算法交流与学习

Latent Dirichlet Allocation (LDA)

合工大管院电子商务研究所 钱洋

2018-03-20

主要内容

- LDA 应用场景
- LDA 涉及知识
- LDA 生成模型
- LDA模型推理及实现
- LDA模型的扩展

LDA应用场景

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Latent dirichlet allocation

DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org

Abstract We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying

☆ 99 被引用次数：22158 相关文章 所有 113 个版本

LDA抽象出文章所包含的主题分布以及每个主题词分布，这些信息可用于推荐、文档相似性计算、文档信息降维、文档聚类、排序、特征生成等。

David M. Blei, Andrew Y. Ng et al. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003)

主要内容

- LDA 应用场景
- LDA 涉及知识
- LDA 生成模型
- LDA模型推理及实现
- LDA模型的扩展

贝叶斯框架

$$p(\vartheta|X) = \frac{p(X|\vartheta) \cdot p(\vartheta)}{p(X)}$$

贝叶斯定理

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

先验分布+数据知识=后验分布

共轭先验分布

✓ 设θ为总体分布的参数(或参数向量)，$\pi(\theta)$是θ的先验密度函数，假如由抽样信息计算得到的后验概率密度函数与$\pi(\theta)$具有相同的形式，则称$\pi(\theta)$是θ的共轭先验分布。

$$p(\vartheta|X)=\frac{p(X|\vartheta)\cdot p(\vartheta)}{p(X)} \longrightarrow p(\theta|\mathrm{x}) \propto p(x|\theta)\cdot p(\theta)$$

✓ 采用共轭先验的原因：**先验分布和后验分布**的形式相同。

注：共轭先验分布式对某一分布中的参数而言的，如正太均值、正太方差、泊松分布均值等。

共轭先验分布

Bernoulli实验：

$$p(C=c|p) = p^c\,(1-p)^{1-c} \triangleq \mathrm{Bern}(c|p)$$ → 发生的概率为p

先验：$$p(p|\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \triangleq \mathrm{Beta}(p|\alpha,\beta),$$ $$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

计算后验：

$$p(p|C,\alpha,\beta) = \frac{\prod_{i=1}^{N} p(C=c_i|p)\; p(p|\alpha,\beta)}{\int_0^1 \prod_{i=1}^{N} p(C=c_i|p)\; p(p|\alpha,\beta)\,\mathrm{d}p}$$

$$= \frac{p^{n^{(1)}}(1-p)^{n^{(0)}}\;\frac{1}{\mathrm{B}(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{Z}$$

$$= \frac{p^{[n^{(1)}+\alpha]-1}(1-p)^{[n^{(0)}+\beta]-1}}{\mathrm{B}(n^{(1)}+\alpha, n^{(0)}+\beta)}$$

$$= \mathrm{Beta}(p|n^{(1)}+\alpha, n^{(0)}+\beta)$$

共轭先验分布

常见的共轭先验分布：

总体分布	参数	共轭先验分布
二项分布	成功概率	贝塔分布 $Beta(\alpha,\beta)$
泊松分布	均值	伽马分布 $\Gamma(k,\theta)$
指数分布	均值的倒数	伽马分布 $\Gamma(k,\theta)$
正态分布(方差已知)	均值	正态分布 $N(\mu,\sigma^2)$
正态分布(方差未知)	方差	逆伽马分布 $IGa(\alpha,\beta)$

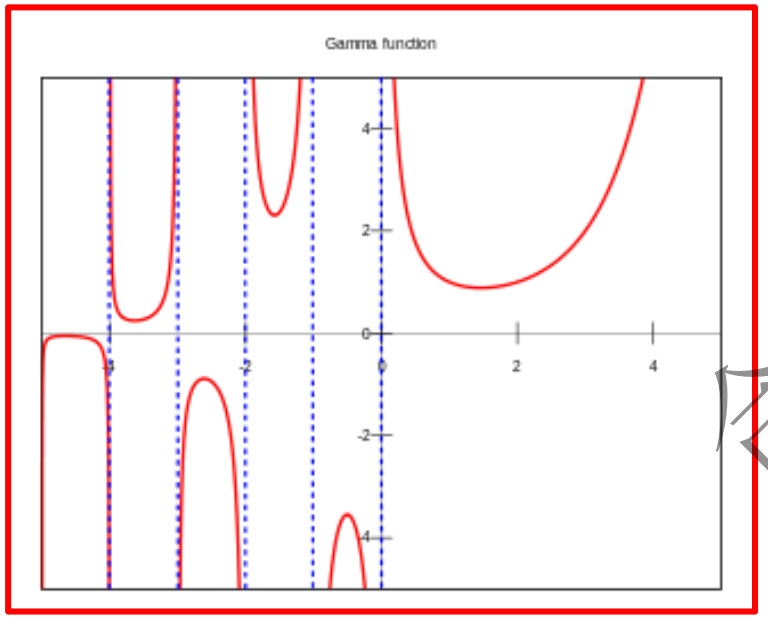
参考：（1）http://blog.csdn.net/qy20115549/article/details/53307535(案例及**beta-Bernoulli process**)

(2)Murphy K P. Conjugate Bayesian analysis of the Gaussian distribution[J]. def, 2007(各种有关正太分布方面的推理)

两个函数

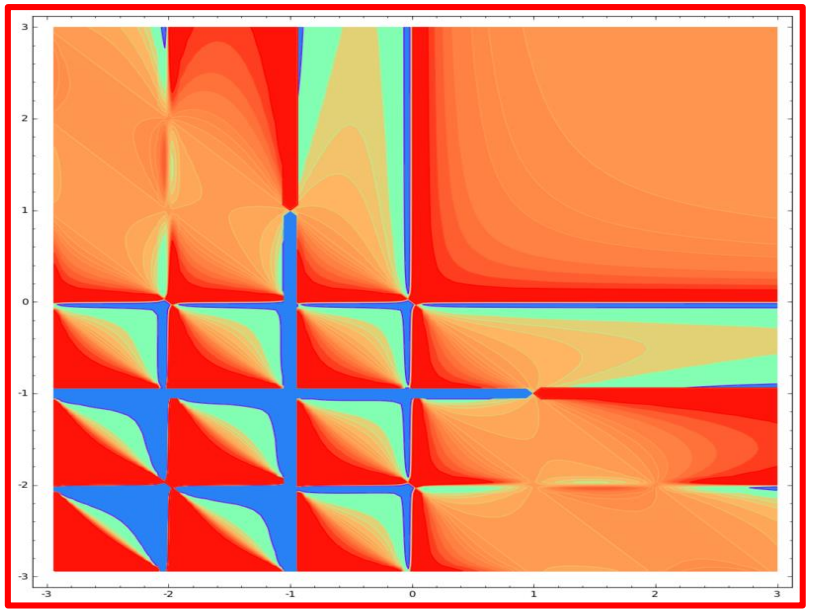
$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

$$\Gamma(x+1) = x\Gamma(x)$$



$$\mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$$

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

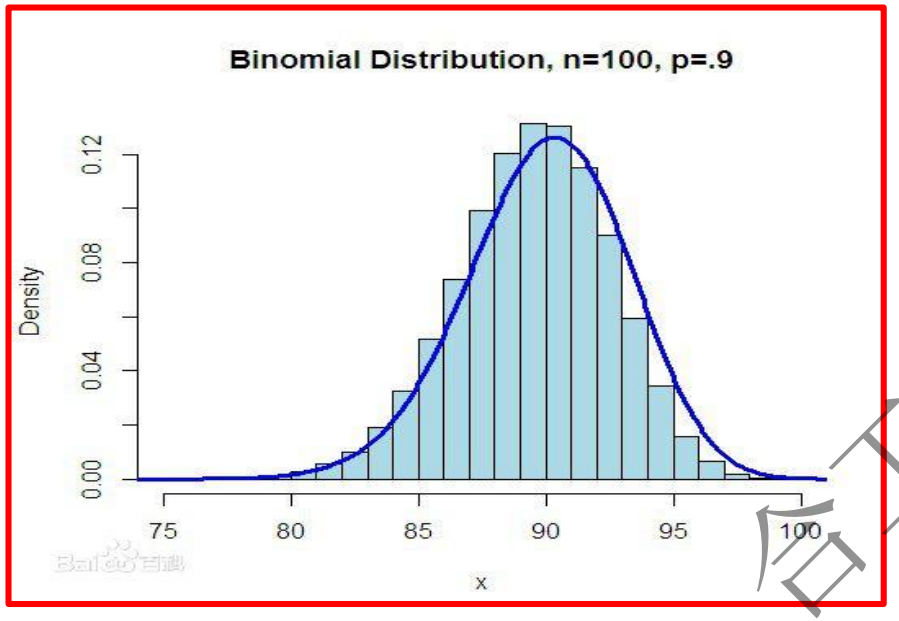


来自：wikipedia

两对分布

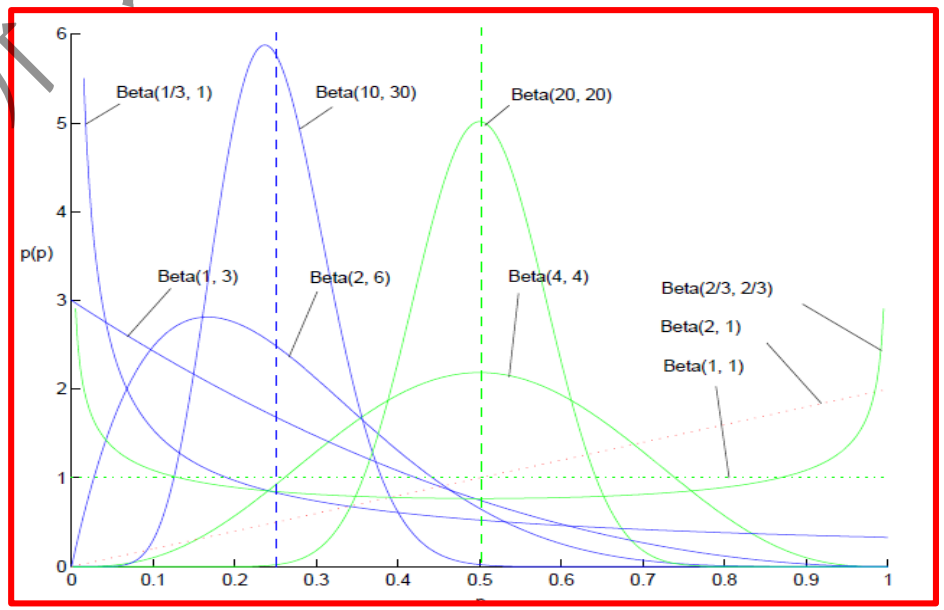
二项分布（n重伯努利分布）：

$$f(x) = P(X = x) = P(X = x|n, p) = C_n^x p^x (1-p)^{n-x}$$



Beta分布：

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



$$Beta(p|\alpha,\beta) + BinomCount(m_1, m_2) = Beta(p|\alpha + m_1, \beta + m_2)$$

