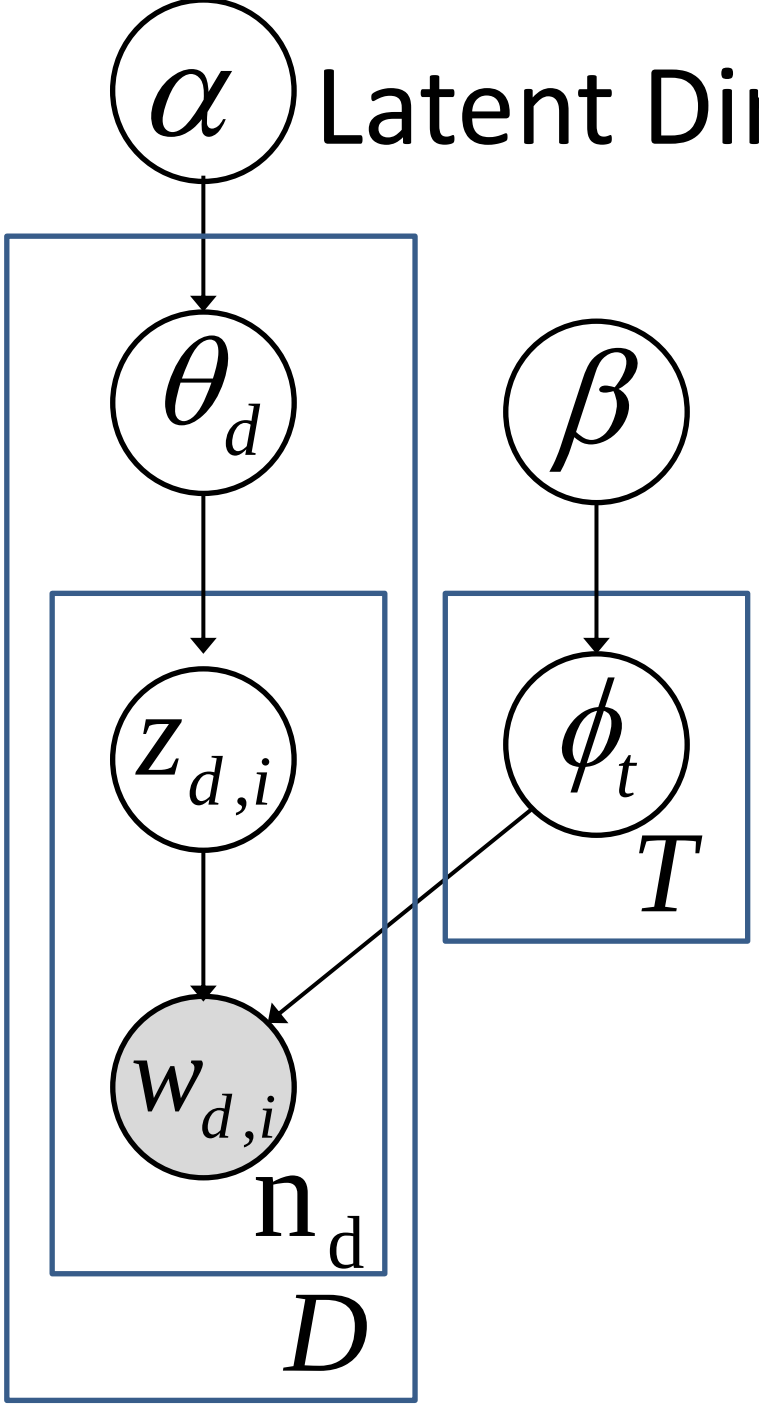
簡潔なLDAの学習

佐藤一誠(東大)@issei_sato
岡野原大輔(PFI)@hillbig

