

# Latent Dirichlet Allocation

介紹者 :Joyce Jhang

## 原作 & 出處

- David M. Blei, Andrew Y. Ng and Michael I. Jordan
- Journal of Machine Learning Research 3 (2003) 993-1022

大綱

Introduction

# Introduction

The goal

To find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.

# LATENT DIRICHLET ALLOCATION

- 隱含狄利克雷分布（英語：Latent Dirichlet allocation，簡稱 LDA）
- 主題模型，它可以將文檔集中每篇文檔的主題按照概率分布的形式給出。同時它是一種無監督學習算法
- 正如Beta分布是二項式分布的共軛先驗概率分布，狄利克雷分布作為多項式分布的共軛先驗概率分布。

## “Arts”

## “Budgets”

## “Children”

## “Education”

NEW  
FILM  
SHOW  
MUSIC  
MOVIE  
PLAY  
MUSICAL  
BEST  
ACTOR  
FIRST  
YORK  
OPERA  
THEATER  
ACTRESS  
LOVE

MILLION  
TAX  
PROGRAM  
BUDGET  
BILLION  
FEDERAL  
YEAR  
SPENDING  
NEW  
STATE  
PLAN  
MONEY  
PROGRAMS  
GOVERNMENT  
CONGRESS

CHILDREN  
WOMEN  
PEOPLE  
CHILD  
YEARS  
FAMILIES  
WORK  
PARENTS  
SAYS  
FAMILY  
WELFARE  
MEN  
PERCENT  
CARE  
LIFE

SCHOOL  
STUDENTS  
SCHOOLS  
EDUCATION  
TEACHERS  
HIGH  
PUBLIC  
TEACHER  
BENNETT  
MANIGAT  
NAMPHY  
STATE  
PRESIDENT  
ELEMENTARY  
HAITI

## pLSI model

- Hofmann (1999)
- The probabilistic LSI (pLSI) model
- Multinomial( $\theta$ )
- 詞袋方法，無關詞與詞之先後順序



主题数  $K=3$



教育

经济

交通

单词数  $V=3$



# pLSI model

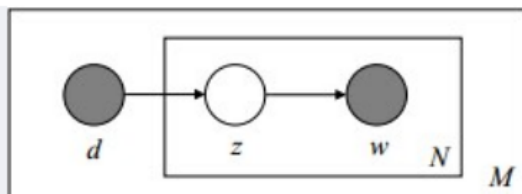
利用上述的第1、3、4个概率，我们便可以按照如下的步骤得到“文档-词项”的生成模型：

1. 按照概率 $P(d_i)$ 选择一篇文档 $d_i$
2. 选定文档 $d_i$ 后，从主题分布中按照概率 $P(z_k|d_i)$ 选择一个隐含的主题类别 $z_k$
3. 选定 $z_k$ 后，从词分布中按照概率 $P(w_j|z_k)$ 选择一个词 $w_j$

所以pLSA中生成文档的整个过程便是选定文档生成主题，确定主题生成词。

反过来，既然文档已经产生，那么如何 **根据已经产生好的文档反推其主题**呢？这个利用看到的文档推断其隐藏的主题（分布）的过程（其实也就是产生文档的逆过程），便是主题建模的目的：自动地发现文档集中的主题（分布）。

文档 $d$ 和单词 $w$ 自然是可被观察到的，但主题 $z$ 却是隐藏的。如下图所示（图中被涂色的 $d$ 、 $w$ 表示可观测变量，未被涂色的 $z$ 表示未知的隐变量， $N$ 表示一篇文档中总共 $N$ 个单词， $M$ 表示 $M$ 篇文档）：



# pLSI model

由于 $P(d_i)$ 可事先计算求出，而 $P(w_j|z_k)$ 和 $P(z_k|d_i)$ 未知，所以 $\theta = (P(w_j|z_k), P(z_k|d_i))$ 就是我们要估计的参数（值），通俗点说，就是要最大化这个 $\theta$ 。

用什么方法进行估计呢，常用的参数估计方法有极大似然估计MLE、最大后验估计MAP、贝叶斯估计等等。因为该待估计的参数中含有隐变量 $z$ ，所以我们可以考虑EM算法。

## 4.2.2 EM算法的简单介绍

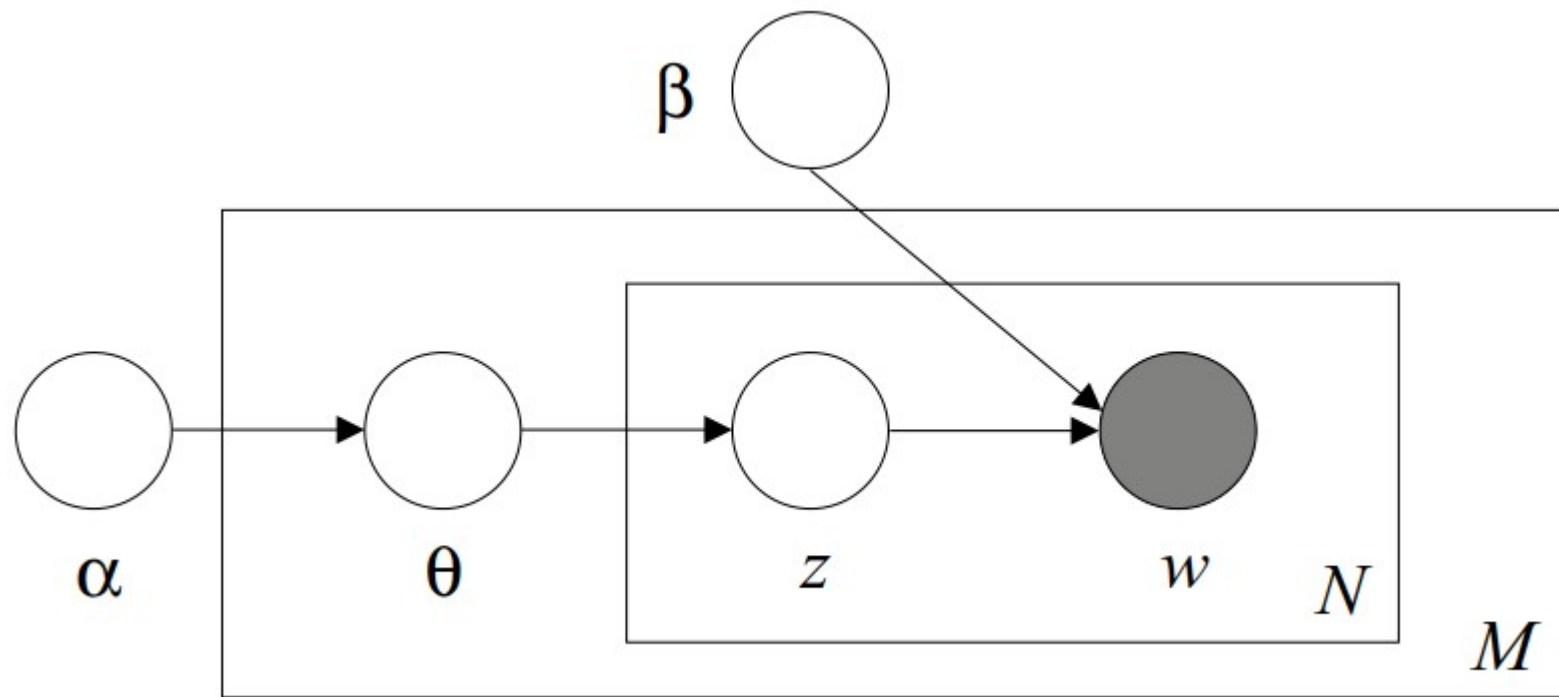
EM算法，全称为Expectation-maximization algorithm，为期望最大算法，其基本思想是：首先随机选取一个值去初始化待估计的值 $\theta^{(0)}$ ，然后不断迭代寻找更优的 $\theta^{(n+1)}$ 使得其似然函数likelihood  $L(\theta^{(n+1)})$ 比原来的 $L(\theta^{(n)})$ 要大。换言之，假定现在得到了 $\theta^{(n)}$ ，想求 $\theta^{(n+1)}$ ，使得

