信息检索与数据挖掘

补充：概率图及主题模型

课程内容

- 第1章 绪论
- 第2章 布尔检索及倒排索引
- 第3章 词项词典和倒排记录表
- 第4章 索引构建和索引压缩
- 第5章 向量模型及检索系统
- 第6章 检索的评价
- 第7章 相关反馈和查询扩展
- 第8章 概率模型
- 第9章 基于语言建模的检索模型
- 第10章 文本分类
- 第11章 文本聚类
- 补充：概率图及主题模型
- 补充：数据挖掘经典算法概述
- 第12章 Web搜索
- 第13章 多媒体信息检索
- ~~第14章 其他应用简介~~

后续课程安排

- 4月29日，补充：概率图及主题模型
- 5月6日，补充：数据挖掘经典算法概述(1)
- 5月8日，补充：数据挖掘经典算法概述(2)
- 5月13日，第12章 Web搜索
- 5月15日，第13章 多媒体信息检索
- 5月20日，复习
- 5月22日，同学们文献阅读报告
- 5月27日，同学们文献阅读报告
- 6月3日，期末考试【暂定】

概率图及主题模型
Probabilistic Graphical Models / Topic Model

- 什么是Graphical Model
  - 定义、示例
  - Representation、Inference、Learning
- 主题模型与分类
  - LSA (Latent Semantic Analysis), 1990
  - pLSA (probabilistic Latent Semantic Analysis), 1999
  - LDA(Latent Dirichlet Allocation), 2003
  - Hierarchical Bayesian model
- 主题模型的R语言实现示例

主要参考书目

- **Probabilistic Graphical Models** (Principles and Applications)
- Luis Enrique Sucar, 2015
- **Probabilistic Graphical** Models (Principles and Techniques)
- Daphne Koller & Nir Friedman, 2009

- **Pattern Recognition and Machine Learning**
- Christopher M.Bishop, 2006

概率图及主题模型
Probabilistic Graphical Models / Topic Model

- 什么是Graphical Model
  - 定义、示例
  - Representation、Inference、Learning
- 主题模型与分类
  - LSA (Latent Semantic Analysis), 1990
  - pLSA (probabilistic Latent Semantic Analysis), 1999
  - LDA(Latent Dirichlet Allocation), 2003
  - Hierarchical Bayesian model
- 主题模型的R语言实现示例

Graphical Model（概率图模型）
Probabilistic Graphical Models (PGMs)

- 概率图模型是一类用图形模式表达基于概率相关关系的模型的总称。概率图模型结合概率论与图论的知识，利用图来表示与模型有关的变量的联合概率分布。近10年它已成为不确定性推理的研究热点，在人工智能、机器学习和计算机视觉等领域有广阔的应用前景。
- 概率图理论共分为三个部分
  - Representation：概率图模型表示理论
  - Inference：概率图模型推理理论
  - Learning：概率图模型学习理论

概率图示例

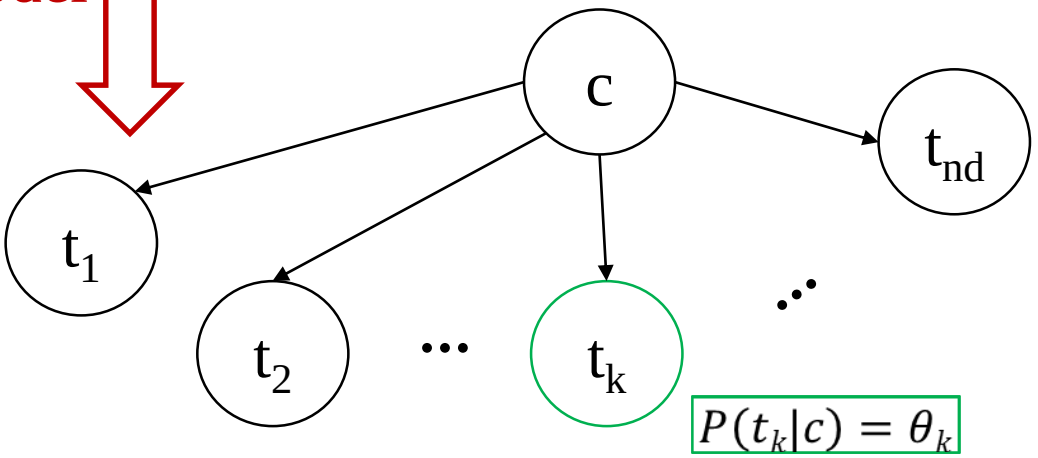
朴素贝叶斯分类器→概率图

独立性的假设

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

Bayes公式：P(c|d) → P(d|c)