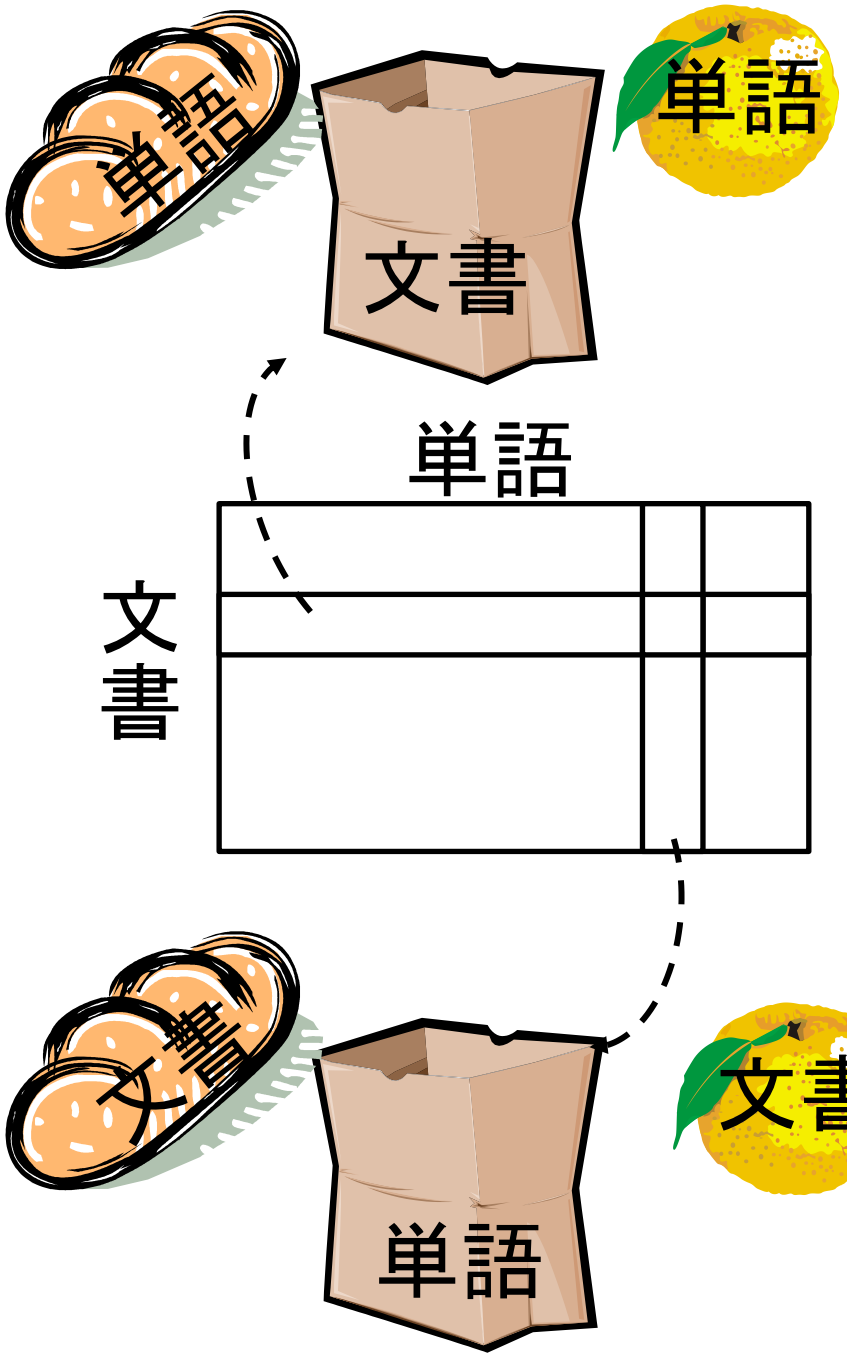
Latent Dirichlet Allocation

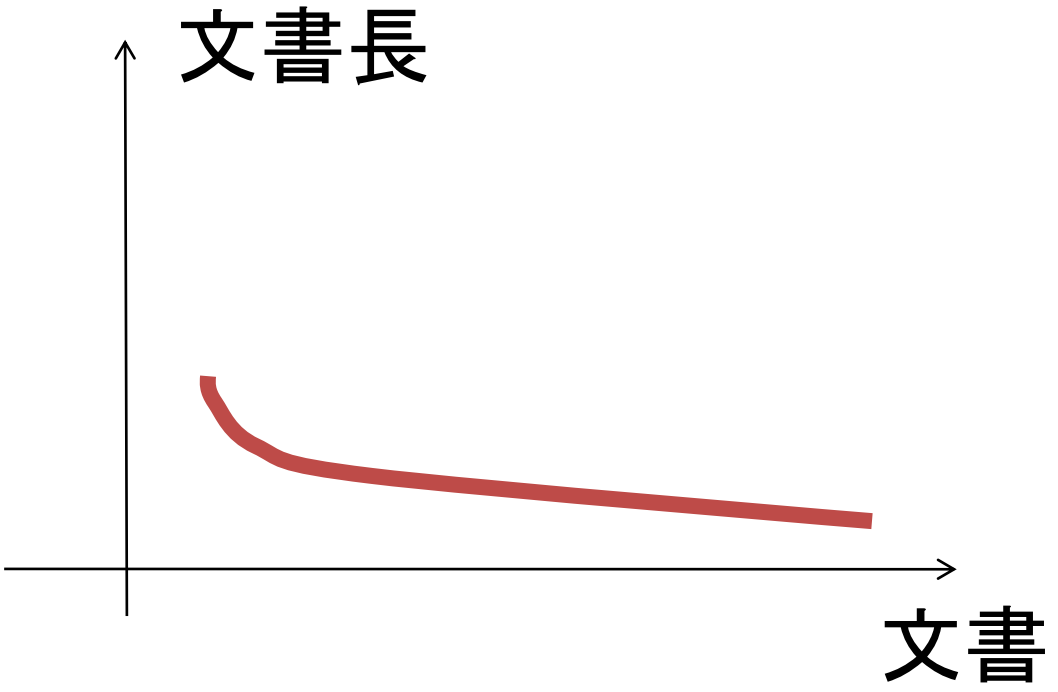
[D.Blei+, 2001]