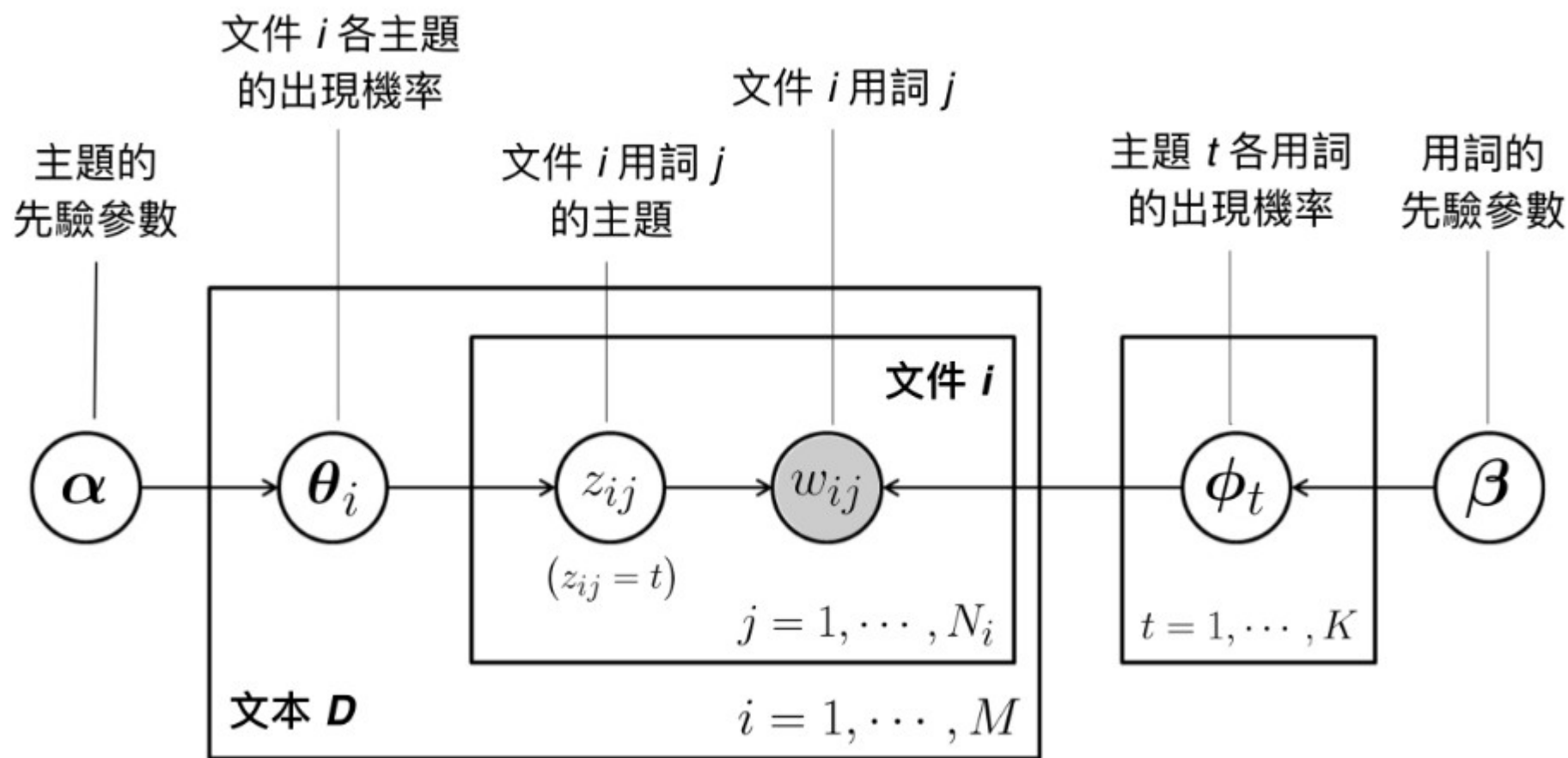
# pLSI model

综上，在pLSA中：

1. 由于 $P(w_j|z_k)$ 和 $P(z_k|d_i)$ 未知，所以我们用EM算法去估计 $\theta = (P(w_j|z_k), P(z_k|d_i))$ 这个参数的值。
2. 而后，用 $\phi_{k,j}$ 表示词项 $w_j$ 出现在主题 $z_k$ 中的概率，即 $P(w_j|z_k) = \phi_{k,j}$ ，用 $\theta_{i,k}$ 表示主题 $z_k$ 出现在文档 $d_i$ 中的概率，即 $P(z_k|d_i) = \theta_{i,k}$ ，从而把 $P(w_j|z_k)$ 转换成了“主题-词项”矩阵 $\Phi$ （主题生成词），把 $P(z_k|d_i)$ 转换成了“文档-主题”矩阵 $\Theta$ （文档生成主题）。
3. 最终求解出 $\phi_{k,j}$ 、 $\theta_{i,k}$ 。

# LDA Model






# LDA Model

- A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by  $\{1, \dots, V\}$ . We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the  $v$ th word in the vocabulary is represented by a  $V$ -vector  $w$  such that  $w^v = 1$  and  $w^u = 0$  for  $u \neq v$ .
- A document is a sequence of  $N$  words denoted by  $\mathbf{w} = (w_1, w_2, \dots, w_N)$ , where  $w_n$  is the  $n$ th word in the sequence.
- A corpus is a collection of  $M$  documents denoted by  $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ .

