Latent Dirichlet Allocation

吴小宝
清华大学软件学院

目录

- 基础知识
- 生成过程
- 求解过程
- 评估工具
- 参考文献和工具

Bayes Theorem

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

后验(posterior) ∝ 似然(likelihood) x 先验(priori)

概率分布

- 伯努利分布(Bernouli)-硬币

$$f_X(x) = p^x (1-p)^{1-x} = \begin{cases} p & \text{if } x = 1, \\ q & \text{if } x = 0. \end{cases}$$

- 二项分布(Binomial)-n重伯努利试验

$$P(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- 多项分布(Multinomial)-骰子

$$P(x_1, x_2, \ldots, x_k; n, p_1, p_2, \ldots, p_k) = \frac{n!}{x_1! \ldots x_k!} {p_1}^{x_1} \ldots {p_k}^{x_k}$$

概率分布

- Beta分布

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$\frac{1}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

- Dirichlet分布-Beta分布的推广

$$f(x_1, x_2, \ldots, x_k; \alpha_1, \alpha_2, \ldots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} {x_i}^{\alpha^i - 1}$$

$$0 \le x_i \le 1 \text{ and } \sum_i x_i = 1$$

概率分布

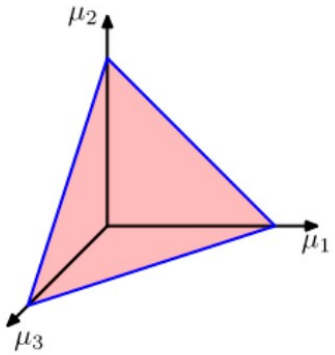


图 2.4: 三个变量μ_1, μ_2, μ_3上的狄利克雷分布被限制在一个单纯形中，如图所示。这是由于限制条件$0 \leq \mu_k \leq 1$和$\sum_k \mu_k = 1$的存在所造成的。

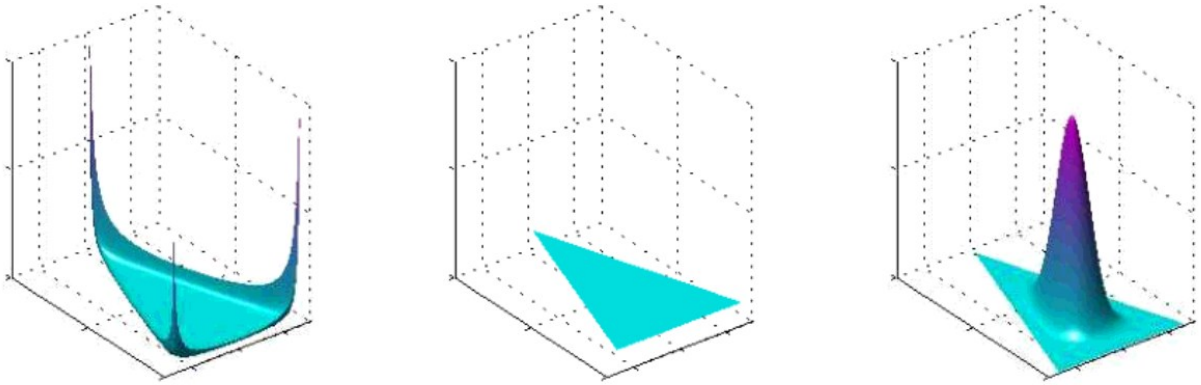


图 2.5: 三个变量上的狄利克雷分布的图像，其中两个水平轴是单纯形平面上的坐标轴，垂直轴对应于概率密度的值。这里$\{\alpha_k\} = 0.1$对应于左图，$\{\alpha_k\} = 1$对应于中图，$\{\alpha_k\} = 10$对应于右图。

