Big Data Analytics & Applications

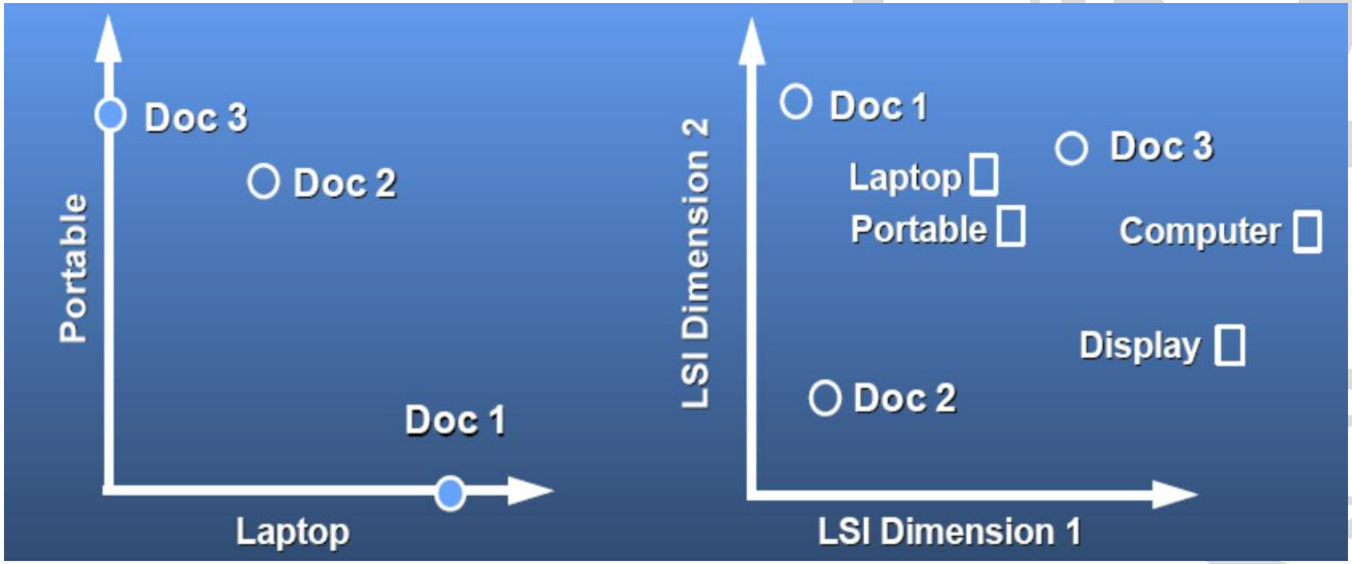
Bin Li

School of Computer Science

Fudan University

Latent Semantic Analysis

- After applying SVD to the word-document co-occurrence matrix and obtain the factorization $A = USV^\top$
  - U: similar words have large inner products
  - V: similar documents have large inner products
  - Related word and document have large inner products



Latent Semantic Analysis

- LSA applies singular value decomposition (SVD) to find latent concepts $A = USV^\top$
  - A: $m \times n$ word-document co-occurrence matrix
  - U: $m \times k$ orthogonal matrices for representing words
  - V: $n \times k$ orthogonal matrices for representing documents
  - S: $k \times k$ diagonal singular value matrix
  - Select $k' \ll n, k' \ll m$ for a low-rank approximation of A

A = U x S x Vt

A:

	d1	d2	d3	d4
a	6	7	1	0
b	8	6	0	1
c	6	9	8	5
d	0	1	8	8
e	2	0	9	7
f	2	0	7	7

U:

	f1	f2	f3	f4
a	0.24	-0.51	0.08	0.06
b	0.25	-0.54	-0.64	-0.23
c	0.58	-0.28	0.57	0.13
d	0.42	0.37	0.16	-0.68
e	0.44	0.34	-0.24	0.66
f	0.39	0.29	-0.40	-0.09

S:

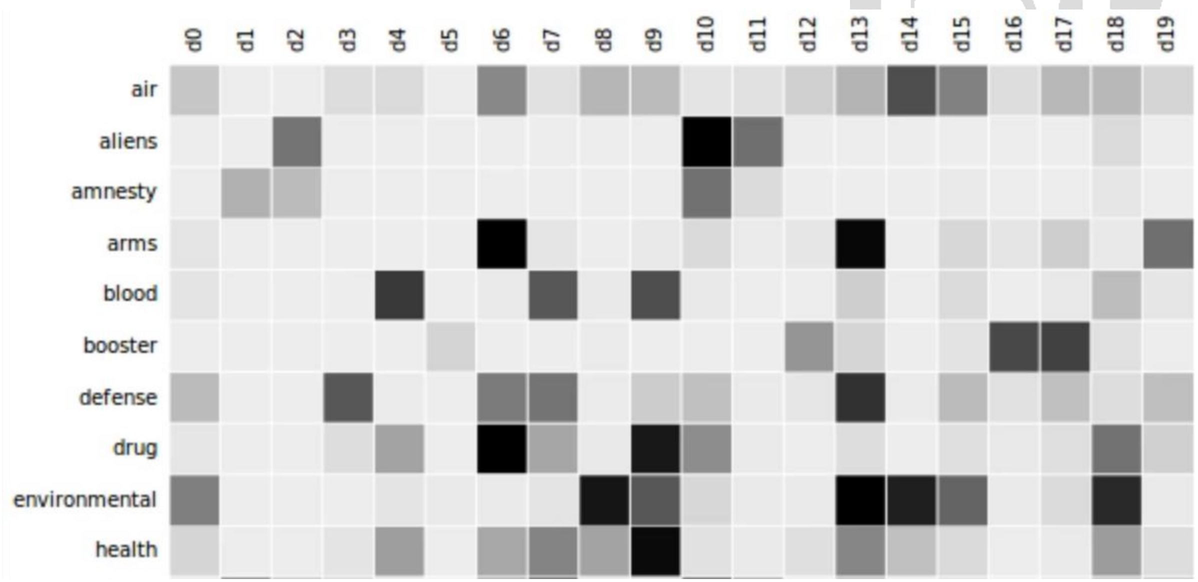
	f1	f2	f3	f4
f1	23.1	0	0	0
f2	0	14.3	0	0
f3	0	0	3.5	0
f4	0	0	0	1.5

Vt:

	d1	d2	d3	d4
f1	0.37	0.38	0.65	0.53
f2	-0.55	-0.63	0.37	0.38
f3	-0.69	0.59	0.27	-0.21
f4	0.26	-0.29	0.59	-0.69

Probabilistic LSA

- Probabilistic LSA (PLSA) is a statistical technique for the analysis of co-occurrence matrix.
- Compared to standard LSA stemming from a low-rank decomposition (SVD), PLSA is based on a mixture decomposition derived from a latent class model



PLSA Model

- Observations in the form of co-occurrences (w, d) of words and documents
- PLSA models the probability of (w, d) as a mixture of conditionally independent multinomial distributions

$$p(w, d) = \sum_z p(z)p(d|z)p(w|z) = p(d) \sum_z p(w|z)p(z|d)$$

- An advantage of PLSA is that the latent variable z can be interpreted as a topic
  - z: topic (latent class)
  - $p(z|d)$: each document has a distribution over K latent topics
  - p$(w|z)$: each topic has a distribution over the vocabulary

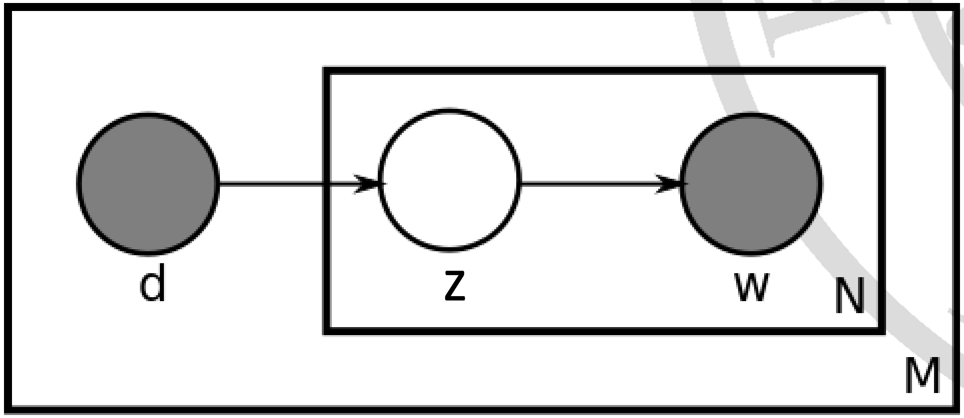
[1] Hofmann (1999). "Probabilistic Latent Semantic Indexing".