两对分布

多项式分布：

$$P(X_1=n_1,\cdots,X_k=n_k)=\begin{cases}\frac{n!}{n_1!\cdots n_k!}p_1^{n_1}\cdots p_k^{n_k} & ,\sum_{i=1}^k n_i=n\\ 0 & ,\text{otherwise}\end{cases}$$

用另一种形式写为：

$$P(X_1=n_1,\cdots,X_k=n_k)=\begin{cases}n!\prod_{i=1}^k\frac{p_i^{n_i}}{n_i!} & ,\sum_{i=1}^k n_i=n\\ 0 & ,\text{otherwise}\end{cases}$$

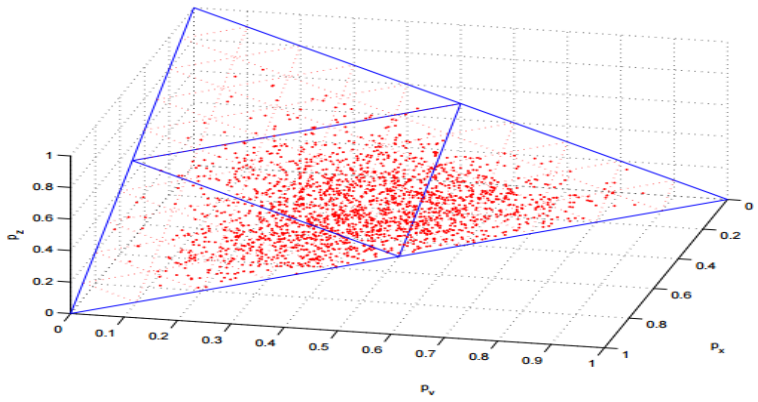
$$Dir(\vec{p}\,|\,\vec{\alpha})+MultCount(\vec{m})=Dir(p|\,\vec{\alpha}+\vec{m})$$

Dirichlet分布：

$$f(x_1,\cdots,x_K;\alpha_1,\cdots,\alpha_K)=\frac{1}{B(\alpha)}\prod_{K}^{i=1}x_i^{\alpha_i-1}$$

其中：$B(\alpha)=\frac{\prod_{i=1}^K\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K\alpha_i)},\alpha=(\alpha_1,\cdots,\alpha_K)$

$Dir(4,4,2)$ 采样2000个点，分布如下：



Beta分布和Dirichlet分布期望

如果$p \sim Beta(t|\alpha,\beta)$, 则

$$
\begin{aligned}
E(p) &= \int_0^1 t * Beta(t|\alpha,\beta)dt \\
&= \int_0^1 t * \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1}dt \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 t^{\alpha}(1-t)^{\beta-1}dt
\end{aligned}
$$

上式右边的积分对应到概率分布$Beta(t|\alpha+1,\beta)$，对于这个分布，我们有

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

把上式带入$E(p)$的计算式，得到

$$
\begin{aligned}
E(p) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}
$$

如果$\vec{p} \sim Dir(\vec{t}|\vec{\alpha})$

$$
E(\vec{p}) = \left(\frac{\alpha_1}{\sum_{i=1}^K \alpha_i}, \frac{\alpha_2}{\sum_{i=1}^K \alpha_i}, \cdots, \frac{\alpha_K}{\sum_{i=1}^K \alpha_i} \right)
$$

LDA及其扩展模型中经常使用

Beta分布和Dirichlet分布参考

参考：

(1) LDA数学八卦

(2) Parameter estimation for text analysis

Code：

(1) [beta分布的采样或抽样(java程序)]

(2) [二项分布的采样或抽样（java实现）]

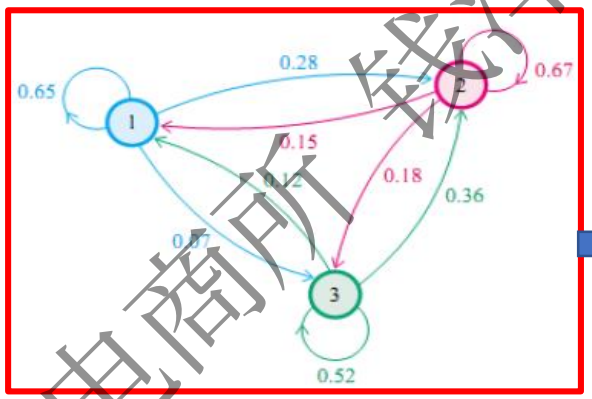
(3) [Gamma函数(伽玛函数)的一阶导数、二阶导数公式推导及java程序]

(4) [多元正态分布的后验采样(包含程序)]

马氏链

马氏链： $P(X_{t+1}=x|X_t,X_{t-1},\cdots)=P(X_{t+1}=x|X_t)$ 状态的转义概率只依赖于上一状态

		子代		
	State	1	2	3
父代	1	0.65	0.28	0.07
	2	0.15	0.67	0.18
	3	0.12	0.36	0.52



子女的收入比例

$\pi_n=\pi_{n-1}P=\pi_0P^n$

初始概率分布

$\pi_0=[0.21,0.68,0.11]$

第n代人	下层	中层	上层
0	0.210	0.680	0.110
1	0.252	0.554	0.194
2	0.270	0.512	0.218
3	0.278	0.497	0.225
4	0.282	0.490	0.226
5	0.285	0.489	0.225
6	0.286	0.489	0.225
7	0.286	0.489	0.225
8	0.289	0.488	0.225
9	0.286	0.489	0.225
10	0.286	0.489	0.225
...	...	...	...

初始概率分布

$\pi_0=[0.75,0.15,0.1]$

第n代人	下层	中层	上层
0	0.75	0.15	0.1
1	0.522	0.347	0.132
2	0.407	0.426	0.167
3	0.349	0.459	0.192
4	0.318	0.475	0.207
5	0.303	0.482	0.215
6	0.295	0.485	0.220
7	0.291	0.487	0.222
8	0.289	0.488	0.225
9	0.286	0.489	0.225
10	0.286	0.489	0.225
...	...	...	...

Gibbs采样

Algorithm 7 二维Gibbs Sampling 算法

1: 随机初始化$X_0 = x_0 Y_0 = y_0$

2: 对$t = 0, 1, 2, \cdots$ 循环采样

1. $y_{t+1} \sim p(y|x_t)$

2. $x_{t+1} \sim p(x|y_{t+1})$

Algorithm 8 n维Gibbs Sampling 算法

1: 随机初始化$\{x_i : i = 1, \cdots, n\}$

2: 对$t = 0, 1, 2, \cdots$ 循环采样

1. $x_1^{(t+1)} \sim p(x_1|x_2^{(t)}, x_3^{(t)}, \cdots, x_n^{(t)})$

2. $x_2^{(t+1)} \sim p(x_2|x_1^{(t+1)}, x_3^{(t)}, \cdots, x_n^{(t)})$

3. $\cdots$

4. $x_j^{(t+1)} \sim p(x_j|x_1^{(t+1)}, \cdots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)}, \cdots, x_n^{(t)})$

5. $\cdots$

6. $x_n^{(t+1)} \sim p(x_n|x_1^{(t+1)}, x_2^t, \cdots, x_{n-1}^{(t+1)})$

Andrieu C, De Freitas N, Doucet A, et al. An introduction to MCMC for machine learning[J]. Machine learning, 2003, 50(1-2): 5-43.

Code: https://github.com/endymecy/MCMC-sampling

主要内容

- LDA 应用场景
- LDA 涉及知识
- LDA 生成模型
- LDA模型推理及实现
- LDA模型的扩展

简单的例子

Gangnam Style

From Wikipedia, the free encyclopedia

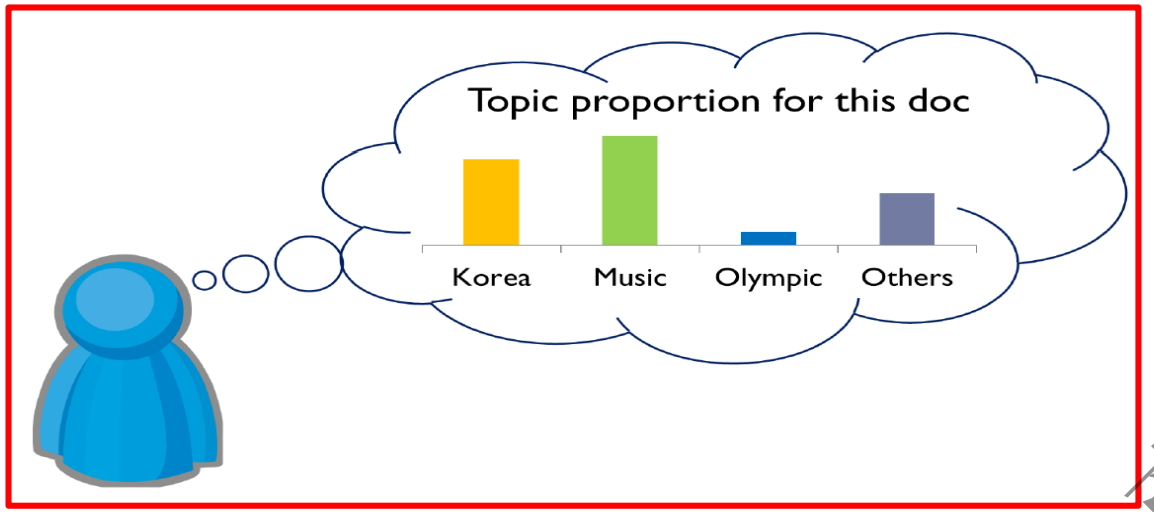
"**Gangnam Style**" (Korean: 강남스타일, IPA: [kaŋnam sɯtʰail]) is a K-pop single by South Korean musician PSY. The song was released on July 15, 2012, as the lead single of his sixth studio album *PSY 6 (Six Rules), Part 1*. "Gangnam Style" debuted at number one on the Gaon Chart, the national record chart of South Korea. As of December 15, 2012, the music video has been viewed over 959 million times on YouTube,[5] and is the site's most watched video after surpassing Justin Bieber's single "Baby".[6][7][8]