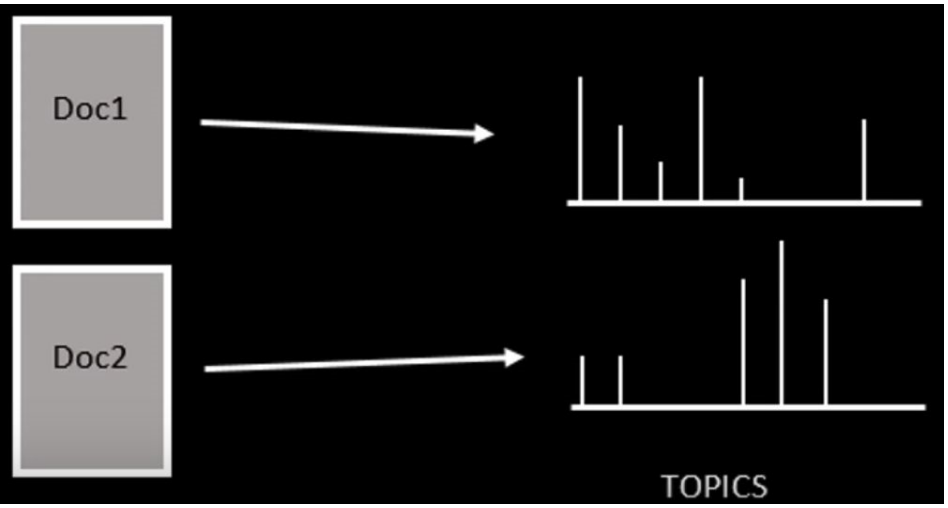
- Pattern Recognition and Machine Learning

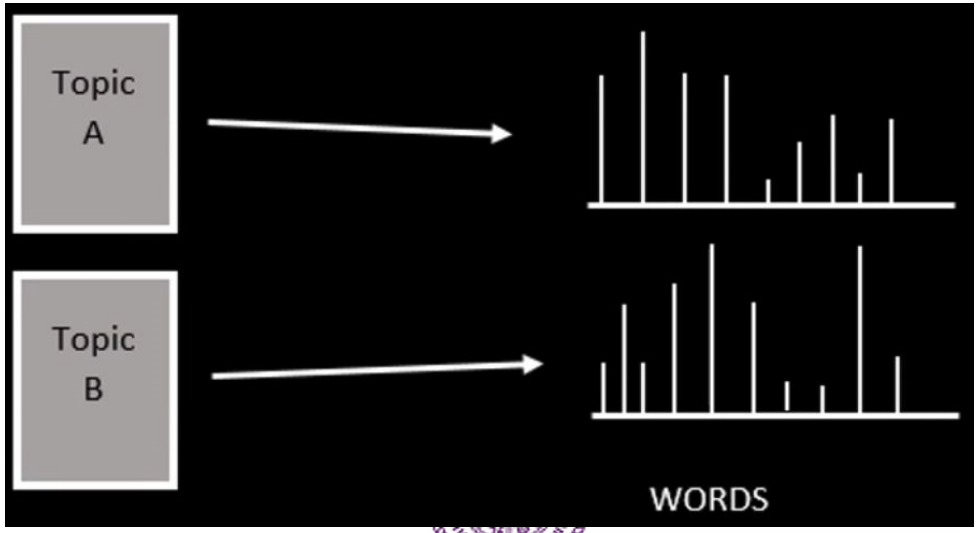
概率分布

- 共轭分布 Conjugate Distribution
  - 后验$p(\theta|x)$ 和先验 $p(\theta)$ 具有共同的相同的概率形式
  - 先验分布被称为似然函数的共轭先验
  - Beta分布是Binomial分布的共轭先验分布
  - Dirichlet分布是Multinomial分布的共轭先验分布