PLSA Model

- PLSA is a generative model of the documents in the collection it is estimated on

$$p(w,d) = p(d)\sum_z p(w|z)p(z|d)$$

  - For each document d, a topic z is generated conditionally to d according to $p(z|d)$
  - A word is then generated from topic z according to $P(w|z)$



EM Algorithm for Latent Variable Models

- Given a joint distribution $p(X, Z|\Theta)$ over observed variables X and latent variables Z, governed by parameters Θ, the goal is to maximize the likelihood function $p(X|\Theta)$ w.r.t. Θ.
- The general EM algorithm:
  - Initialize the parameters Θ^{old};
  - E-Step: Evaluate $p(Z|X, \Theta^{\text{old}})$;
  - M-Step: Evaluate Θ^{new} given by

$$\Theta^{\text{new}} = \underset{\Theta}{\operatorname{argmax}} \sum_{Z} p(Z|X, \Theta^{\text{old}}) p(X, Z|\Theta)$$

  - Check the convergence of the parameter values; if not convergence condition not satisfied set $\Theta^{\text{old}} = \Theta^{\text{new}}$ and go to E-step.

Learning for PLSA

- The parameters $p(z|d)$ and $p(w|z)$ of PLSA can be learned by using the EM algorithm
- EM algorithm for PLSA:
  - E-Step: Evaluate $p\left(z_k|\, d_i, w_j; \Theta^{\text{old}}\right)$;

$$p\left(z_k|d_i, w_j; \Theta^{\text{old}}\right) = \frac{p(w_j|z_k)p(z_k|d_i)}{\sum_{l=1}^{K} p(w_j|z_l)p(z_l|d_i)}$$

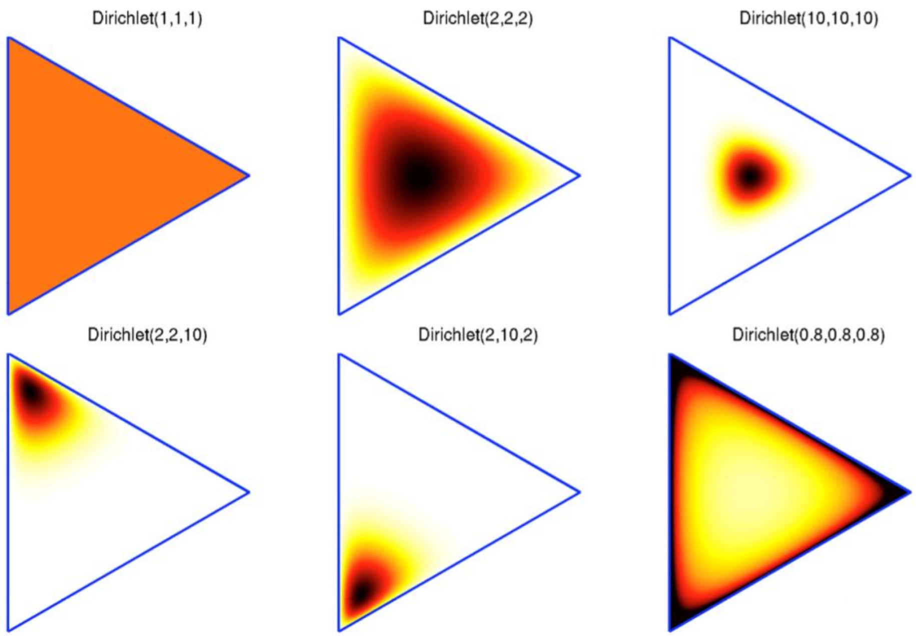
  - M-Step: Evaluate Θ^{new} given by

$$p\left(w_j|z_k\right) = \frac{\sum_{i=1}^{M} p(d_i, w_j)p(z_k|d_i, w_j)}{\sum_{n=1}^{N}\sum_{m=1}^{M} p(d_m, w_n)p(z_k|d_m, w_n)} = \frac{\sum_{i=1}^{M} \#(d_i, w_j)p(z_k|d_i, w_j)}{\sum_{n=1}^{N}\sum_{m=1}^{M} \#(d_m, w_n)p(z_k|d_m, w_n)}$$

$$p(z_k|d_i) = \sum_{j=1}^{N} p(w_j|d_i)p(z_k|d_i, w_j) = \frac{\sum_{j=1}^{N} \#(d_i, w_j)p(z_k|d_i, w_j)}{\#(d_i)}$$