The phrase "Gangnam Style" is a Korean neologism that refers to a lifestyle associated with the Gangnam District of Seoul. The song and its accompanying music video went viral in August 2012 and have influenced popular culture since then. "Gangnam Style" received mixed to positive reviews, with praise going to its catchy beat and PSY's amusing dance moves in the music video and during live performances in various locations such as Madison Square Garden, *The Today Show*, *The Ellen DeGeneres Show*, and Samsung commercials.[9] On September 20, 2012, "Gangnam Style" was recognized by *Guinness World Records* as the most "liked" video in YouTube history.[10] It subsequently won Best Video at the MTV Europe Music Awards held later that year.[11]

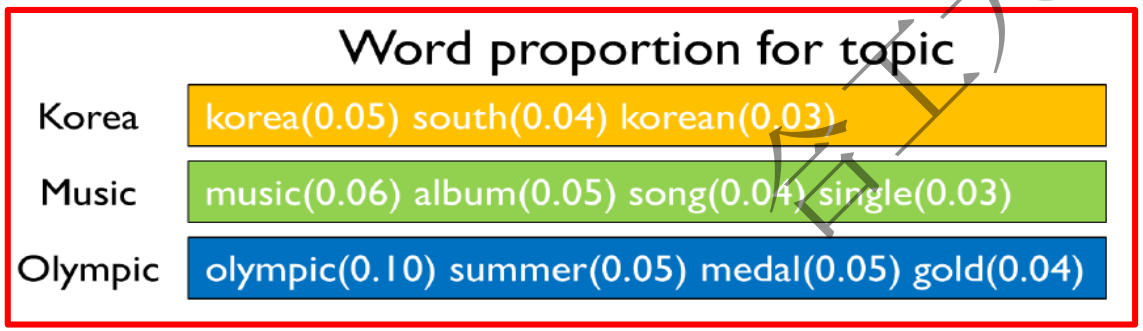
It became a source of parodies and reaction videos by many different groups, organizations, and individuals, including The Oregon Duck, midshipmen from the United States Naval Academy, the North Korean government,[12] and the American space agency NASA, who uploaded a parody shot at its Mission Control Center in Houston, Texas.[13] By the end of 2012, the song had reached the number one position in more than 30 countries including Australia, Canada, France, Germany, Italy, Spain, and the United Kingdom.

한글 漢字 This article contains Korean text. Without proper rendering support, you may see question marks, boxes, or other symbols instead of Hangul or Hanja.

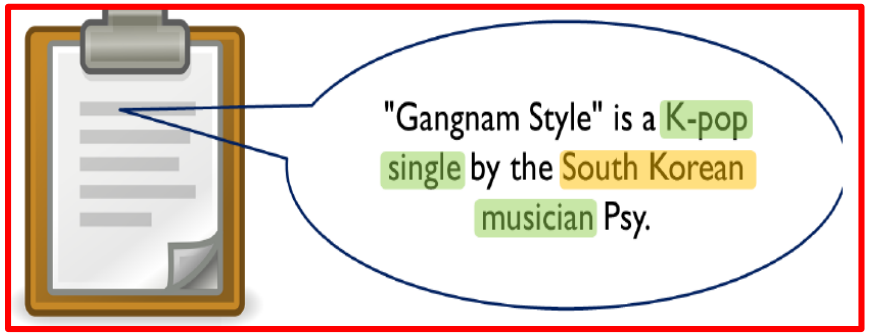
简单的例子



每篇文档包含多个主题，且有一定的比例(文档主题分布)



每个主题包含一些列的单词(主题词分布)



文档中的每个单词都有其来自的主题(topic index)

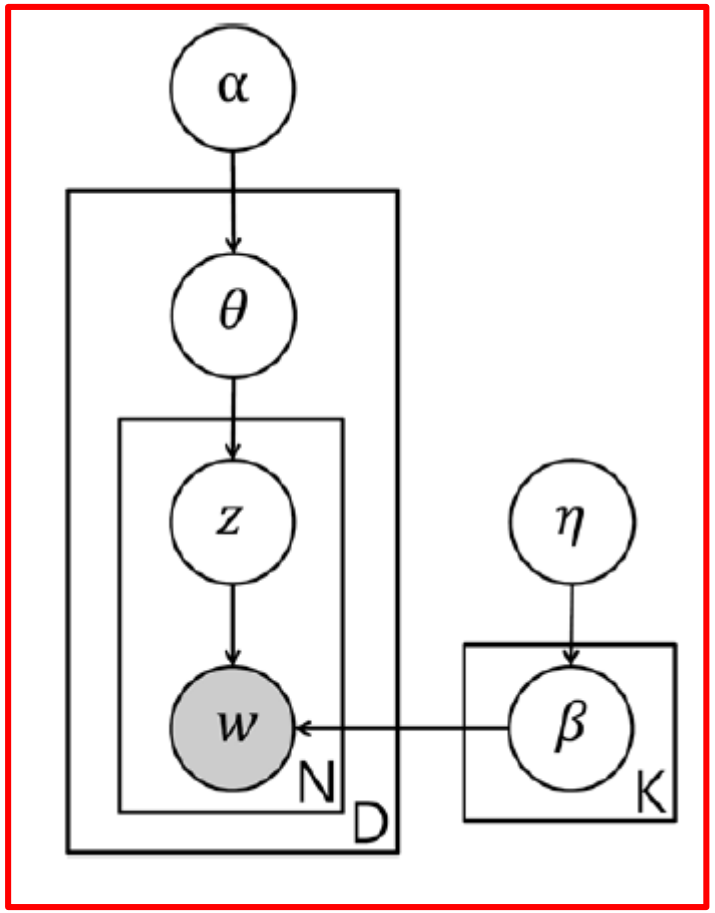
设计思想

- LDA的设计思想
  - 文档是为了描述一定的主题的；每个文档可以描述多个主题。
  - 一个主题是一组固定的词汇表的随机分布。
  - 文档的所有主题服从某种随机分布。
- 我们是如何形成一个文档的？
  - 选择一个或多个主题。
  - 从描述相应的主题词汇表中选择相应的词汇。

注：LDA所面对的是已知的文档集，基于已知的文档集发掘这些文档是怎么生成的！

David M. Blei, Andrew Y. Ng et al. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003)

生成过程

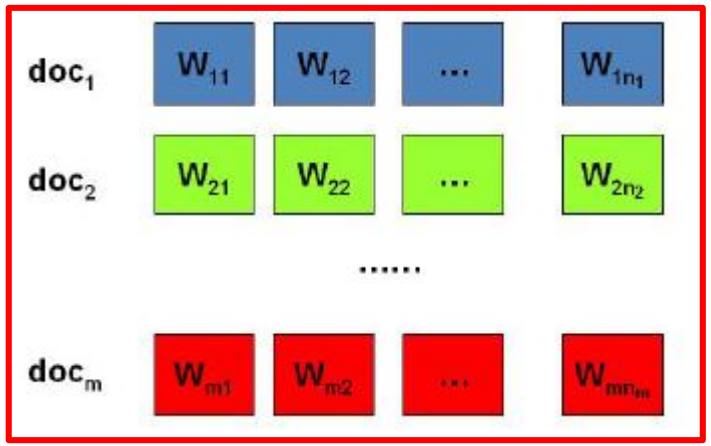


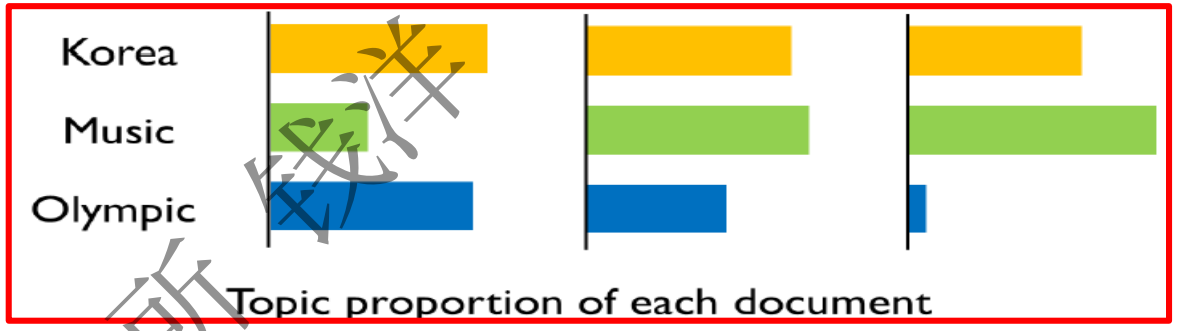
LDA的概率图表示

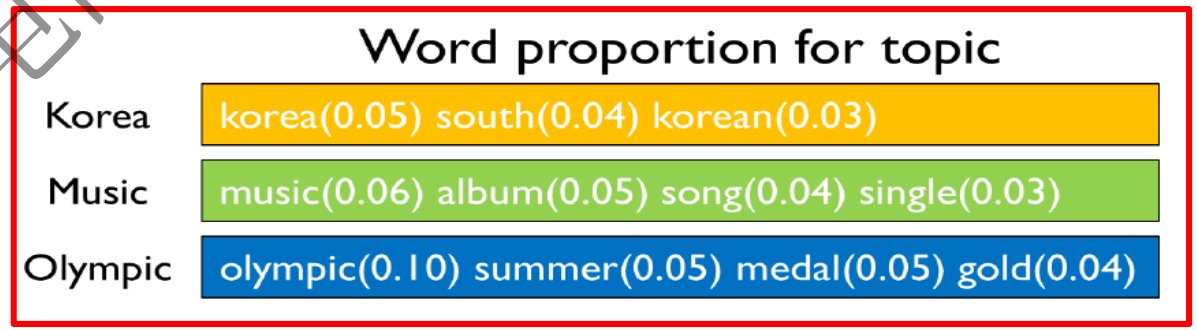
- For each topic $k \in \{1, \ldots, K\}$:
  - Draw word distributions $\beta_k \sim Dir(\eta)$
- For each document $d \in \{1, \ldots, D\}$:
  - Draw topic proportions $\theta_d \sim Dir(\alpha)$
  - For each word in a document $n \in \{1, \ldots, N\}$:
    - ★ Draw a topic index $z_{dn} \sim Mult(\theta)$
    - ★ Generate word from chosen topic $w_{dn} \sim Mult(\beta_{z_{dn}})$