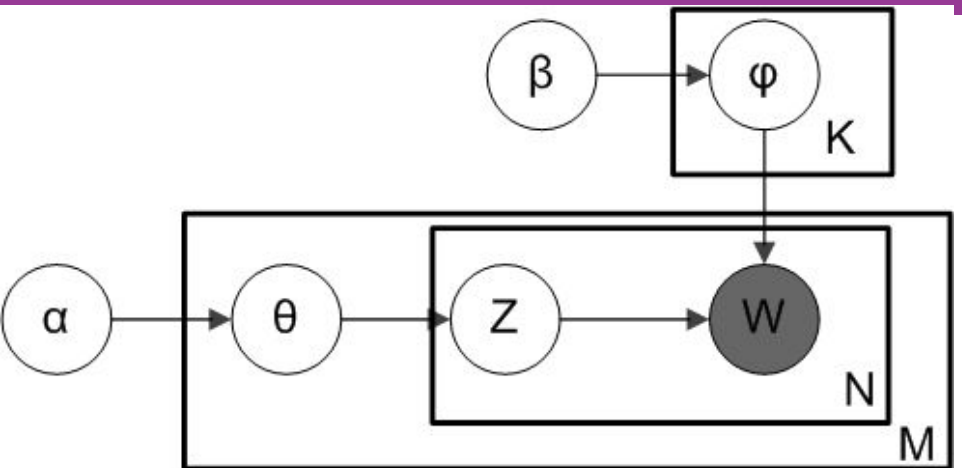
LDA

- Assumptions
  - 词袋(bag-of-words): 顺序无关
  - 文档是潜在主题的概率分布
  - 主题是单词的概率分布





概率图模型表示



- For each document i , draw $\theta_i \sim Dirichlet(\alpha), d = 1 \dots D$
- For each word j in document i :
  - Sample from θ_i , draw a topic index $z_{ij} \sim Multinomial(\theta_i)$
  - Sample from $\varphi_{z_{ij}}$, draw the observed word $w_{ij} \sim Multinomial(\varphi_{z_{ij}})$

- M ：语料库中文档的数量
- N ：文档中单词的数量
- K ：语料库中主题的数量
- α ：每篇文档的主题分布 (先验狄利克雷分布) 的参数，是一个 K 维向量，K 是主题数量
  - 越高，代表每篇文档可能包含更多的主题，而不只是包含一个或者两个特定的主题
  - 越低，代表每篇文档包含的主题数越少
- θ_i ：文档 i 的主题分布
- z_{ij} ：文档 i 中第 j 个单词的主题编号，服从多项式分布
- β ：每个主题的单词分布 (先验狄利克雷分布) 的参数，是一个 V 维向量，V 是文档中单词的数量
  - 越高，代表每个主题可能包含更多的单词
  - 越低，代表每个主题包含的单词数越少
- $\varphi_{z_{ij}}$ ：主题 z_{ij} 的单词分布
- w_{ij} ：特定的单词，服从多项式分布

共轭分布

$$p(\vec{\theta}_m|\vec{z}_m) = \frac{p(\vec{z}_m|\vec{\theta}_m)p(\vec{\theta}_m|\vec{\alpha})}{p(\vec{z}_m|\vec{\alpha})}$$

$$p(\vec{z}_m|\vec{\theta}_m)p(\vec{\theta}_m|\vec{\alpha}) = \prod_{k=1}^{K}\theta_k^{n_{m,k}}\frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$$

$$= \frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}\theta_k^{n_{m,k}+a_k-1}$$

$$p(\vec{z}_m|\vec{\alpha}) = \int \frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}\theta_k^{n_{m,k}+a_k-1}d\vec{\theta}_m$$

$$= \frac{1}{\Delta(\vec{\alpha})}\int \prod_{k=1}^{K}\theta_k^{n_{m,k}+a_k-1}d\vec{\theta}_m$$

$$= \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})}$$

共轭分布

$$p(\vec{\theta}_m|\vec{z}_m) = \frac{p(\vec{z}_m|\vec{\theta}_m)p(\vec{\theta}_m|\vec{\alpha})}{p(\vec{z}_m|\vec{\alpha})}$$

$$= \frac{1}{\Delta(\vec{n}_m + \vec{\alpha})} \prod_{k=1}^{K} \theta_k^{n_{m,k}+a_k-1}$$