Latent Dirichlet Allocation

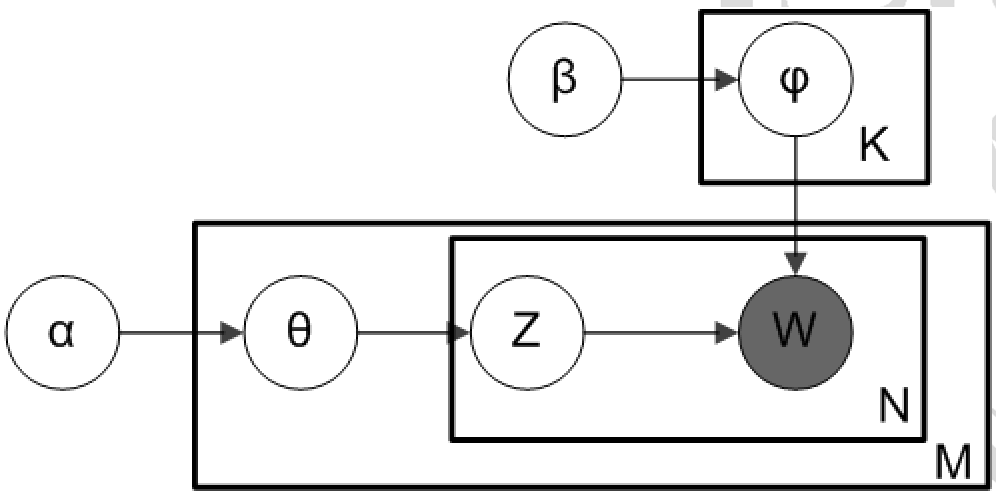
- PLSA is not a generative model of new documents
- LDA is identical to PLSA except that in LDA
  - Document-topic distribution ← sparse Dirichlet prior
  - Topic-word distribution ← sparse Dirichlet prior



[1] Blei et al. (2003). "Latent Dirichlet Allocation".

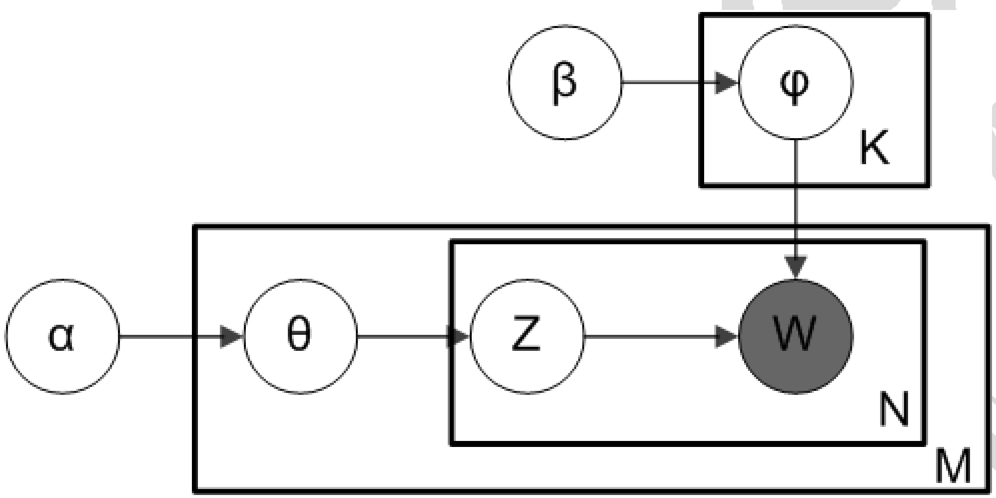
Latent Dirichlet Allocation

- θ and φ are matrices through decomposing the document-word co-occurrence matrix
  - θ: $N \times K$ matrix for distribution for a document over topics
  - φ: $K \times M$ matrix for distribution for a topic over words
  - α and β are fixed hyper-parameters
  - z is latent variable



Generative Process of LDA

- Sample $\theta_i \sim Dir(\alpha)$ for each document (typically $\alpha < 1$)
- Sample $\varphi_k \sim Dir(\beta)$ for each topic (typically $\beta < 1$)
- For each of the word positions (i, j)
  - Sample a topic $z_{i,j} \sim Multinomial(\theta_i)$
  - Sample a word $w_{i,j} \sim Multinomial(\varphi_{z_{i,j}})$



Properties of Dirichlet

- Dirichlet distribution is the conjugate prior of the multinomial distribution

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$Multi(m_1, \ldots, m_K|\mu, N) = \frac{n!}{m_1!\cdots m_K!} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\begin{aligned} p(\mu|D,\alpha) &= Dir(\mu|\alpha + m) \\ &= \frac{\Gamma(\alpha_1 + \cdots + \alpha_K + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \end{aligned}$$