# LDA Model

LDA assumes the following generative process for each document  $w$  in a corpus  $D$ :

1. Choose  $N$   $\sim \text{Poisson}(\xi)$ . 
2. Choose  $\theta$   $\sim \text{Dir}(\alpha)$ .
3. For each of the  $N$  words  $w_n$ :
  - (a) Choose a topic  $z_n$   $\sim \text{Multinomial}(\theta)$ .
  - (b) Choose a word  $w_n$  from  $p(w_n | z_n, \beta)$ , a multinomial probability conditioned on the topic  $z_n$ .



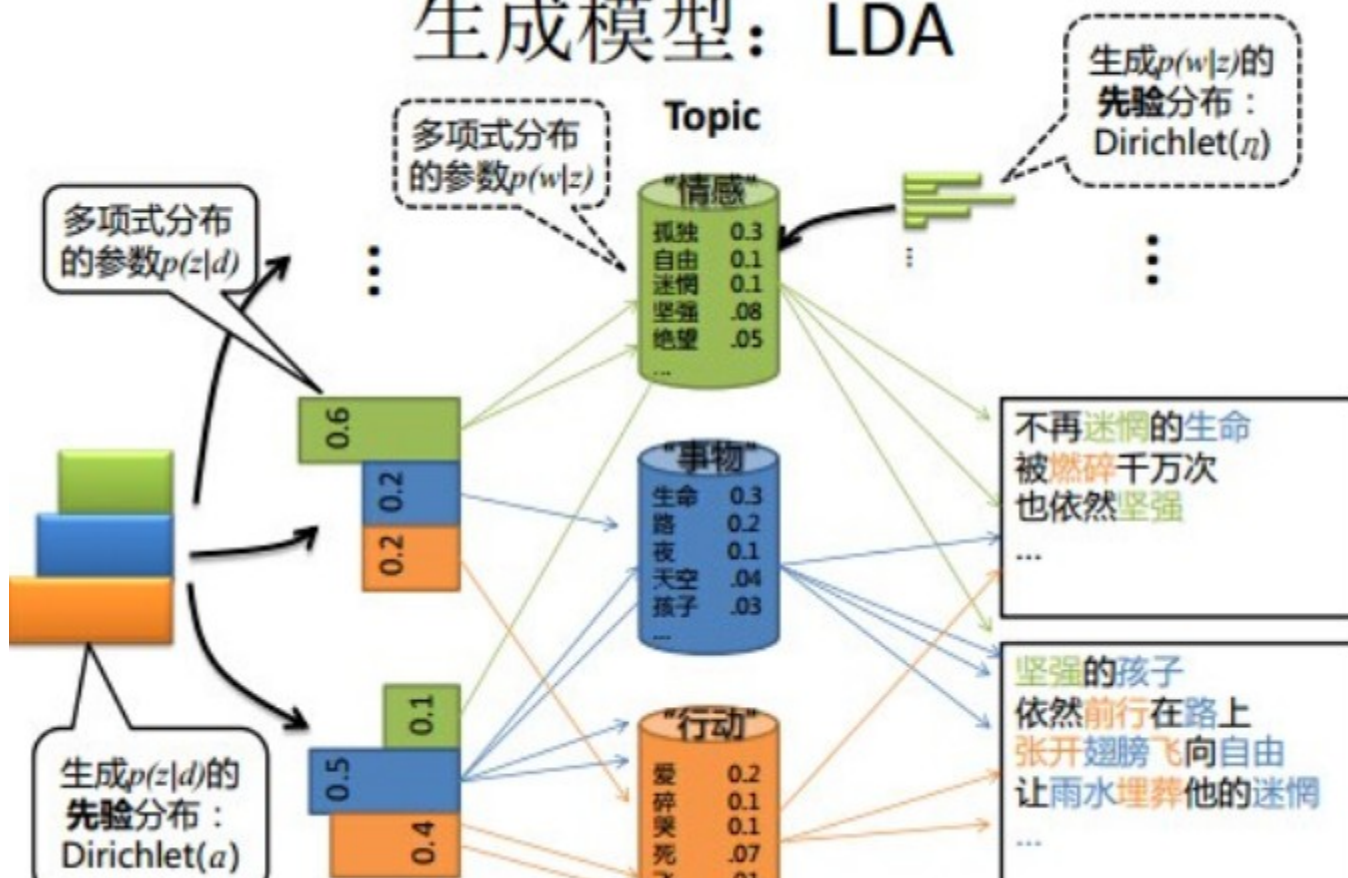
## LDA Model

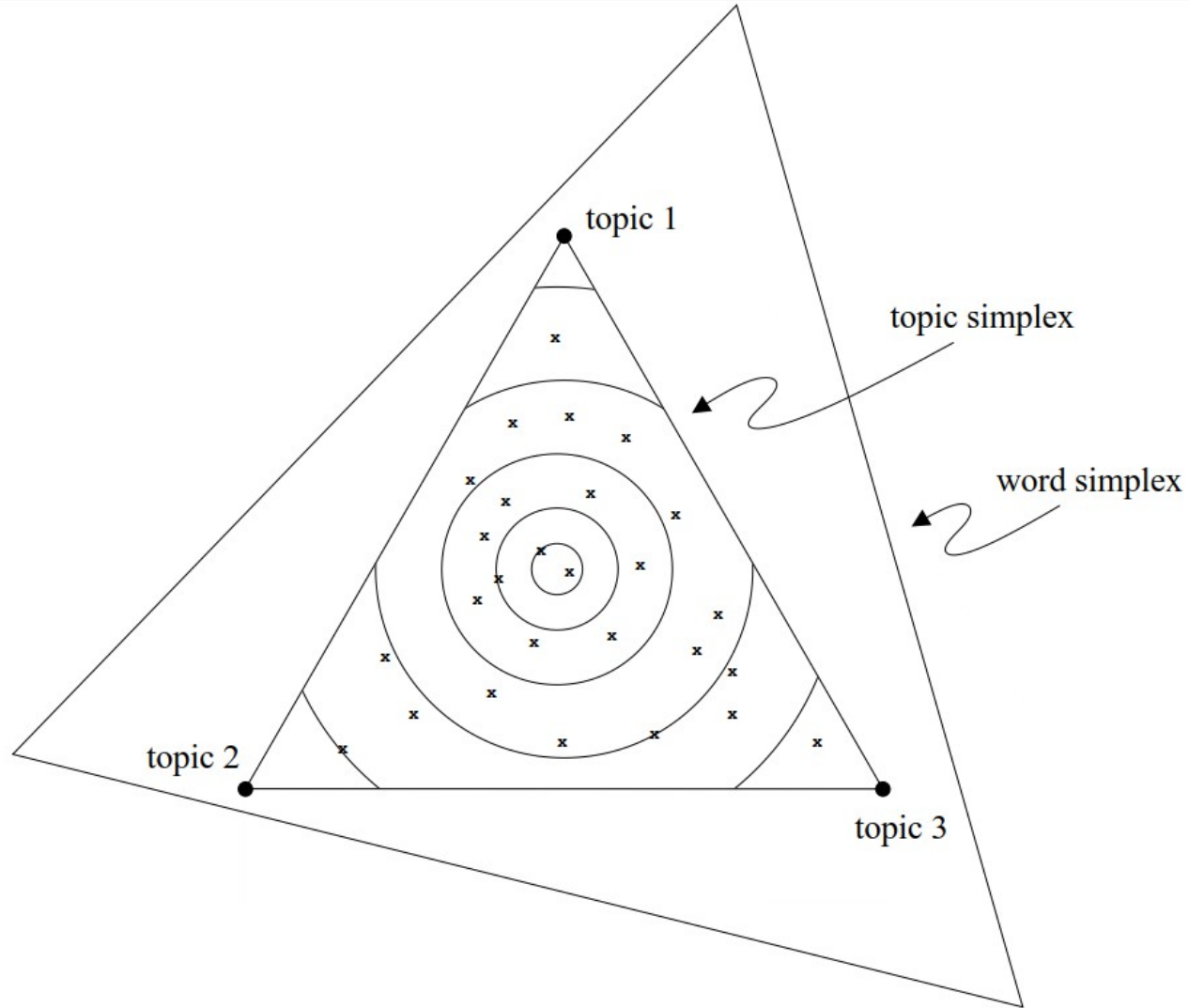
- 從狄利克雷分布 $\alpha$ 中取樣生成文檔 $i$ 的主題分布 $\theta_i$
- 從主題的多項式分布 $\theta_i$ 中取樣生成文檔 $i$ 第 $j$ 個詞的主題 $z_{i,j}$
- 從狄利克雷分布 $\beta$ 中取樣生成主題 $z_{i,j}$ 的詞語分布 $\phi_{z_{i,j}}$
- 從詞語的多項式分布 $\phi_{z_{i,j}}$ 中採樣最終生成詞語 $w_{i,j}$