LDA的生成流程

算法输入与输出



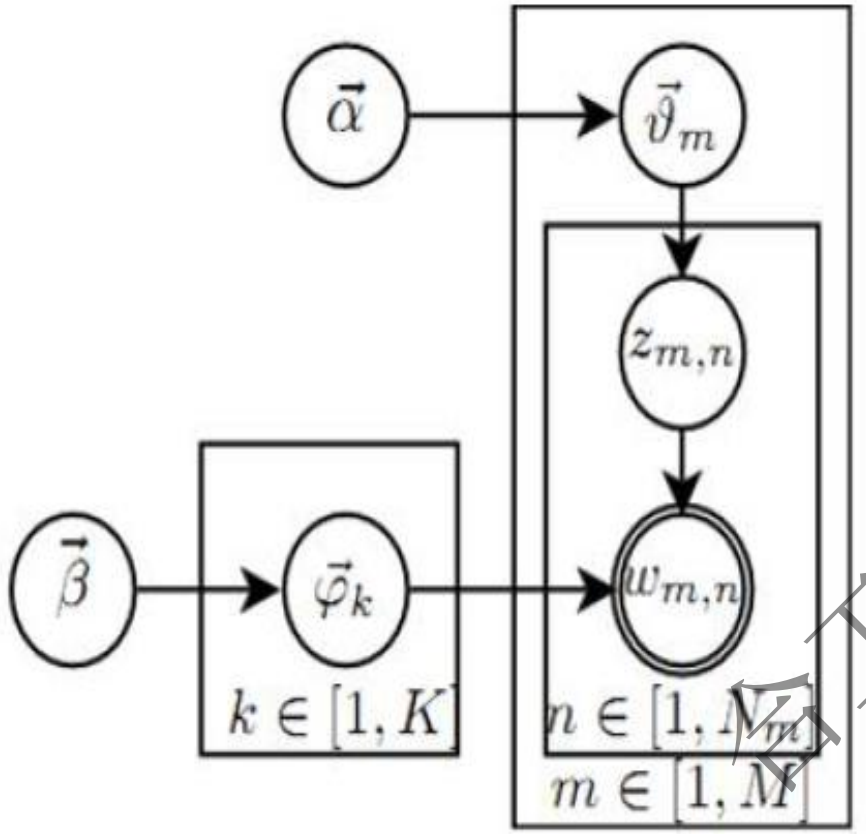




主要内容

- LDA 应用场景
- LDA 涉及知识
- LDA 生成模型
- LDA模型推理及实现
- LDA模型的扩展

LDA模型推理(一)



Step1：估计的目标分布

$$\mathrm{p}(z_i = k|\overrightarrow{z_{-i}}, \vec{w}, \vec{\alpha}, \vec{\beta}) = \frac{p(z_i, \overrightarrow{z_{-i}}, \vec{w}, \vec{\alpha}, \vec{\beta})}{\mathrm{p}(\overrightarrow{z_{-i}}, \vec{w}, \vec{\alpha}, \vec{\beta})}$$

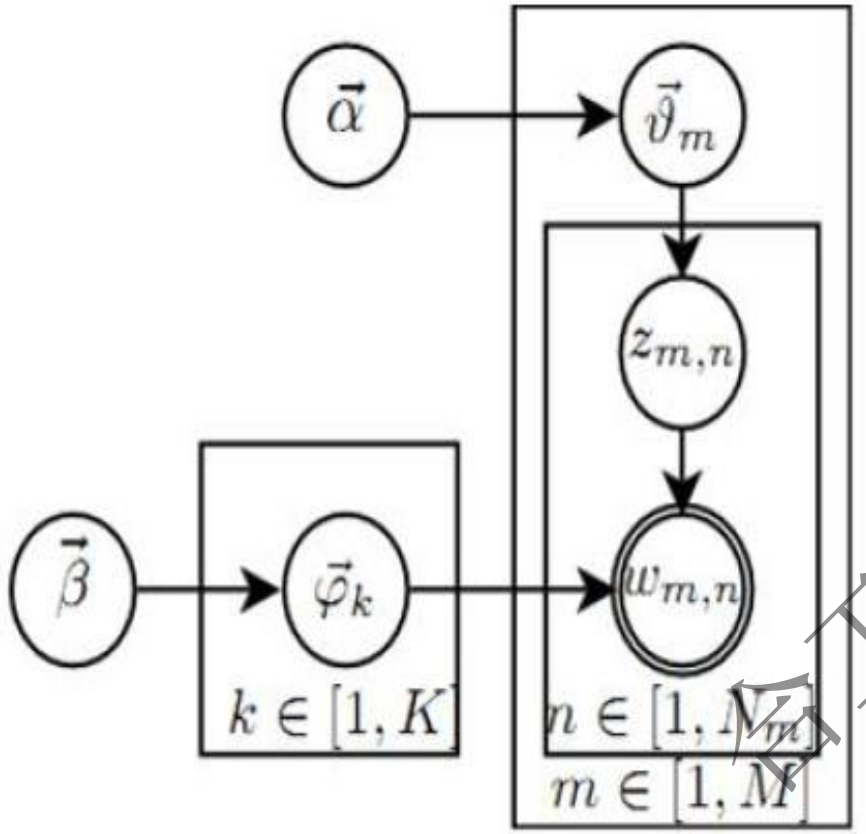
$$= \frac{p(\vec{z}, \vec{w}|\vec{\alpha}, \vec{\beta})}{\mathrm{p}(\overrightarrow{z_{-i}}, \overrightarrow{w_{-i}}, w_i|\vec{\alpha}, \vec{\beta})} = \frac{p(\vec{z}, \vec{w}|\vec{\alpha}, \vec{\beta})}{\mathrm{p}(\overrightarrow{z_{-i}}, \overrightarrow{w_{-i}}|\vec{\alpha}, \vec{\beta})p(w_i)}$$

$$\propto \frac{p(\vec{z}, \vec{w}|\vec{\alpha}, \vec{\beta})}{\mathrm{p}(\overrightarrow{z_{-i}}, \overrightarrow{w_{-i}}|\vec{\alpha}, \vec{\beta})}$$

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

转化成了求联合概率分布

LDA模型推理(一)



Step2: 联合概率分布

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

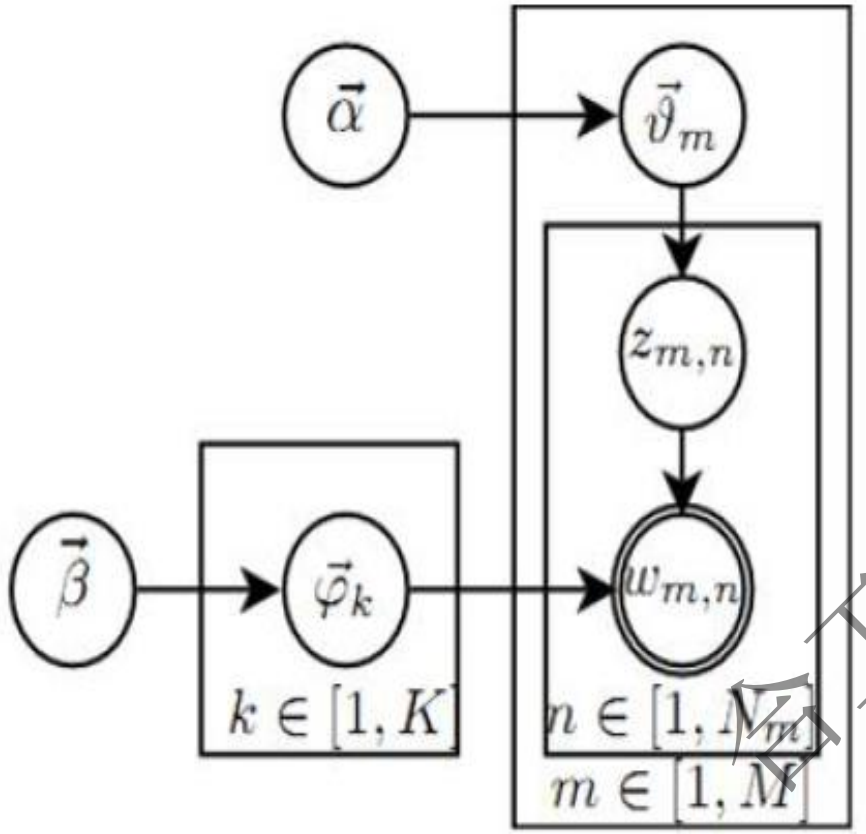
$$= p(\vec{w}|\vec{z},\vec{\beta})p(\vec{z}|\vec{\alpha})$$

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

$$= \int \prod_{k=1}^{K}\prod_{t=1}^{W} \varphi_{k,t}^{n_k^{(t)}} \prod_{k=1}^{K} \frac{\Gamma(\sum_{t=1}^{W}\beta_t)}{\prod_{t=1}^{W}\Gamma(\beta_t)} \prod_{t=1}^{W} \varphi_{k,t}^{\beta_t-1}\, d\,\vec{\varphi}_k$$

$$= \left(\frac{\Gamma(\sum_{t=1}^{W}\beta_t)}{\prod_{t=1}^{W}\Gamma(\beta_t)}\right)^{K} \prod_{k=1}^{K} \int \prod_{t=1}^{W} \varphi_{k,t}^{n_k^{(t)}+\beta_t-1}\, d\,\vec{\varphi}_k$$

LDA模型推理(一)



Step2: 联合概率分布

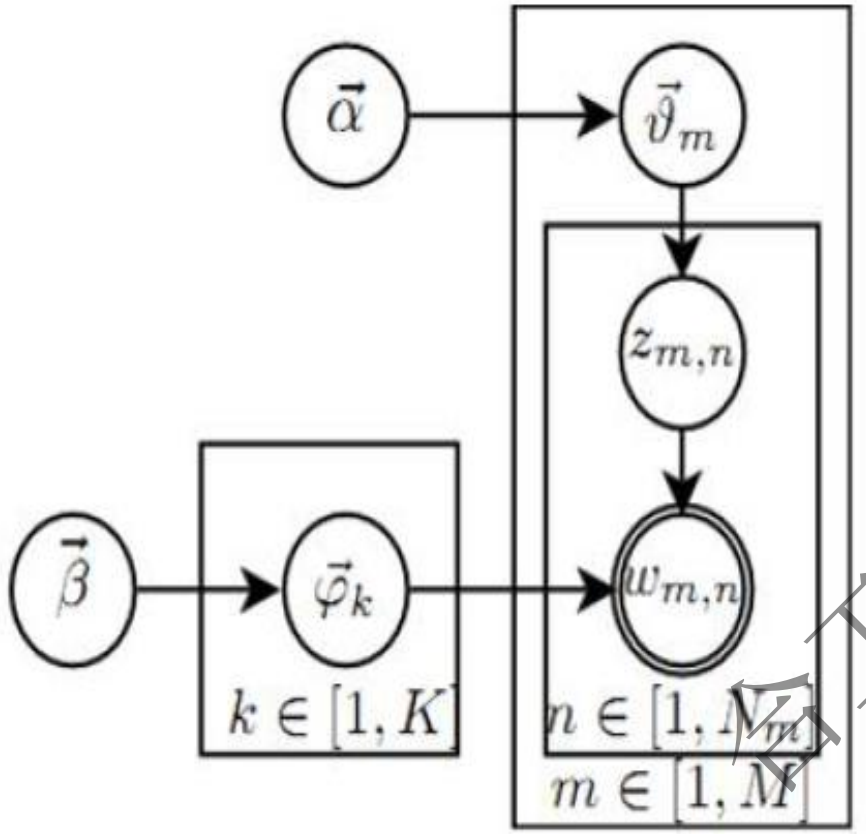
$$\int \prod_{t=1}^{W} \varphi_{k,t}^{n_k^{(t)}+\beta_t-1} \, d\,\vec{\varphi}_k = \frac{\prod_{t=1}^{W} \Gamma\left(n_k^{(t)}+\beta_t\right)}{\Gamma\left(\sum_{t=1}^{W}\left(n_k^{(t)}+\beta_t\right)\right)}$$