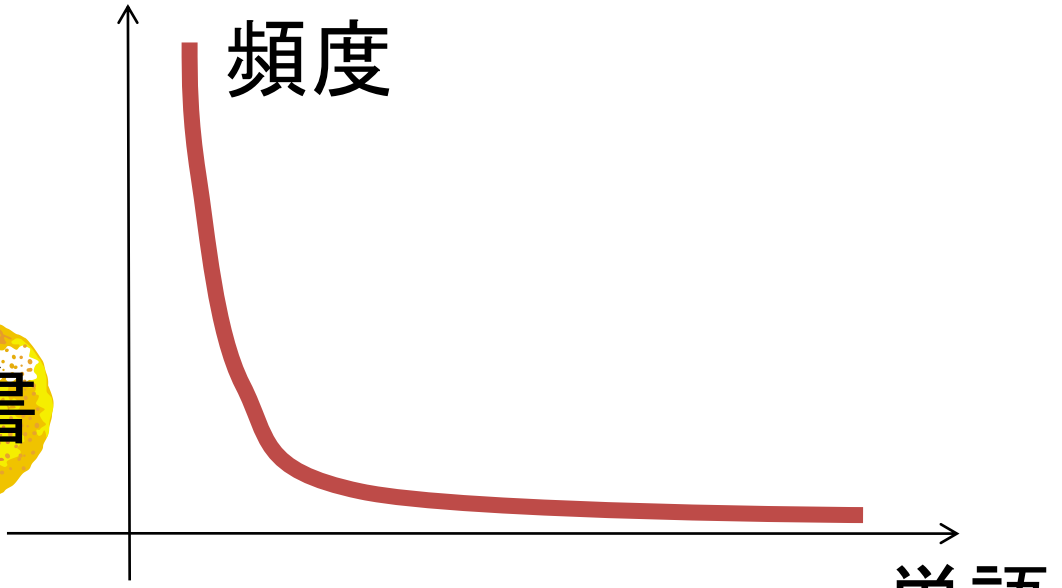


文書

Bag of Words







※この偏りを利用

- Notation-

$w_{d,i}$ 文書dのi番目の単語

$z_{d,i}$ 文書dのi番目の単語の潜在変数

$n_{t,v}$ トピックtから単語vが生成された回数

$n_{d,t}$ 文書dでトピックtが生成された回数

V 語彙数　　T トピック数　　D 文書数

$$n_t = \sum_{v=1} n_{t,v} \quad n_d = \sum_{t=1} n_{d,t} \quad n = \sum_{d=1} n_d$$

トピック出現数　　文書長　　総単語数

Gibbs sampler for LDA

For d=1,…,D

For i=1,…,Nd

$$z_{d,i} \sim p(z_{d,i} = t | w_{d,i} = v, W^{-d,i}, Z^{-d,i})$$

$$\propto \frac{n_{t,v}^{-d,i} + \beta}{n_t^{-d,i} + V * \beta} (n_{d,t}^{-d,i} + \alpha)$$

※必要なサイズ:2*n+T*(V+D)

目標: 情報論的下限へ

Bag of Wordsの簡潔な表現

文書dが語彙{1,..., V}の中の
n_d 個の要素を含んでいるとする

dに対する問い合わせを考えると
dのデータ構造のサイズの
情報理論的下限は

$$n_d \log_2(\frac{eV}{n_d})$$

文書全体での情報理論的下限

$$\sum_{d=1}^{D} n_d \log_2(\frac{eV}{n_d})$$

$$= n \log_2(\frac{eV}{n}) - n \sum_{d=1}^{D} \frac{n_d}{n} \log(\frac{n_d}{n})$$

$$= n \log_2(\frac{eV}{n}) + nH(\frac{n_d}{n})$$

の場合:

$$\sum_{d=1}^{D} n_d \log_2(\frac{eV}{n_d}) = n \log_2(\frac{eV}{n}) + nH(\frac{n_d}{n})$$

の場合:

$$\sum_{v=1}^{V} n_v \log_2(\frac{eD}{n_v}) = n \log_2(\frac{eD}{n}) + nH(\frac{n_v}{n})$$

	D<<V	D<V	D>V
APコーパス	D=1万, n=約224万, V=約6.7万	D=10万, n=約2217万, V=約14,3万	D=約24万, N=約5425万, V17.8万
H(nd/n)	13.08	16.40	17.68
H(nv/n)	12.56	12.60	12.63
通常MB(文書毎)	4.67	49.08	122.10
通常MB(単語毎)	3.93	47.71	125.00
下限MB(文書毎)	2.63	27.93	70.55
下限MB(単語毎)	1.66	16.52	40.54

D>Vでは,通常MB(文書毎)<通常MB(単語毎)となってしまうが、圧縮することで
D>Vでも,下限MB(文書毎)>下限MB(単語毎)となる。

まとめと今後の課題

2*n + T*(V+D)

今回の話　今後の課題

-T*Vは容易に可能-

トピック毎に単語を格納

$$\sum_{t=1}^{T} n_t \log_2(\frac{eV}{n_t}) = n \log_2(\frac{eV}{n}) + nH(\frac{n_t}{n})$$

単語毎にトピックを格納

$$\sum_{v=1}^{V} n_v \log_2(\frac{eT}{n_v}) = n \log_2(\frac{eT}{n}) + nH(\frac{n_v}{n})$$