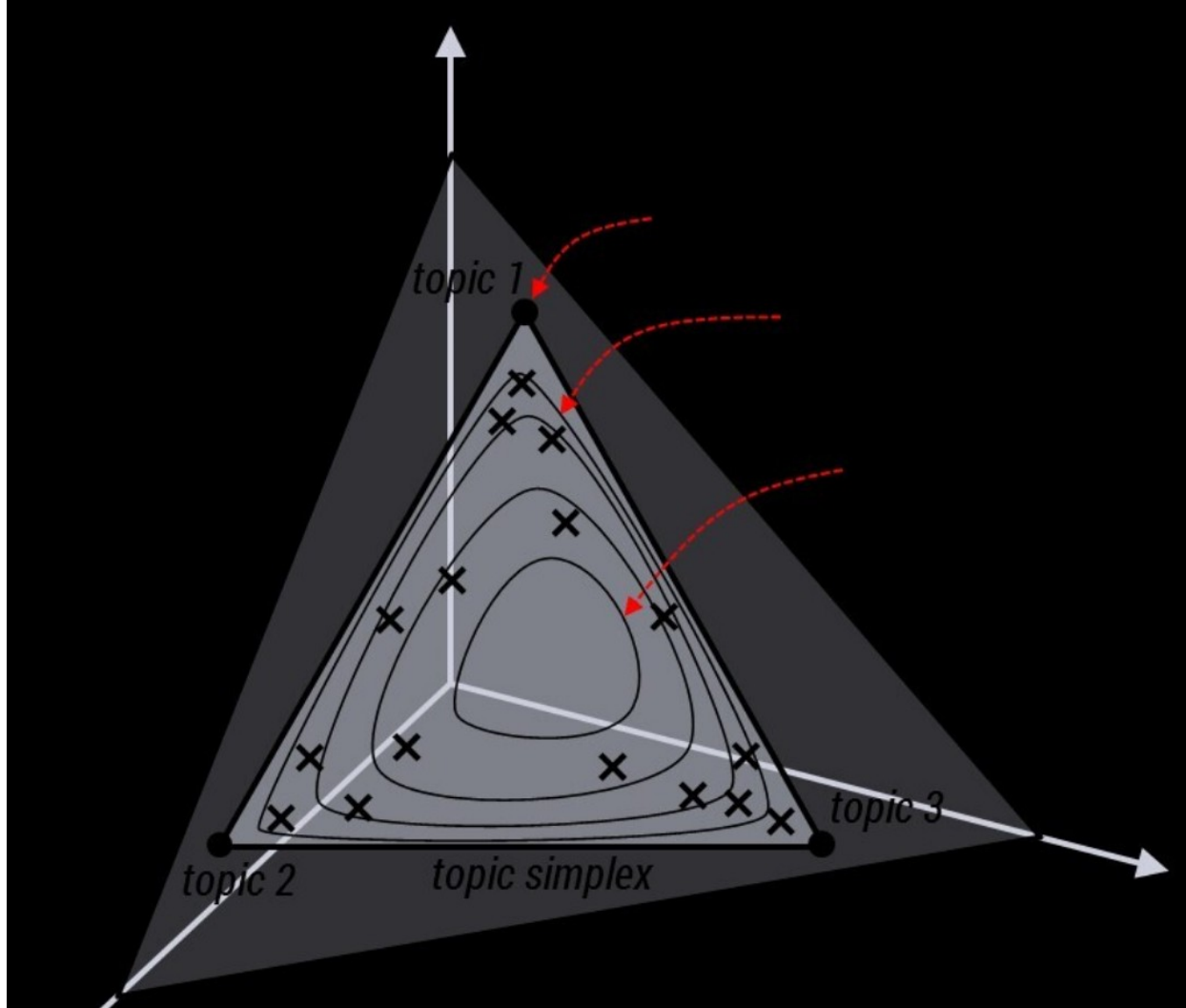
## LAD v.s. pLSI

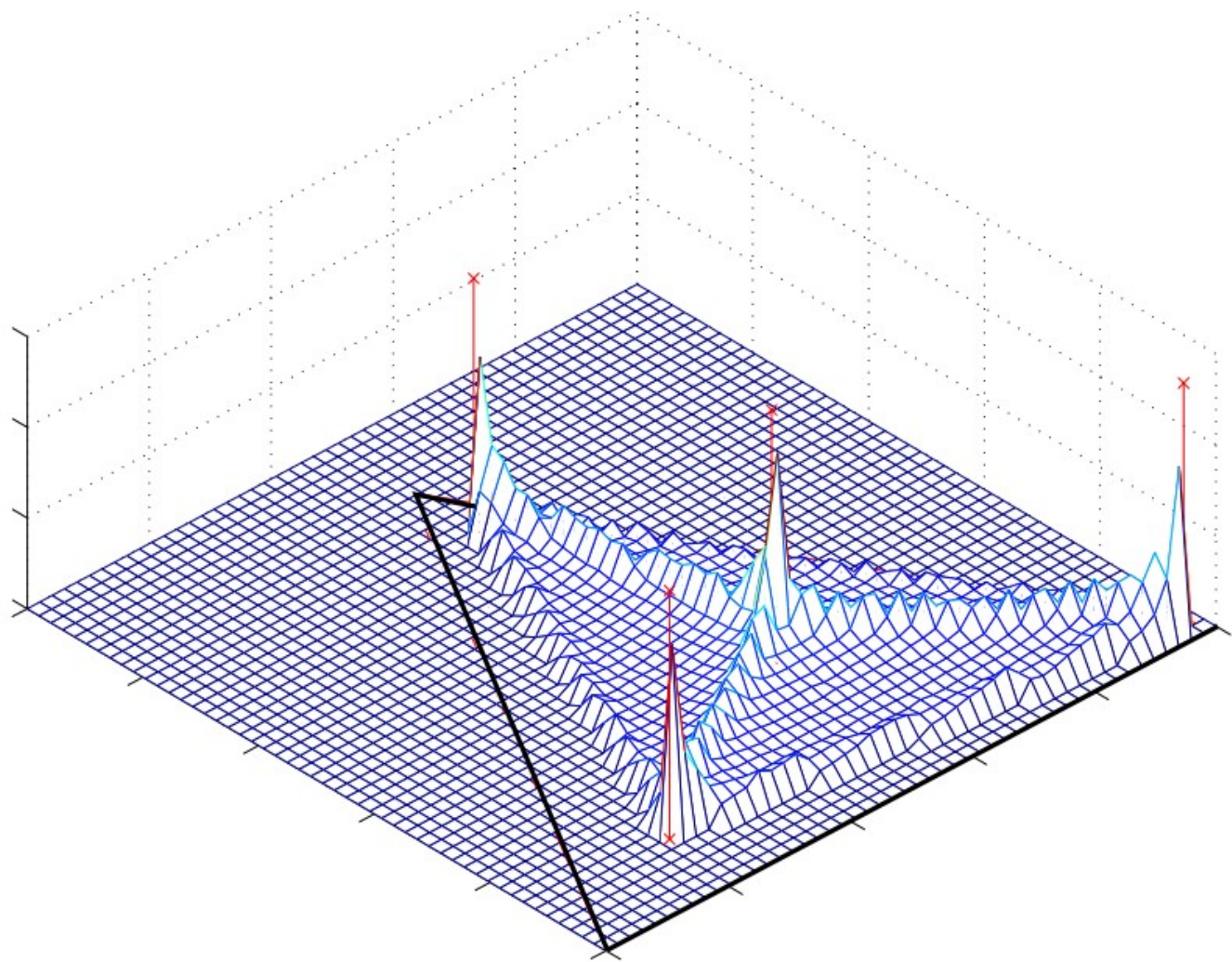
- LDA 中，主题分布和词分布不再唯一确定不變

# 生成模型：LDA



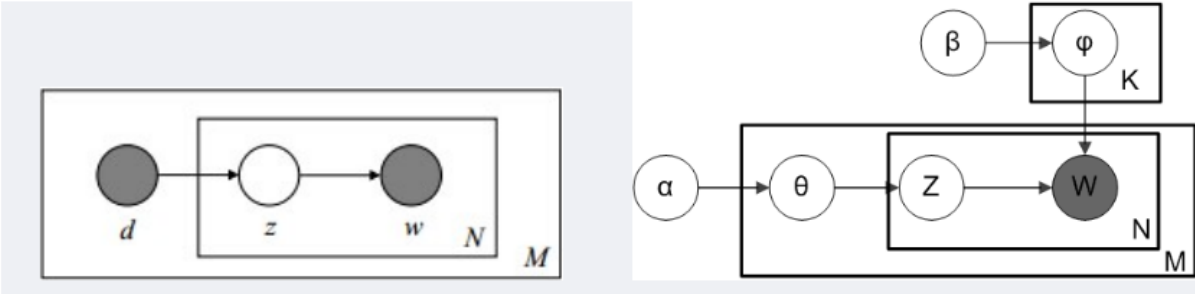






### 4.3.3 pLSA跟LDA的概率图对比

接下来，对比下LDA跟pLSA的概率模型图模型，左图是pLSA，右图是LDA（右图不太规范，z跟w都得是小写，其中，阴影圆圈表示可观测的变量，非阴影圆圈表示隐变量，箭头表示两变量间的条件依赖性conditional dependency，方框表示重复抽样，方框右下角的数字代表重复抽样的次数）：



对应到上面右图的LDA，只有 $W / w$ 是观察到的变量，其他都是隐变量或者参数，其中， $\Phi$ 表示词分布， $\Theta$ 表示主题分布， $\alpha$ 是主题分布 $\Theta$ 的先验分布（即Dirichlet分布）的参数， $\beta$ 是词分布 $\Phi$ 的先验分布（即Dirichlet分布）的参数， $N$ 表示文档的单词总数， $M$ 表示文档的总数。