$$= \mathrm{Dir}(\vec{n}_m + \vec{\alpha})$$

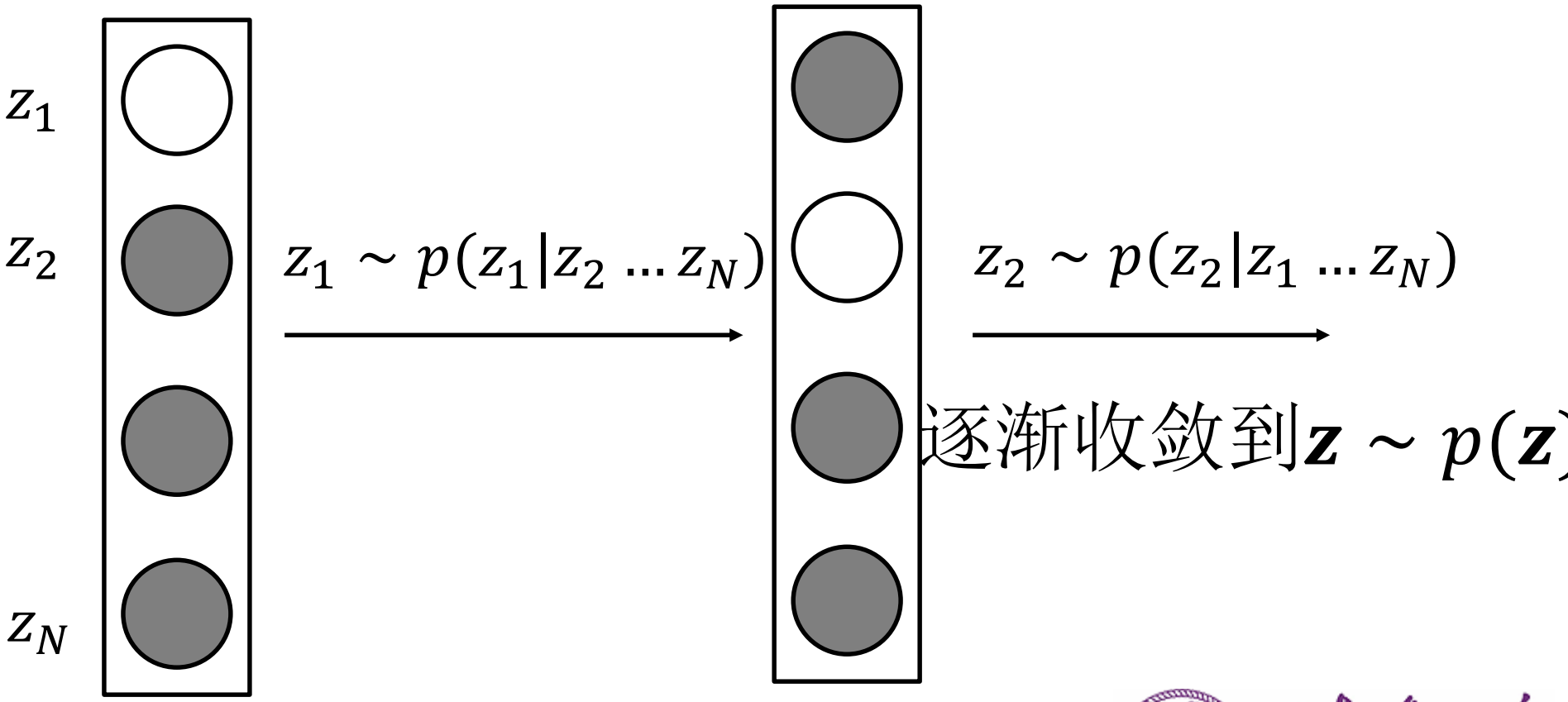
$$p(\vec{\phi}_k|\vec{w}) = \mathrm{Dir}(\vec{n}_k + \vec{\beta})$$

求解过程

$$P(z, \theta, \varphi | w, \alpha, \beta) = \frac{P(z, \theta, \varphi | \alpha, \beta)}{P(w | \alpha, \beta)}$$

Gibbs Sampling

- 一种Markov Chain Monte Carlo算法



Gibbs Sampling

$$p(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta}) = p(\vec{w}|\vec{z}, \vec{\beta})p(\vec{z}|\vec{\alpha})$$

$$p(\vec{w}|\vec{z}, \vec{\beta}) = \int p(\vec{w}|\vec{z}, \Phi)p(\Phi|\vec{\beta})d\Phi$$

$$= \int \prod_{z=1}^{K} \frac{1}{\Delta(\vec{\beta})} \prod_{t=1}^{V} \phi_{z,t}^{n_z^{(t)}+\beta_t-1} d\vec{\phi}_z$$

$$= \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{\beta})}$$

Gibbs Sampling

$$p(\vec{z}|\vec{\alpha}) = \int p(\vec{z}|\Theta)p(\Theta|\vec{\alpha})d\Theta$$

$$= \int \prod_{m=1}^{M} \frac{1}{\Delta(\alpha)} \prod_{k=1}^{K} \theta_{m,k}^{n_m^{(k)}+\alpha_k-1} d\vec{\theta}_m$$

$$= \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

Gibbs Sampling

$$p(\vec{z}, \vec{w}|\vec{\alpha}, \vec{\beta}) = p(\vec{w}|\vec{z}, \vec{\beta})p(\vec{z}|\vec{\alpha})$$

$$= \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{\beta})} \cdot \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