欧拉积分

$$p\left(\vec{w}|\vec{z},\vec{\beta}\right) = \left(\frac{\Gamma(\sum_{t=1}^{W}\beta_t)}{\prod_{t=1}^{W}\Gamma(\beta_t)}\right)^K \prod_{k=1}^{K} \frac{\prod_{t=1}^{W} \Gamma\left(n_k^{(t)}+\beta_t\right)}{\Gamma\left(\sum_{t=1}^{W}\left(n_k^{(t)}+\beta_t\right)\right)}$$

$$= \prod_{k=1}^{K} \frac{\nabla\left(\overrightarrow{n_k}+\vec{\beta}\right)}{\nabla\left(\vec{\beta}\right)}$$

LDA模型推理(一)



Step2: 联合概率分布

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

$$= \int \prod_{m=1}^{M}\prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \vartheta_{m,k}^{\alpha_k-1}\,\mathrm{d}\vec{\vartheta}_m$$

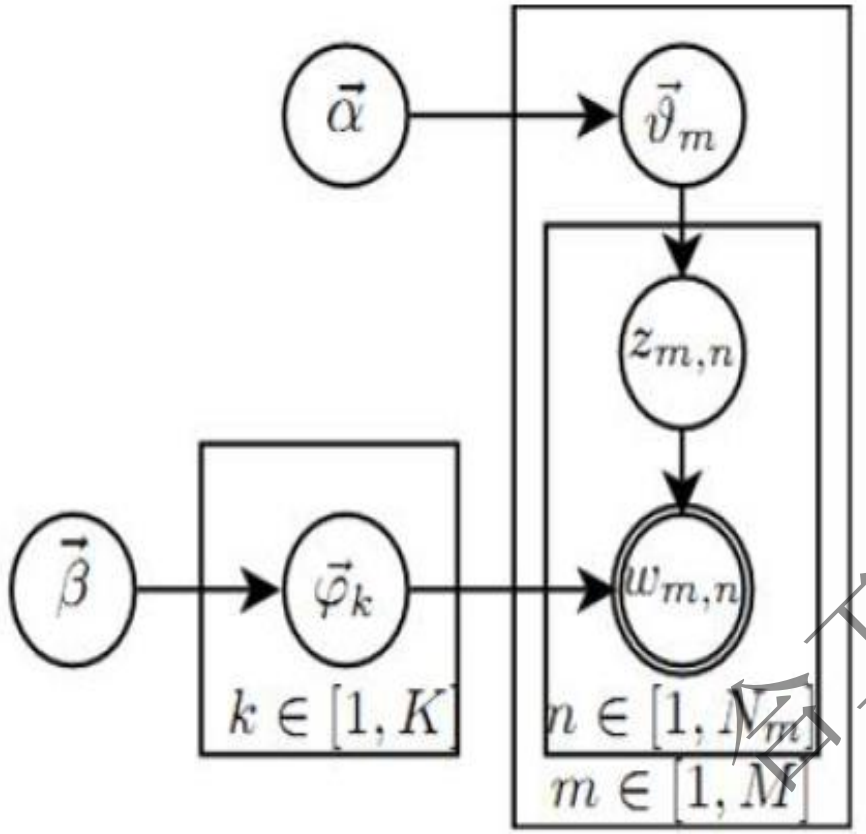
$$= \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \int \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}+\alpha_k-1}\,\mathrm{d}\vec{\vartheta}_m$$

$$= \left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)^{M} \prod_{m=1}^{M} \frac{\prod_{k=1}^{K}\Gamma\left(n_m^{(k)}+\alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_m^{(k)}+\alpha_k\right)\right)}$$

$$= \prod_{m=1}^{M} \frac{\nabla(\overrightarrow{n_m}+\vec{\alpha})}{\nabla(\vec{\alpha})}$$

$$\Delta(\vec{\alpha}) = \frac{\prod_{k=1}^{K}\Gamma(\alpha)}{\Gamma\left(\sum_{k=1}^{K}\alpha\right)}$$

LDA模型推理(一)

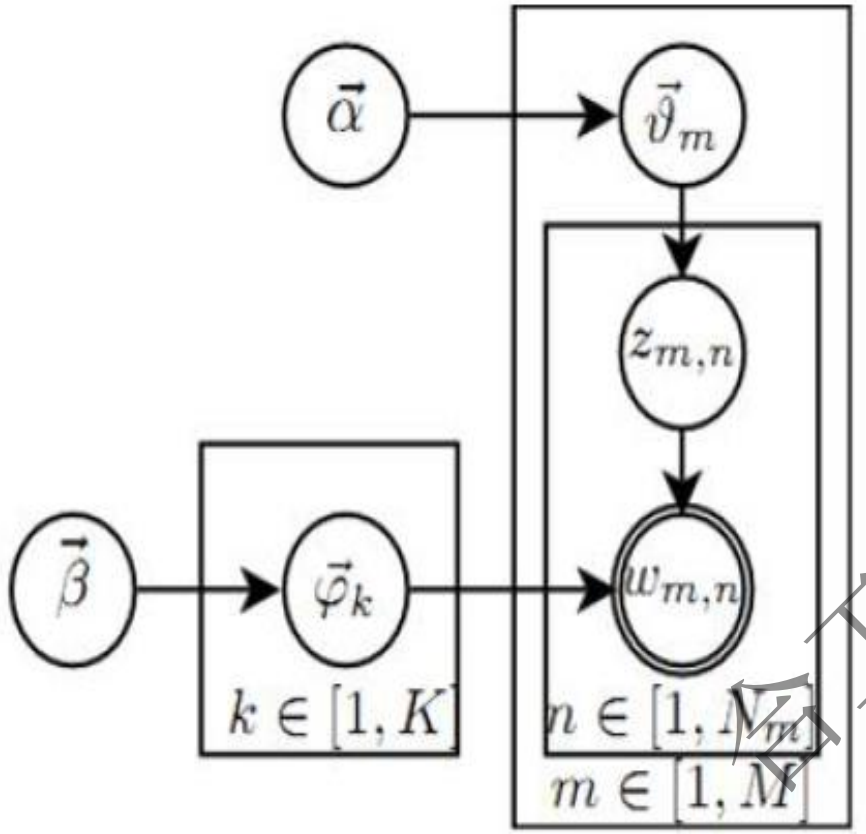


Step3: 约分化简

$$p(z_i = k|\overrightarrow{z_{-i}}, \vec{w}, \vec{\alpha}, \vec{\beta}) = \frac{p(z_i, \overrightarrow{z_{-i}}, \vec{w}, \vec{\alpha}, \vec{\beta})}{p(\overrightarrow{z_{-i}}, \vec{w}, \vec{\alpha}, \vec{\beta})}$$

$$= \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^{W}(n_{k,-i}^{(t)} + \beta_t)} \cdot \frac{n_{m,-i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,-i}^{(k)} + \alpha_k)}$$

LDA模型推理(二)



$$p(z_i = k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}) \propto p(z_i = k, w_i = t|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i})$$

$$= \int p(z_i = k, w_i = t, \vec{\theta}_m, \vec{\varphi}_k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) d\vec{\theta}_m d\vec{\varphi}_k$$

$$= \int p(z_i = k, \vec{\theta}_m|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) \cdot p(w_i = t, \vec{\varphi}_k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) d\vec{\theta}_m d\vec{\varphi}_k$$

$$= \int p(z_i = k|\vec{\theta}_m) p(\vec{\theta}_m|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) \cdot p(w_i = t|\vec{\varphi}_k) p(\vec{\varphi}_k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) d\vec{\theta}_m d\vec{\varphi}_k$$

$$= \int p(z_i = k|\vec{\theta}_m) Dir(\vec{\theta}_m|\vec{n}_{m,\neg i} + \vec{\alpha}) d\vec{\theta}_m$$

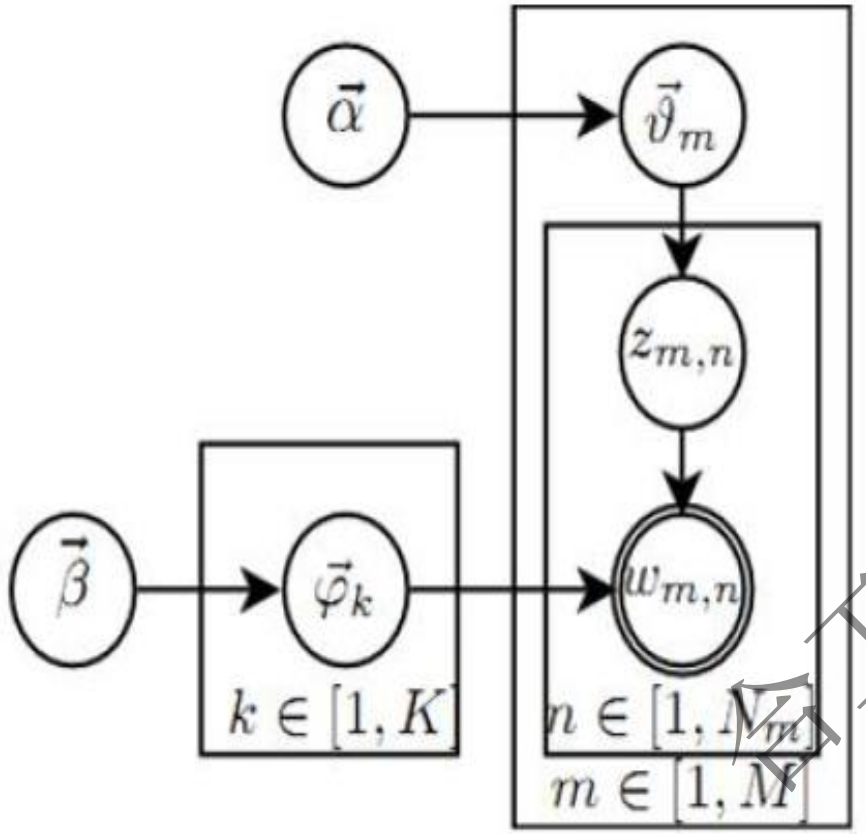
$$\cdot \int p(w_i = t|\vec{\varphi}_k) Dir(\vec{\varphi}_k|\vec{n}_{k,\neg i} + \vec{\beta}) d\vec{\varphi}_k$$

$$= \int \theta_{mk} Dir(\vec{\theta}_m|\vec{n}_{m,\neg i} + \vec{\alpha}) d\vec{\theta}_m \cdot \int \varphi_{kt} Dir(\vec{\varphi}_k|\vec{n}_{k,\neg i} + \vec{\beta}) d\vec{\varphi}_k$$

$$= E(\theta_{mk}) \cdot E(\varphi_{kt})$$

$$= \hat{\theta}_{mk} \cdot \hat{\varphi}_{kt}$$

LDA模型推理(二)



$$\hat{\theta}_{mk} = \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(t)} + \alpha_k)}$$

$$\hat{\varphi}_{kt} = \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)} + \beta_t)}$$

$$p(z_i = k|\vec{\mathrm{z}}_{\neg i}, \vec{\mathrm{w}}) \propto \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(t)} + \alpha_k)} \cdot \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)} + \beta_t)}$$