所以，对于一篇文档 $d$ 中的每一个单词，LDA根据先验知识 $\alpha$ 确定某篇文档的主题分布 $\theta$ ，然后从该文档所对应的多项分布（主题分布） $\theta$ 中抽取一个主题 $z$ ，接着根据先验知识 $\beta$ 确定当前主题的词分布 $\phi$ ，然后从主题 $z$ 所对应的多项分布（词分布） $\phi$ 中抽取一个单词 $w$ 。然后将这个过程重复 $N$ 次，就产生了文档 $d$ 。

1. 假定语料库中共有M篇文章，每篇文章下的Topic的**主题分布**是一个**从参数为 $\alpha$ 的Dirichlet先验分布中采样得到的Multinomial分布**，每个Topic下的**词分布**是一个**从参数为 $\beta$ 的Dirichlet先验分布中采样得到的Multinomial分布**。
2. 对于某篇文章中的第n个词，首先从该文章中出现的每个主题的多项式分布（**主题分布**）**中选择或采样一个主题**，然后再在这个主题**对应的词**的多项式分布（**词分布**）**中选择或采样一个词**。不断重复这个随机生成过程，直到M篇文章全部生成完成。



综上，M 篇文档会对应于 M 个独立的 Dirichlet-Multinomial 共轭结构，K 个 topic 会对应于 K 个独立的 Dirichlet-Multinomial 共轭结构。

- 其中， $\alpha \rightarrow \theta \rightarrow z$  表示生成文档中的所有词对应的主题，显然  $\alpha \rightarrow \theta$  对应的是 Dirichlet 分布， $\theta \rightarrow z$  对应的是 Multinomial 分布，所以整体是一个 Dirichlet-Multinomial 共轭结构，如下图所示：

$$\vec{\alpha} \xrightarrow{\text{Dirichlet}} \vec{\theta}_m \xrightarrow{\text{Multinomial}} \vec{z}_m$$

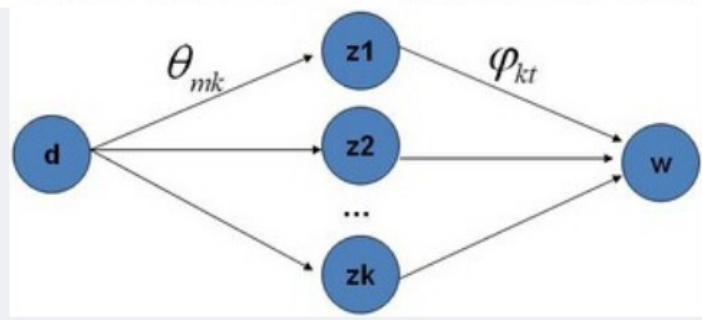
- 类似的， $\beta \rightarrow \varphi \rightarrow w$ ，容易看出，此时  $\beta \rightarrow \varphi$  对应的是 Dirichlet 分布， $\varphi \rightarrow w$  对应的是 Multinomial 分布，所以整体也是一个 Dirichlet-Multinomial 共轭结构，如下图所示：

$$\vec{\beta} \xrightarrow{\text{Dirichlet}} \vec{\varphi}_k \xrightarrow{\text{Multinomial}} \vec{w}_{(k)}$$

## LDA 後驗分布 -Gibbs 採樣

- 马尔可夫链蒙特卡尔理论 ( MCMC ) 中用来获取一系列近似等于指定多维概率分布 ( 比如 2 个或者多个随机变量的联合概率分布 ) 观察样本的算法。
- 馬可夫鏈蒙地卡羅 ( 英語 : **Markov chain Monte Carlo** , **MCMC** ) 方法 ( 含 隨機漫步蒙地卡羅 方法 ) 是一組用馬氏鏈從隨機分布取樣的演算法，之前步驟的作為底本。

仔细观察上述结果，可以发现，式子的右半部分便是  $p(topic|doc) \cdot p(word|topic)$ ，这个概率的值对应着  $doc \rightarrow topic \rightarrow word$  的路径概率。如此，K 个 topic 对应着 K 条路径，Gibbs Sampling 便在这 K 条路径中进行采样，如下图所示：



何等奇妙，就这样，Gibbs Sampling 通过求解出主题分布和词分布的后验分布，从而成功解决主题分布和词分布这两参数未知的问题。

## “Arts”

## “Budgets”

## “Children”

## “Education”

NEW  
FILM  
SHOW  
MUSIC  
MOVIE  
PLAY  
MUSICAL  
BEST  
ACTOR  
FIRST  
YORK  
OPERA  
THEATER  
ACTRESS  
LOVE

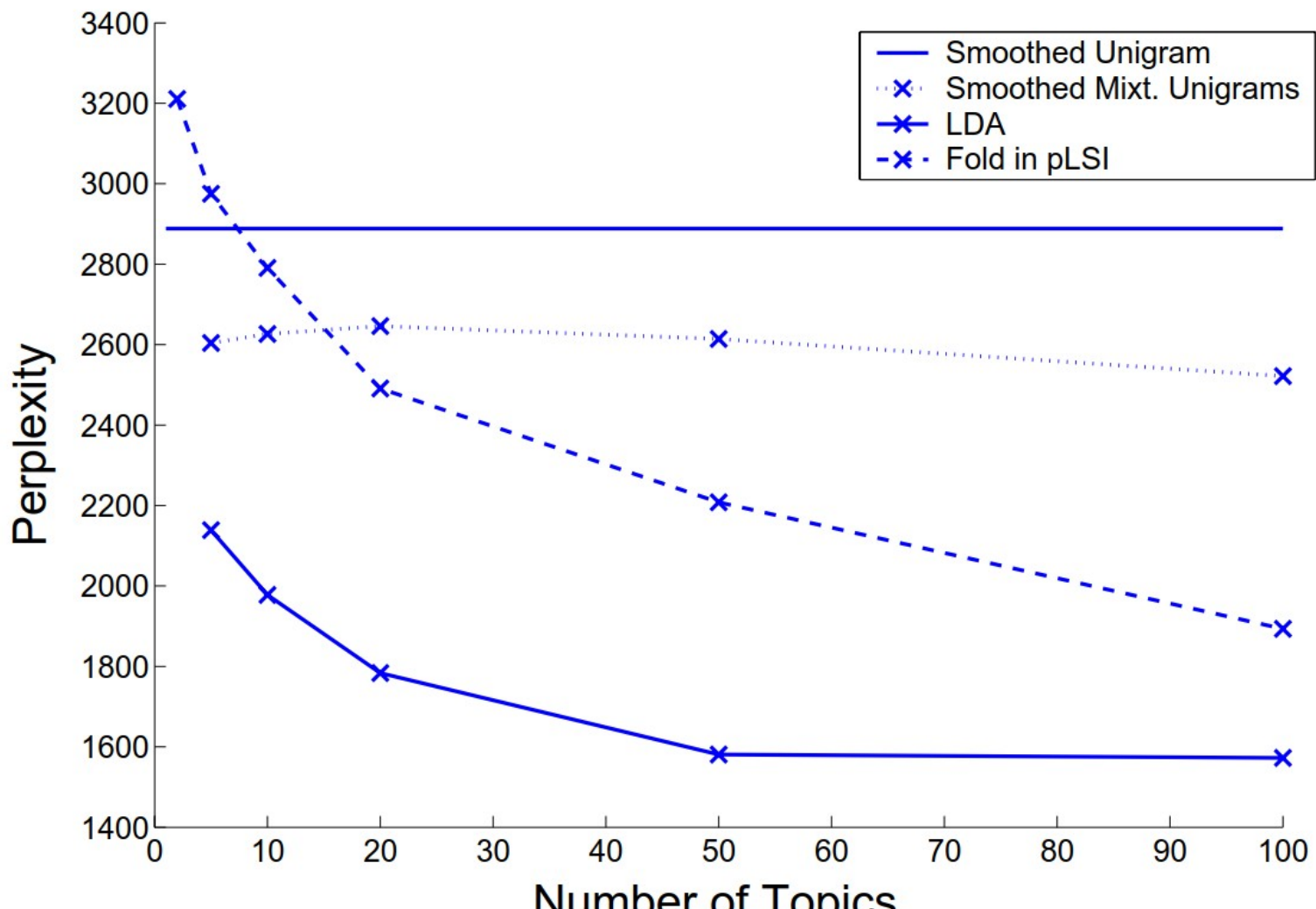
MILLION  
TAX  
PROGRAM  
BUDGET  
BILLION  
FEDERAL  
YEAR  
SPENDING  
NEW  
STATE  
PLAN  
MONEY  
PROGRAMS  
GOVERNMENT  
CONGRESS