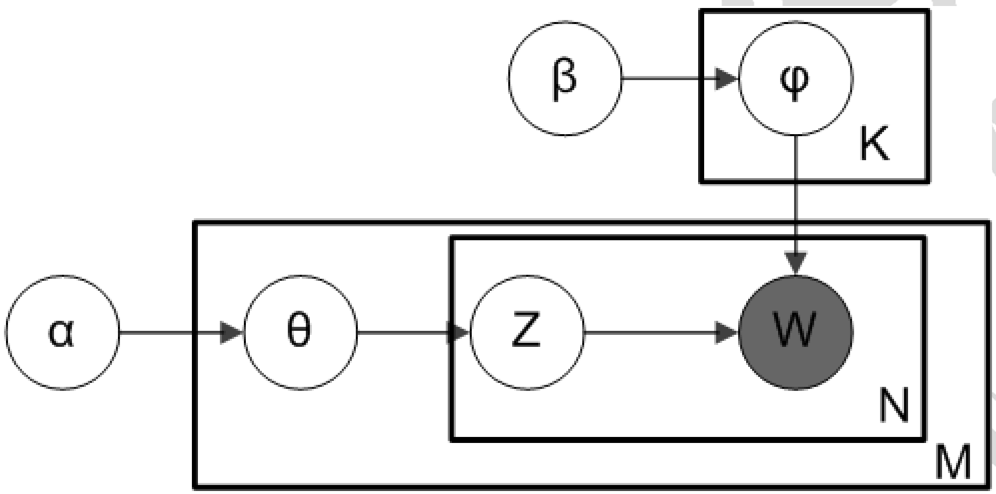
- The expectation of Dirichlet is $E(\mu_k) = \dfrac{\alpha_k}{\alpha_1 + \cdots + \alpha_K}$

Gibbs Sampling

- Suppose we want to obtain k samples of $(x_1, \dots, x_n)$ from a joint distribution $p(x_1, \dots, x_n)$, we can sample x_i in order in each iteration $p(x_i^{(t+1)} | x_1^{(t+1)}, \dots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \dots, x_n^{(t)})$
- Gibbs sampling
  - Sample x_1 conditioned on x_2, x_3
  - Sample x_2 conditioned on x_1, x_3
  - Sample x_3 conditioned on x_1, x_2
- Collapsed Gibbs sampling
  - Sample x_1 conditioned on x_3
  - Sample x_3 conditioned on x_1
  - x_2 is collapsed out during the sampling process

Collapsed Gibbs Sampling for LDA

- In the generative process of LDA, latent variable $z_{d,j}$ is used to choose a topic for the jth word of the dth document
  - If we know $\{z_{d,j}\}$ it is easy to estimate $\{\theta_d\}$ and $\{\varphi_k\}$
  - If we can integrate out $\{\theta_d\}$ and $\{\varphi_k\}$ the problem can be simplified to estimate $p(z_{d,j} = k|\{z_{-(d,j)}\},\{w_{d,j}\})$



Collapsed Gibbs Sampling for LDA

- By integrating out $\{\theta_d\}$ and $\{\varphi_k\}$ the Gibbs sampling procedure boils down to estimate

$$p(z_i = k|\{z_{-i}\},\{w_i\}) = \frac{p(z_i = k,\{z_{-i}\},\{w_i\})}{p(\{z_{-i}\},\{w_i\})}$$

$$p(z_i = k|\{z_{-i}\},\{w_i\}) \propto p(z_i = k,\{z_{-i}\},\{w_i\})$$

$$\propto p(w_i|z_i = k,\{z_{-i}\},\{w_{-i}\})p(z_i = k|\{z_{-i}\},\{w_{-i}\})$$

$$= p(w_i|z_i = k,\{z_{-i}\},\{w_{-i}\})p(z_i = k|\{z_{-i}\})$$

  - The first term is the likelihood
  - The second term is like a prior

Collapsed Gibbs Sampling for LDA

- Look at the first term in $p(z_i = k|\{z_{-i}\},\{w_i\}) \propto p(w_i|z_i = k,\{z_{-i}\},\{w_{-i}\})p(z_i = k|\{z_{-i}\})$

$$p(w_i|z_i = k,\{z_{-i}\},\{w_{-i}\}) = \int p(w_i|z_i = k,\varphi_k)p(\varphi_k|\{z_{-i}\},\{w_{-i}\})d\varphi_k$$

$$= \int \varphi_{k,w_i}\, p(\varphi_k|\{z_{-i}\},\{w_{-i}\})d\varphi_k$$

where

$$p(\varphi_k|\{z_{-i}\},\{w_{-i}\}) \propto p(\{w_{-i}\}\,|\varphi_k,\{z_{-i}\})p(\varphi_k\,) \sim Dir(\#_{-i,k}^{(w_i)} + \beta)$$