$\vec{n}_z = \left\{n_z^{(t)}\right\}_{t=1}^{V}$ 词t被赋予主题z的次数

$\vec{n}_{\mathrm{m}} = \left\{n_m^{(k)}\right\}_{k=1}^{K}$ 文档m中主题k被赋予的次数

Gibbs Sampling

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) = \frac{p(\vec{w}, \vec{z})}{p(\vec{w}, \vec{z}_{\neg i})} \tag{19}$$

$$= \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})p(w_i)} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})} \tag{20}$$

$$\propto \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})} \tag{21}$$

$$= \frac{\Gamma(n_k^{(t)} + \beta_t)\,\Gamma(\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t)}{\Gamma(n_{k,\neg i}^{(t)} + \beta_t)\,\Gamma(\sum_{t=1}^{V} n_k^{(t)} + \beta_t)} \cdot \frac{\Gamma(n_m^k + \alpha_k)\,\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^k + \alpha_k)}{\Gamma(n_{m,\neg i}^k + \alpha_k)\,\Gamma(\sum_{k=1}^{K} n_m^k + \alpha_k)} \tag{22}$$

$$= \frac{\Gamma(n_{k,\neg i}^{(t)} + \beta_t + 1)\,\Gamma(\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t)}{\Gamma(n_{k,\neg i}^{(t)} + \beta_t)\,\Gamma([\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t] + 1)} \cdot \frac{\Gamma(n_{m,\neg i}^k + \alpha_k + 1)\,\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^k + \alpha_k)}{\Gamma(n_{m,\neg i}^k + \alpha_k)\,\Gamma([\sum_{k=1}^{K} n_m^k + \alpha_k] - 1 + 1)} \tag{23}$$

$$= \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \cdot \frac{n_{m,\neg i}^{(t)} + \alpha_k}{[\sum_{k=1}^{K} n_m^k + \alpha_k] - 1} \tag{24}$$

$$\propto \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t}(n_{m,\neg i}^k + \alpha_k) \tag{25}$$

Gibbs Sampling


```
// Gibbs sampling over burn-in period and sampling period
while not finished do
    for all documents m ∈ [1, M] do
        for all words n ∈ [1, N_m] in document m do
            // for the current assignment of k to a term t for word w_{m,n}:
            decrement counts and sums: n_m^(k) −= 1; n_m −= 1; n_k^(t) −= 1; n_k −= 1
            // multinomial sampling acc. to Eq. 78 (decrements from previous step):
            sample topic index k̃ ~ p(z_i | z⃗_¬i, w⃗)
            // for the new assignment of z_{m,n} to the term t for word w_{m,n}:
            increment counts and sums: n_m^(k̃) += 1; n_m += 1; n_k̃^(t) += 1; n_k̃ += 1
```

Gibbs Sampling

$$\vec{\theta}_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}$$

$$\vec{\phi}_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t}$$

Gibbs Sampling

- 词分布矩阵Φ不变

Algorithm 15 LDA Inference

1: 随机初始化：对当前文档中的每个词w，随机的赋一个topic 编号z;
2: 重新扫描当前文档，按照Gibbs Sampling 公式，对每个词w, 重新采样它的topic;
3: 重复以上过程直到Gibbs Sampling 收敛;
4: 统计文档中的topic分布，该分布就是$\vec{\theta}_{new}$

评估工具

- 混乱度Perplexity

$$exp(-\frac{1}{D}\sum_d \frac{1}{N_d} log\ p(w_d))$$

- **Topic Coherence**
  - NPMI
  - CV
- **Topic Diversity**
  - TU

$$\frac{1}{T}\sum_{t}^{T}\frac{1}{cnt(t,z)}$$

  - 只出现过一次的词的比例

参考文献

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research3.Jan (2003): 993-1022.
- Hoffman, Blei, Bach: Online Learning for Latent Dirichlet Allocation, NIPS 2010.
- LDA数学八卦
- Patameters Estimation of Text Analysis

- gensim LDA model
- sklearn LDA
- lda-project/lda

THANKS

清華大學
Tsinghua University