CHILDREN  
WOMEN  
PEOPLE  
CHILD  
YEARS  
FAMILIES  
WORK  
PARENTS  
SAYS  
FAMILY  
WELFARE  
MEN  
PERCENT  
CARE  
LIFE

SCHOOL  
STUDENTS  
SCHOOLS  
EDUCATION  
TEACHERS  
HIGH  
PUBLIC  
TEACHER  
BENNETT  
MANIGAT  
NAMPHY  
STATE  
PRESIDENT  
ELEMENTARY  
HAITI

# Document modeling

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.



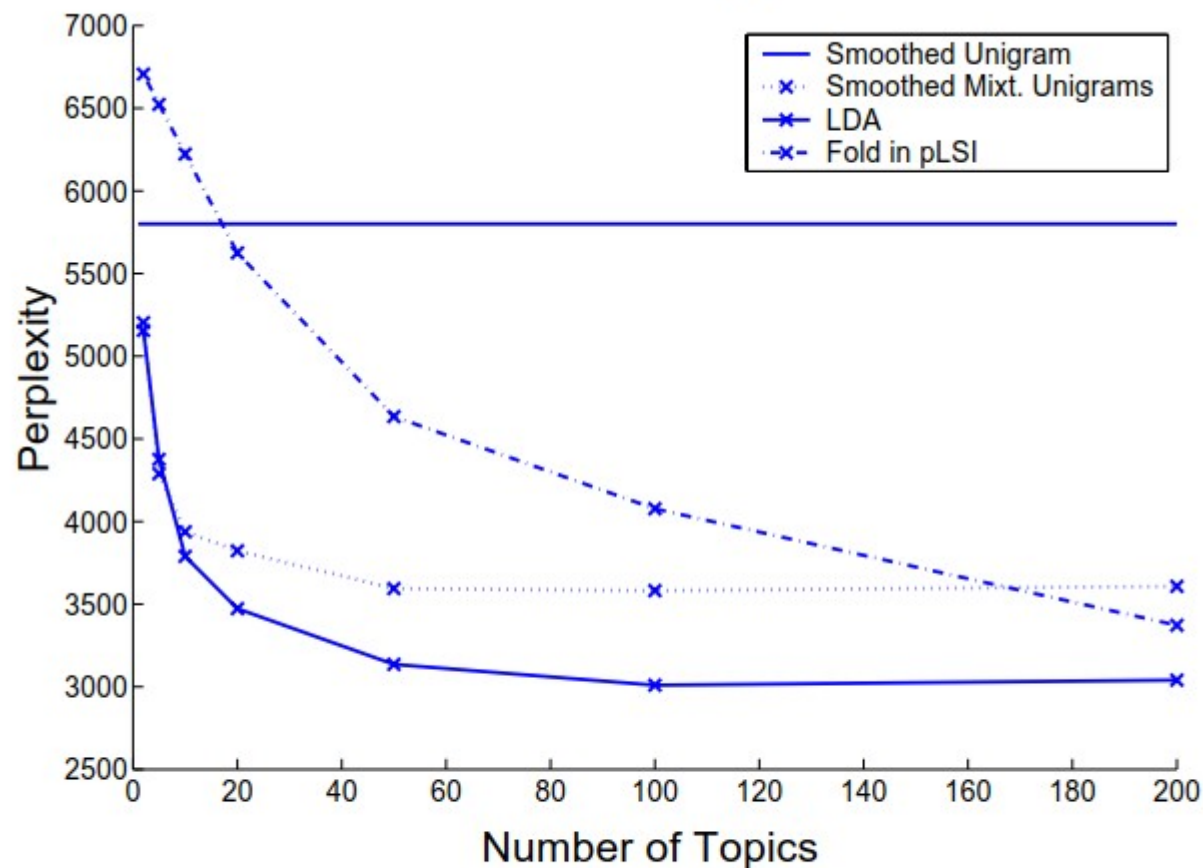


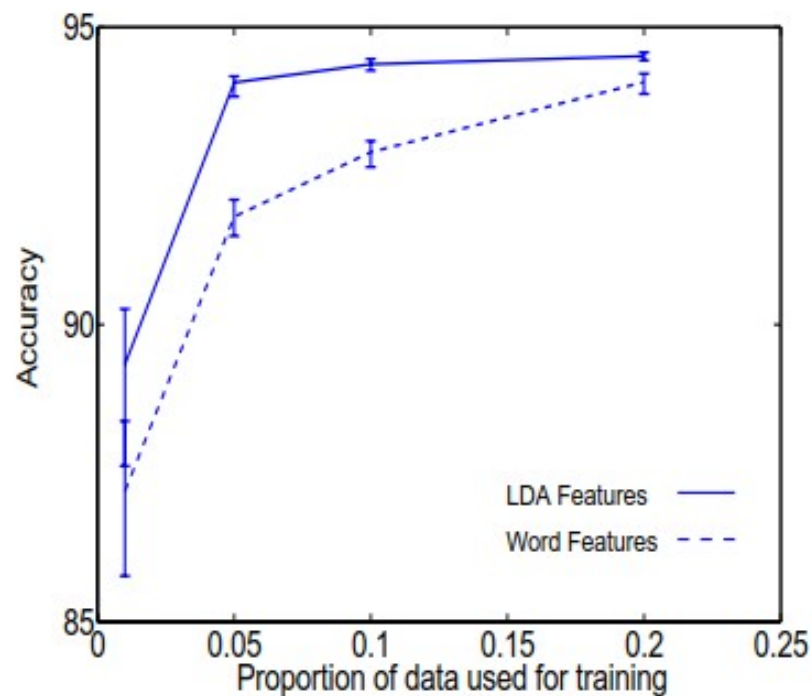
Figure 9: Perplexity results on the nematode (Top) and AP (Bottom) corpora for LDA, the unigram model, mixture of unigrams, and pLSI.



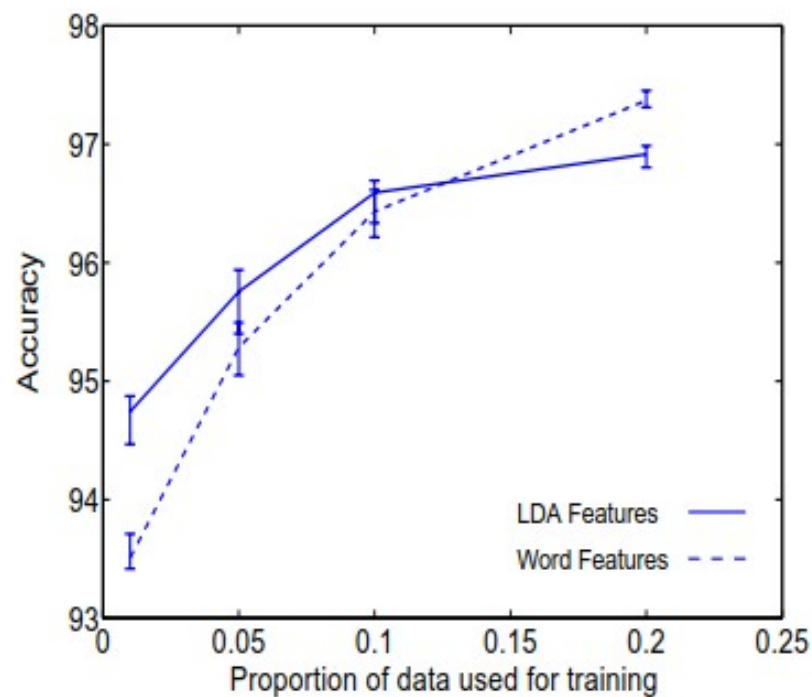
Num. topics ( $k$ )	Perplexity (Mult. Mixt.)	Perplexity (pLSI)
2	22,266	7,052
5	$2.20 \times 10^8$	17,588
10	$1.93 \times 10^{17}$	63,800
20	$1.20 \times 10^{22}$	$2.52 \times 10^5$
50	$4.19 \times 10^{106}$	$5.04 \times 10^6$
100	$2.39 \times 10^{150}$	$1.72 \times 10^7$
200	$3.51 \times 10^{264}$	$1.31 \times 10^7$