LDA Gibbs Sampling

Gibbs sampling算法有三个阶段：初始化、burn-in(Gibbs未收敛阶段)和sampling(Gibbs收敛阶段)

- 算法：LdaGibbs({w,α,β,K})
- 输入：单词向量w，超参数α和β，主题数K
- 全局变量：统计量{n(k)m}、{n(t)k}，以及它们的总数{nm}、{nk}，全部条件概率数组p(zi|·)
- 输出：主题向量{z}，多项分布参数Φ和Θ，超参数估计量α和β
- [初始化] 设置全局变量n(k)m、n(t)k、nm、nk为零
- 对所有文档 m∈[1,M]：
  - 对文档 m 中的所有单词 n∈[1,Nm]：
    - 采样每个单词对应的主题zm,n=k~Mult(1/K)
    - 增加“文档-主题”计数：n(k)m+=1
    - 增加“文档-主题”总数：nm+=1
    - 增加“主题-词项”计数：n(t)k+=1
    - 增加“主题-词项”总数：nk+=1

- 迭代burn-in和sampling步骤：
  - [burn-in] 对所有文档 m∈[1,M]：
    - 对文档 m 中的所有单词 n∈[1,Nm]：
      - 减少计数：n(k)m−=1;nm−=1;n(t)k−=1;nk−=1;
      - 根据公式p(zi=k|z¬i,w) = ...{公式（80）}采样主题：k~p(zi|z¬i,w)
      - 增加计数：n(k~)m+=1;nm+=1;n(t)k~+=1;nk~+=1;
  - [sampling] 如果Markov链收敛：
    - 根据公式φk,t生成参数 Φ
    - 根据公式ϑm,k生成参数 Θ

参考博客：
http://blog.csdn.net/pipisorry/article/details/42649657

LDA源码分析

```
for(int i = 0; i < iterations; i++){
    //输出这是第几次迭代
    System.out.println("Iteration " + i);
    //当迭代次数大于等于开始保存的代数且((i - beginSaveIters) % saveStep) == 0，则
    if((i >= beginSaveIters) && (((i - beginSaveIters) % saveStep) == 0)){
        //Saving the model，当符号上述条件后就开始保存计算的参数了
        System.out.println("Saving model at iteration " + i +" ... ");
        //Firstly update parameters，第一步更新模型参数
        updateEstimatedParameters();
        //Secondly print model variables，第二步打印模型的参数
        saveIteratedModel(i, docSet);
    }

    //Use Gibbs Sampling to update z[][]，使用吉布斯抽样去更新多维矩阵z[][]
    for(int m = 0; m < M; m++){
        //获取第m篇文档的单词数目
        int N = docSet.docs.get(m).docWords.length;
        for(int n = 0; n < N; n++){
            // Sample from p(z_i|z_-i, w)
            int newTopic = sampleTopicZ(m, n);
            z[m][n] = newTopic;
        }
    }
}
```

循环每个文档的每个单词，重新为期分配主题

LDA源码分析

```
private int sampleTopicZ(int m, int n) {
    // Sample from p(z_i|z_-i, w) using Gibbs upde rule根据吉布斯抽样规则进行抽样
    //Remove topic label for w_{m,n}，将第m篇文章的第n个词的topic label删除
    int oldTopic = z[m][n];//获取第m篇文档第n个单词的旧的主题，z[][]中存储的是每个单词对应的topic label
    nmk[m][oldTopic]--;//将m篇文档中，属于主题oldTopic的数量减1.
    nkt[oldTopic][doc[m][n]]--;//减1，doc[][]中存储的是每个单词的index,
    nmkSum[m]--;//将第m篇文章总的单词数减1
    nktSum[oldTopic]--;//

    //上述5行代码特别重要，因为当前的单词w原来所属的topic label必须移除，因为后面的要给w赋新的topic label。

    //Compute p(z_i = k|z_-i, w)
    double [] p = new double[K];//定义double类型的数组
    //下面公式
    for(int k = 0; k < K; k++){
        p[k] = (nkt[k][doc[m][n]] + beta) / (nktSum[k] + V * beta) * (nmk[m][k] + alpha) /
                (nmkSum[m] + K * alpha);
    }
    //Sample a new topic label for w_{m, n} like roulette像轮盘赌一样选择一个标签
    //Compute cumulated probability for p
    for(int k = 1; k < K; k++){
        p[k] += p[k - 1];
    }
    double u = Math.random() * p[K - 1]; //p[] is unnormalised没有规范化
    int newTopic;
    for(newTopic = 0; newTopic < K; newTopic++){
        if(u < p[newTopic]){
            break;
        }
    }

    //Add new topic label for w_{m, n} 给第m篇文档的第n个词添加新的标签
    nmk[m][newTopic]++;//newTopic是一个索引
    nkt[newTopic][doc[m][n]]++;
    nmkSum[m]++;
    nktSum[newTopic]++;
    return newTopic;
}
```

去除该单词，并做相关统计

依据后验概率计算该单词分配到每个主题的概率

轮盘赌分配主题编号

重新做相关统计

轮盘赌算法

轮盘赌算法是一种常用的随机选择算法，在计算机模拟随机过程中广泛应用。

$$P(s_1)=\frac{f(s_1)}{\sum_{j=1}^{N} f(s_j)}=\frac{169}{169+576+64+361}=0.14$$

$$P(s_2)=\frac{f(s_2)}{\sum_{j=1}^{N} f(s_j)}=\frac{576}{169+576+64+361}=0.49$$

$$P(s_3)=\frac{f(s_3)}{\sum_{j=1}^{N} f(s_j)}=\frac{64}{169+576+64+361}=0.06$$

$$P(s_4)=\frac{f(s_4)}{\sum_{j=1}^{N} f(s_j)}=\frac{361}{169+576+64+361}=0.31$$

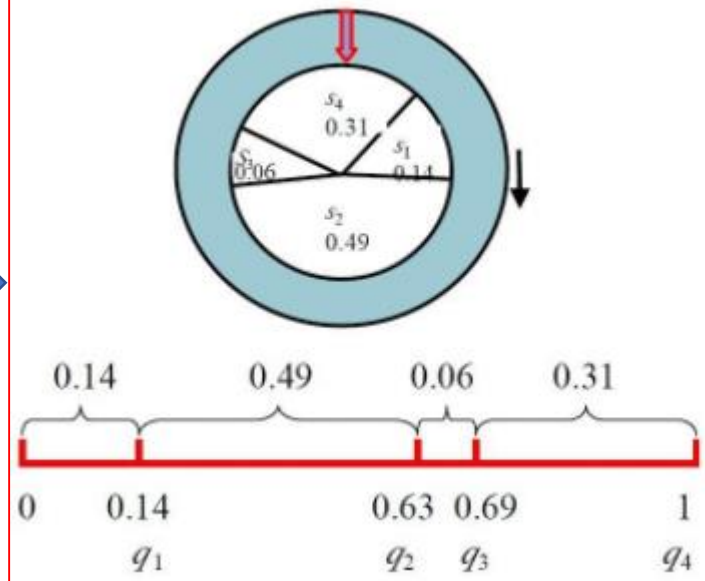
$$q(s_1)=\sum_{j=1}^{N} p(s_j)=0.14$$

$$q(s_2)=\sum_{j=1}^{N} p(s_j)=0.14+0.49=0.63$$

$$q(s_3)=\sum_{j=1}^{N} p(s_j)=0.14+0.49+0.06=0.69$$

$$q(s_4)=\sum_{j=1}^{N} p(s_j)=0.14+0.49+0.06+0.31=1$$

染色体	适应度	选择概率	积累概率	选中次数
s_1=01101	169	0.14	0.14	1
s_2=11000	576	0.49	0.63	2
s_3=01000	64	0.06	0.69	0
s_4=10011	361	0.31	1.00	1



例如设从区间 [0, 1] 中产生4个随机数：

r1 = 0.450126, r2 = 0.110347

r3 = 0.572496, r4 = 0.98503

主题模型评估方法 - Perplexity

$$\mathrm{P}(\tilde{\boldsymbol{W}}|\mathcal{M}) = \prod_{m=1}^{M} p(\vec{\tilde{w}}_{\tilde{m}}|\mathcal{M})^{-\frac{1}{N}} = \exp - \frac{\sum_{m=1}^{M} \log p(\vec{\tilde{w}}_{\tilde{m}}|\mathcal{M})}{\sum_{m=1}^{M} N_m}$$

$$p(\vec{\tilde{w}}_{\tilde{m}}|\mathcal{M}) = \prod_{n=1}^{N_{\tilde{m}}} \sum_{k=1}^{K} p(w_n{=}t|z_n{=}k) \cdot p(z_n{=}k|d{=}\tilde{m})$$

$$= \prod_{t=1}^{V} \left(\sum_{k=1}^{K} \varphi_{k,t} \cdot \vartheta_{\tilde{m},k} \right)^{n_{\tilde{m}}^{(t)}}$$

$$\log p(\vec{\tilde{w}}_{\tilde{m}}|\mathcal{M}) = \sum_{t=1}^{V} n_{\tilde{m}}^{(t)} \log \left(\sum_{k=1}^{K} \varphi_{k,t} \cdot \vartheta_{\tilde{m},k} \right)$$

$$\vartheta_{\tilde{m},k} = \frac{n_{\tilde{m}}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{\tilde{m}}^{(k)} + \alpha_k}.$$