By using the property of expectation of Dirichlet distribution

$$p(w_i|z_i = k,\{z_{-i}\},\{w_{-i}\}) = \frac{\#_{-i,k}^{(w_i)} + \beta}{\#_{-i,k} + W\beta}$$

Collapsed Sampling for LDA

- Look at the second term in $p(z_i = k|\{z_{-i}\},\{w_i\}) \propto p(w_i|z_i = k,\{z_{-i}\},\{w_{-i}\})p(z_i = k|\{z_{-i}\})$

$$p(z_i = k|\{z_{-i}\}) = \int p(z_i = k|\theta_d)p(\theta_d|\{z_{-i}\})d\theta_d$$
$$= \int \theta_{d,z_i}\, p(\theta_d|\{z_{-i}\})d\theta_d$$

where

$$p(\theta_d|\{z_{-i}\}) \propto p(\{z_{-i}\}|\theta_d)p(\theta_d) \sim Dir(\#_{-i,k}^{(d)} + \alpha)$$

By using the property of expectation of Dirichlet distribution

$$p(z_i = k|\{z_{-i}\}) = \frac{\#_{-i,k}^{(d)} + \alpha}{\#_{-i,*}^{(d)} + K\alpha}$$

Collapsed Gibbs Sampling for LDA

- Recall that the Gibbs sampling for LDA

$$p(z_i = k|\{z_{-i}\},\{w_i\}) \propto p(w_i|z_i = k,\{z_{-i}\},\{w_{-i}\}) \times p(z_i = k|\{z_{-i}\})$$

$$\propto \frac{\#_{-i,k}^{(w_i)} + \beta}{\#_{-i,k}^{(*)} + W\beta} \times \frac{\#_{-i,k}^{(d)} + \alpha}{\#_{-i,*}^{(d)} + K\alpha}$$

- Now the problem has been simplified to count
  - $\#_{-i,k}^{(w_i)}$: number of w_i apperaring in topic k
  - $\#_{-i,k}^{(*)}$: number of words in topic k
  - $\#_{-i,k}^{(d)}$: number of words of document d in topic k
  - $\#_{-i,*}^{(d)}$: number of words of document d (constant)

PLSA vs LDA

- PLSA

$$p(w_j|z_k) = \frac{\sum_{i=1}^{M} \#(d_i, w_j) p(z_k|d_i, w_j)}{\sum_{n=1}^{N} \sum_{m=1}^{M} \#(d_m, w_n) p(z_k|d_m, w_n)}$$

$$p(z_k|d_i) = \frac{\sum_{j=1}^{N} \#(d_i, w_j) p(z_k|d_i, w_j)}{\#(d_i)}$$

- LDA

$$p(z_i = k|\{z_{-i}\}, \{w_i\}) \propto p(w_i|z_i = k, \{z_{-i}\}, \{w_{-i}\}) \times p(z_i = k|\{z_{-i}\})$$

$$\propto \frac{\#_{-i,k}^{(w_i)} + \beta}{\#_{-i,k}^{(*)} + W\beta} \times \frac{\#_{-i,k}^{(d)} + \alpha}{\#_{-i,*}^{(d)} + K\alpha}$$

Topic Visualization

- How to interpret the following topic visualization?

$$\varphi_{k,w_i} = \frac{\#_{*,k}^{(w_i)} + \beta}{\#_{*,k}^{(*)} + W\beta}$$

Project: Topic Modeling

- Dataset:
  - Public available topic modeling datasets (e.g., https://www.kaggle.com/jaykrishna/topic-modeling-enron-email-dataset)
  - Or document data collected by yourself
- Method:
  - Use Probabilistic LSA or LDA with Collapsed Gibbs Sampling for topic modeling
- Experiments:
  - Obtain the topic modeling results and visualize the topics using word clouds
  - And discuss the observations from the visualization

Thanks

Email: libin@fudan.edu.cn