Table 1: Overfitting in the mixture of unigrams and pLSI models for the AP corpus. Similar behavior is observed in the nematode corpus (not reported).





(a)



(b)

Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

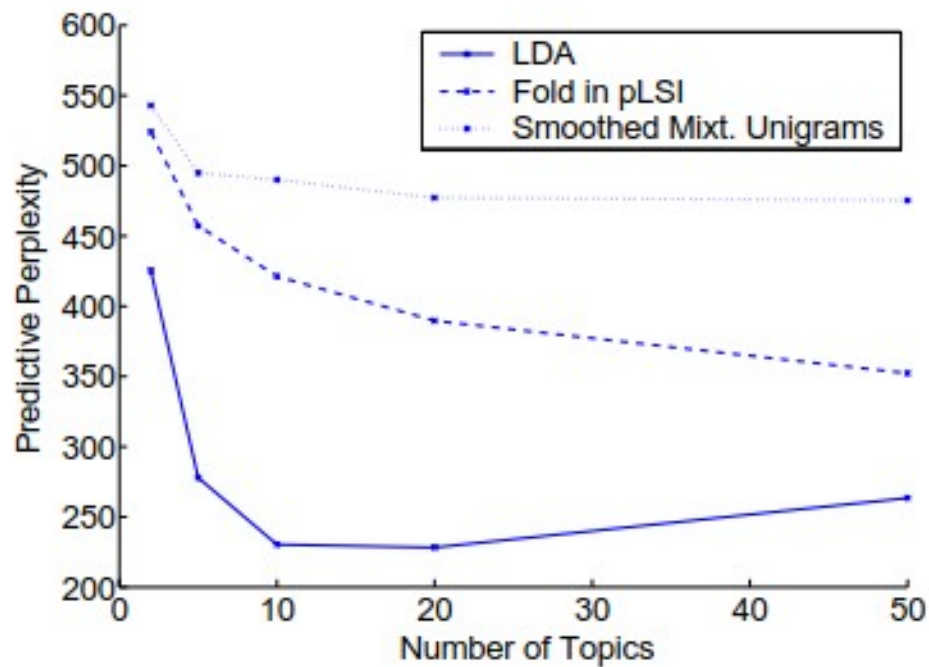


Figure 11: Results for collaborative filtering on the EachMovie data.

## 結論

a variety of extensions of LDA can be considered in which the distributions on the topic variables are elaborated. For example, we could arrange the topics in a time series, essentially relaxing the full exchangeability assumption to one of partial exchangeability. We could also consider partially exchangeable models in which we condition on exogenous variables; thus, for example, the topic distribution could be conditioned on features such as “paragraph” or “sentence,” providing a more powerful text model that makes use of information obtained from a parser.

Q & A

## 來源

- 通俗理解LDA主题模型
- Latent Dirichlet Allocation (隱狄利克雷分配模型)
- 生成模型與文字探勘：利用 LDA 建立文件主題模型