测试语料似然的均值

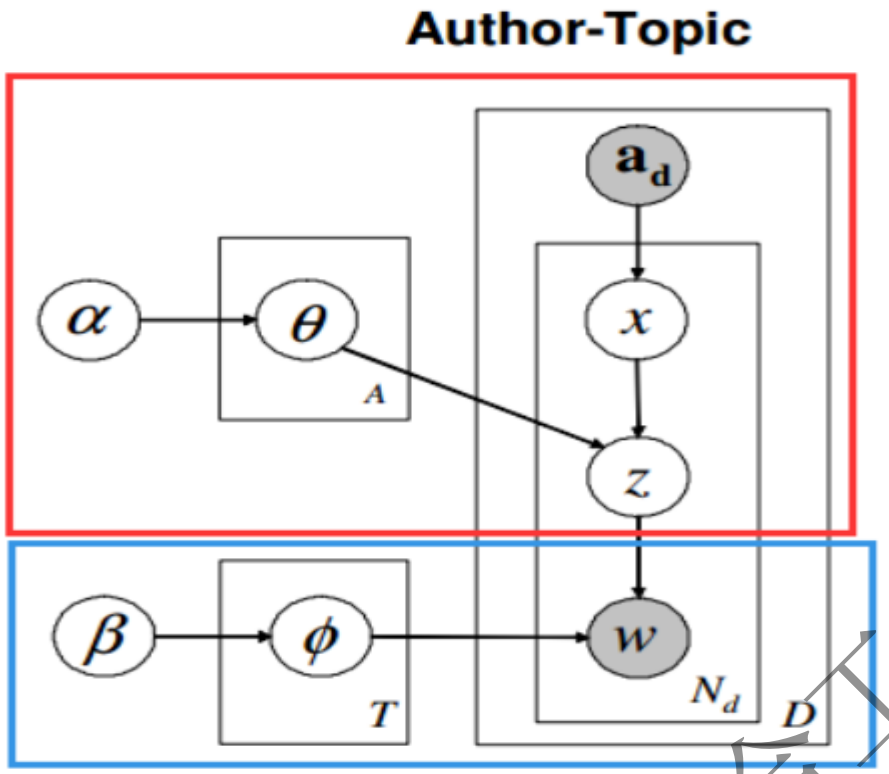
主要参考内容

- Gregor Heinrich, Parameter estimation for text analysis. 2008
- David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet Allocation, 2003
- Steyvers M, Griffiths T. Probabilistic topic models[J]. Handbook of latent semantic analysis, 2007, 427(7): 424-440
- 靳志辉，LDA数学八卦，2013
- 博客：http://blog.csdn.net/qy20115549/article/
- 博客：http://www.datalearner.com/blog_list

主要内容

- LDA 应用场景
- LDA 涉及知识
- LDA 生成模型
- LDA模型推理及实现
- LDA模型的扩展

Author-Topic Model



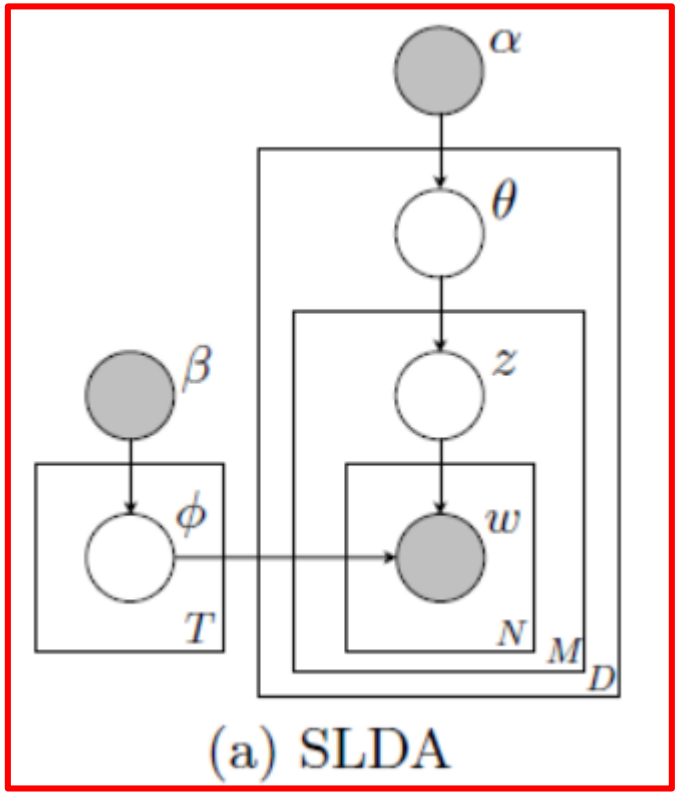
模型将作者的兴趣偏好及文档的内容信息融合在一起。

- a_d向量表示决定写文档d的一群作者。
- 一个作者对应多个主题。
- 文档中的一个单词对应的作者来自于均匀分布。
- 主题的抽取来自于作者对应的主题分布，而词的抽取来自于该主题对应的词分布。

http://blog.csdn.net/qy20115549/article/details/54407099

Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004: 487-494.(UAI A+)

Sentence-LDA



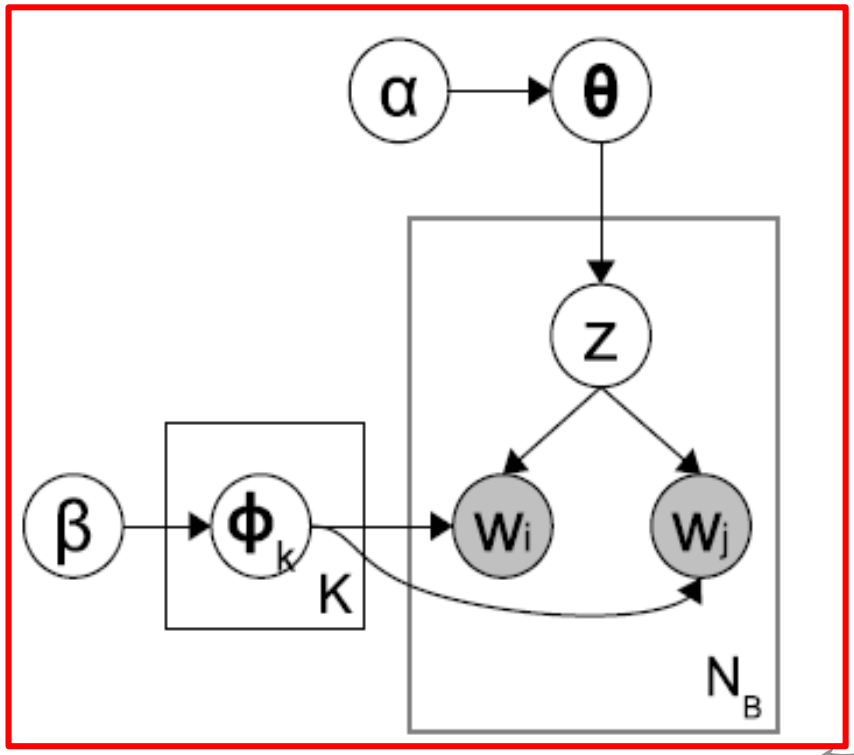
(a) SLDA

Jo Y, Oh A H. Aspect and sentiment unification model for online review analysis[C]//Proceedings of the fourth ACM international conference on Web search and data mining. **ACM**, 2011: 815-824.

Balikas G, Amini M R, Clausel M. On a Topic Model for Sentences[C]//Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. **ACM**, 2016: 921-924. (A+ SIGIR)

Büschken J, Allenby G M. Sentence-based text analysis for customer reviews[J]. **Marketing Science**, 2016, 35(6): 953-975.

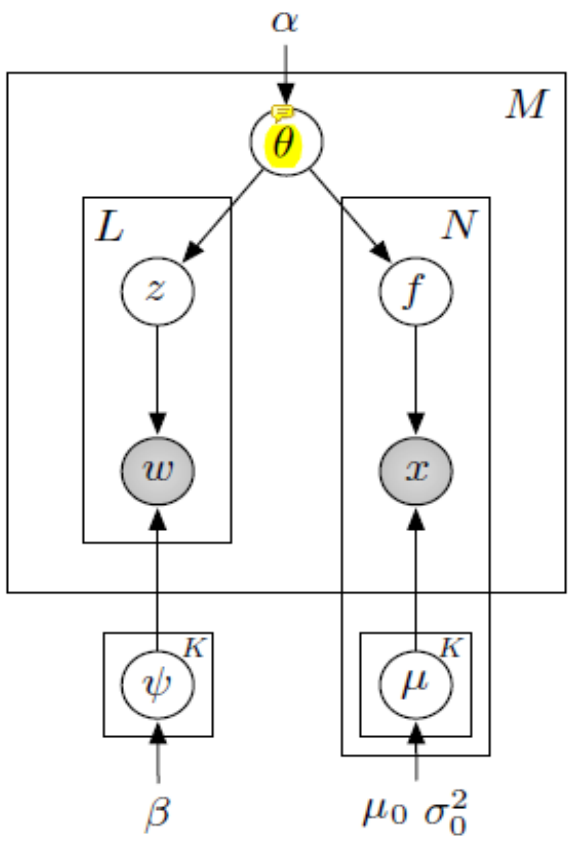
BTM



Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C]//Proceedings of the 22nd international conference on World Wide Web. ACM, 2013: 1445-1456.（A+ WWW）

Cheng X, Yan X, Lan Y, et al. Btm: Topic modeling over short texts[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 2928-2941.(A TKDE)

RMR

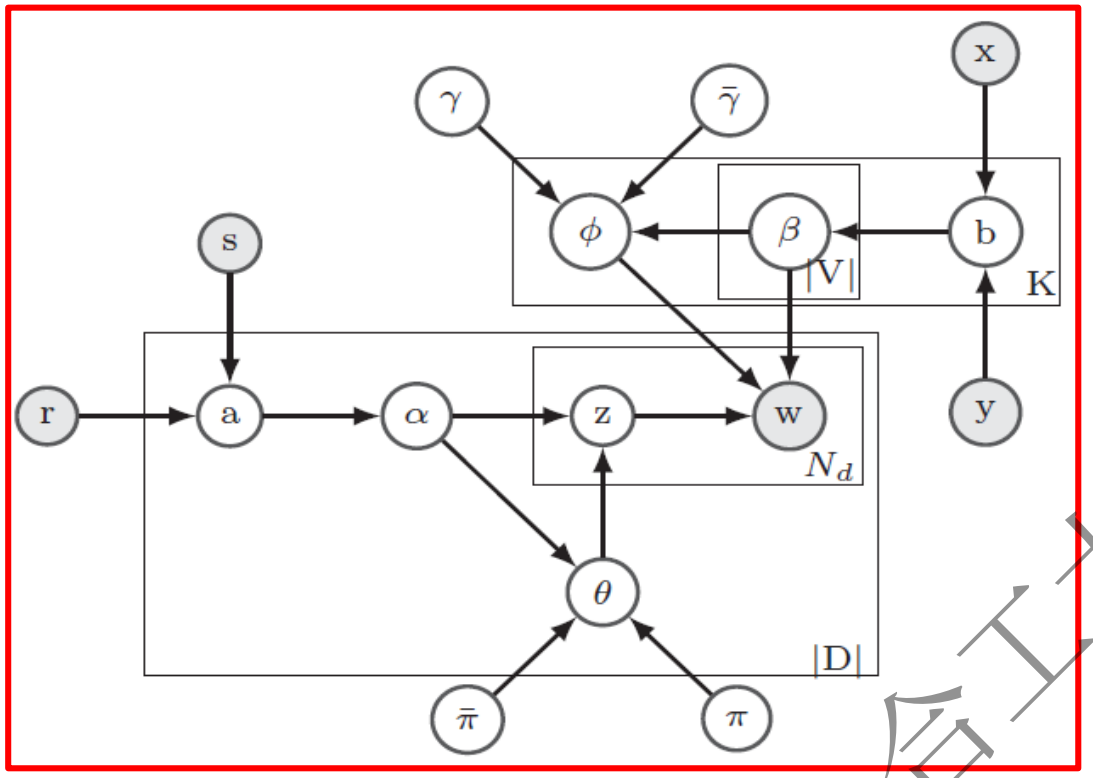


Ratings Meet Reviews

1. For each user $u \in \mathcal{U}$:
   (a) For each latent topic dimension $k \in [1, K]$:
      i. Draw $\mu_{u,k} \sim \text{Gaussian}(\mu_0, \sigma_0^2)$
2. For each latent topic dimension $k \in [1, K]$:
   (a) Draw $\psi_k \sim \text{Dirichlet}(\beta)$
3. For each item $v \in \mathcal{V}$:
   (a) Draw topic mixture proportion $\theta_v \sim \text{Dirichlet}(\alpha)$
   (b) For each description word $w_{v,n}$:
      i. Draw topic assignment $z_{v,n} \sim \text{Multinomial}(\theta_v)$
      ii. Draw word $w_{v,n} \sim \text{Multinomial}(\psi_{z_{v,n}})$
   (c) For each observed rating assigned by u to v:
      i. Draw topic assignment $f_{v,u} \sim \text{Multinomial}(\theta_v)$
      ii. Draw the rating $x_{v,u} \sim \text{Gaussian}(\mu_{u,f_{v,u}}, \sigma^2)$.

Ling G, Lyu M R, King I. Ratings meet reviews, a combined approach to recommend[C]//Proceedings of the 8th ACM Conference on Recommender systems. ACM, 2014: 105-112.

DsparseTM



2. For each topic $k \in \{1, 2, \ldots, K\}$:
   (a) the topic selector $\alpha_{dk} \sim \text{Bernoulli}(a_d)$, $\vec{\alpha}_d = \{\alpha_{dk}\}_{k=1}^{K}$;
   (b) the set of focused topics: $A_d = \{k : \alpha_{dk} = 1\}$;
3. the topic proportion $\vec{\theta}_d \sim \text{Dirichlet}(\pi\vec{\alpha}_d + \bar{\pi}\vec{1})$;

Lin T, Tian W, Mei Q, et al. The dual-sparse topic model: mining focused topics and focused terms in short text[C]//Proceedings of the 23rd international conference on World wide web. ACM, 2014: 539-550.

TTM

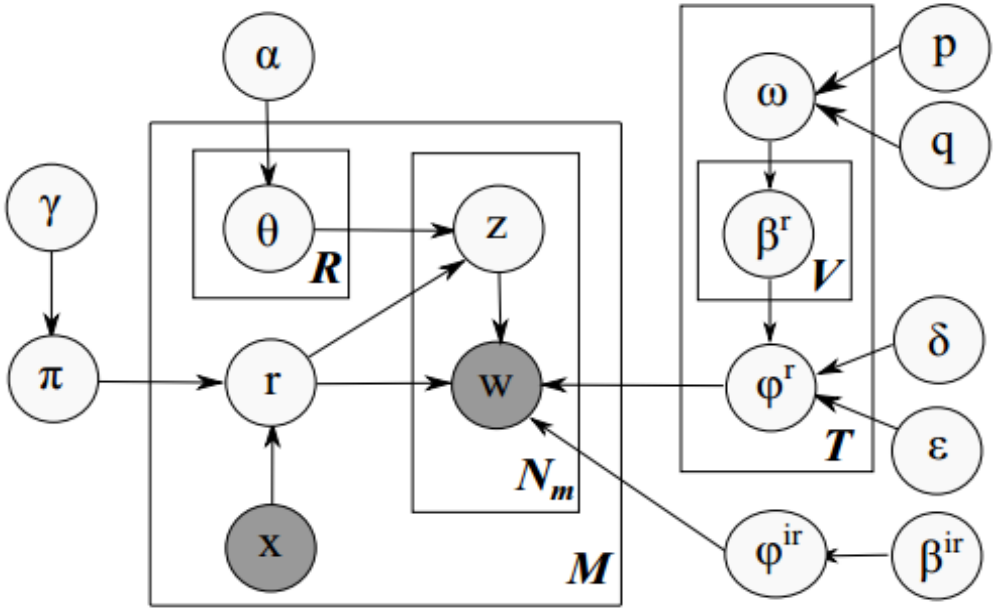
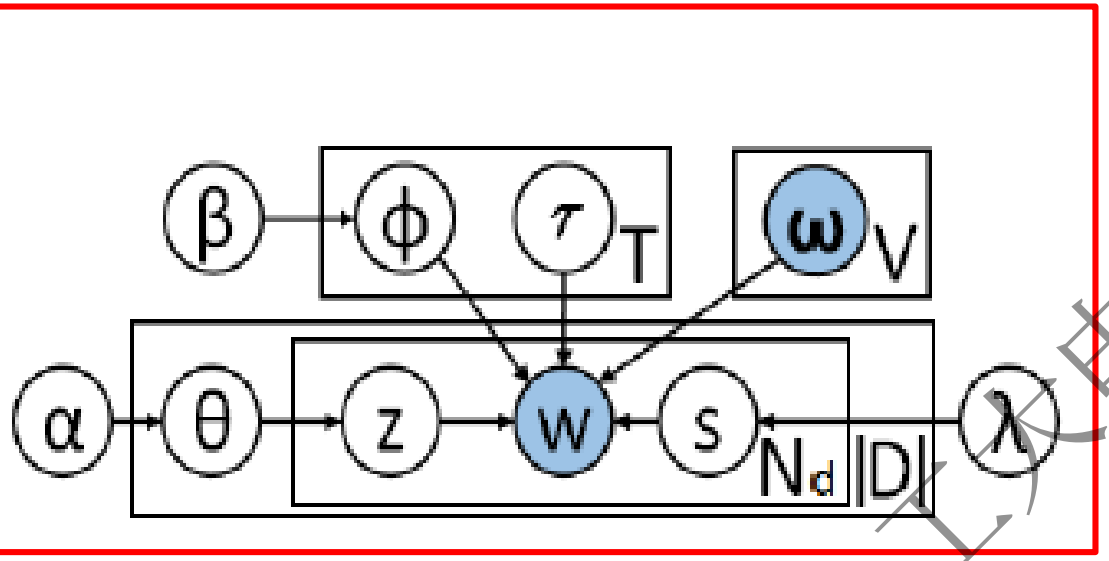


Figure 1: The graphical model of TTM

1. Draw $\varphi^{ir} \sim Dirichlet(\boldsymbol{\beta^{ir}})$ as a word distribution of a irrelevant topic to the targeted aspect;
2. For each target-relevant topic $t \in \{1, 2, ..., T\}$:
   (a) Draw a prior distribution $\omega_t \sim Beta(p, q)$;
   (b) For each term $v \in \{1, 2, ..., V\}$:
      i. Draw a term selector $\beta^r_{t,\,v} \sim Bernoulli(\omega_t)$;
   (c) Draw a word distribution $\varphi^r_t \sim Dirichlet(\boldsymbol{\beta^r_t}\,\boldsymbol{\delta} + \boldsymbol{\epsilon})$;
3. For each document $m \in \{1, 2, ..., M\}$:
   (a) Draw a prior distribution $\pi_m \sim Beta(\boldsymbol{\gamma})$;
   (b) Draw relevance status r based on keyword indicator x and $Bernoulli(\pi_m)$;
   (c) If the document is relevant to the targeted aspect, i.e., $r = 1$:
      i. Draw a topic $z \sim Multinomial(\theta^r)$;
      ii. Emit a word $w_i \sim Multinomial(\varphi^r_z)$.
   (d) If the document is irrelevant to the targeted aspect, i.e., $r = 0$:
      i. Emit a word $w_i \sim Multinomial(\varphi^{ir})$

Wang S, Chen Z, Fei G, et al. Targeted topic modeling for focused analysis[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 1235-1244.

LF-LDA



$$\theta_d \sim \mathrm{Dir}(\alpha) \qquad z_{d_i} \sim \mathrm{Cat}(\theta_d)$$
$$\phi_z \sim \mathrm{Dir}(\beta) \qquad s_{d_i} \sim \mathrm{Ber}(\lambda)$$
$$w_{d_i} \sim (1 - s_{d_i})\mathrm{Cat}(\phi_{z_{d_i}}) + s_{d_i}\mathrm{CatE}(\tau_{z_{d_i}}\, \omega^\top)$$

Nguyen D Q, Billingsley R, Du L, et al. Improving topic models with latent feature word representations[J]. Transactions of the Association for Computational Linguistics, 2015, 3: 299-313.

总结

(1)概率图模型加圈
(2)模型转变为非参
(2)找到对应的管理学问题加以应用(注重解释性)

(1)Puranam D, Narayan V, Kadiyali V. The Effect of Calorie Posting Regulation on Consumer Opinion: A Flexible Latent Dirichlet Allocation Model with Informative Priors[J]. Marketing Science, 2017, 36(5): 726-746.

(2)Nam H, Joshi Y V, Kannan P K. Harvesting brand information from social tags[J]. Journal of Marketing, 2017, 81(4): 88-108.

(3)Büschken J, Allenby G M. Sentence-based text analysis for customer reviews[J]. Marketing Science, 2016, 35(6): 953-975.

(4)Jacobs B J D, Donkers B, Fok D. Model-based purchase predictions for large assortments[J]. Marketing Science, 2016, 35(3): 389-404.

END!