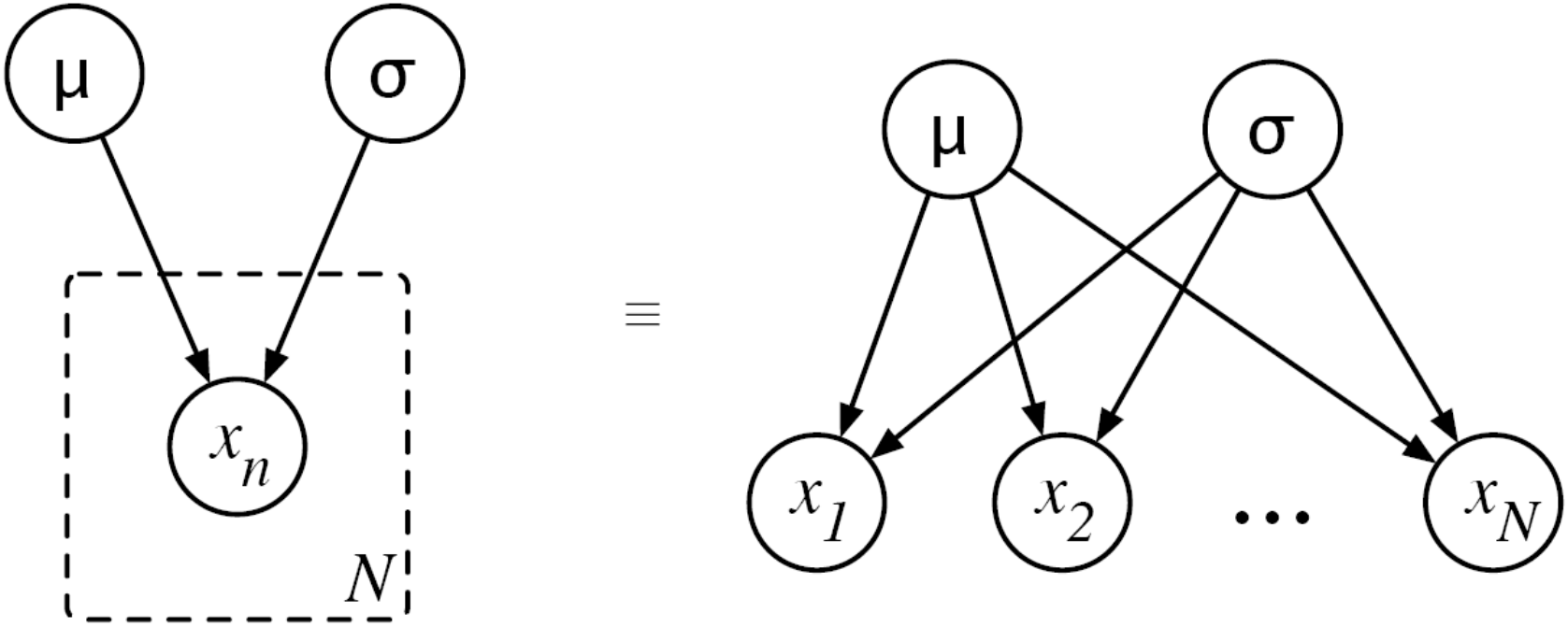
Graphical Model



概率图示例
用有向图表示统计模型

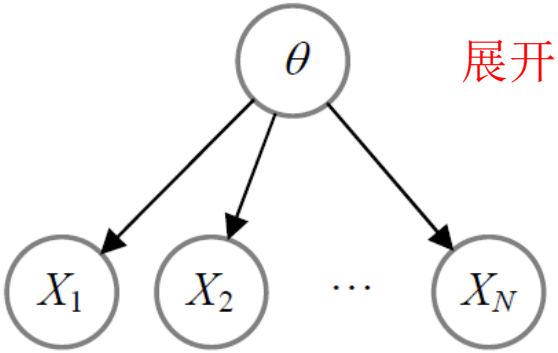
- A data set of N points generated from a Gaussian:

$$p(x_1, \ldots, x_N, \mu, \sigma) = p(\mu)p(\sigma) \prod_{n=1}^{N} p(x_n|\mu, \sigma)$$

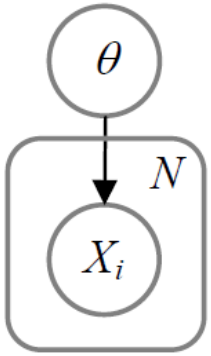


盘式记法

概率图的记法

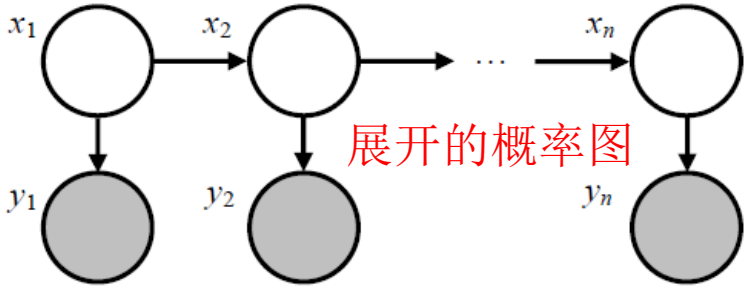


展开的概率图

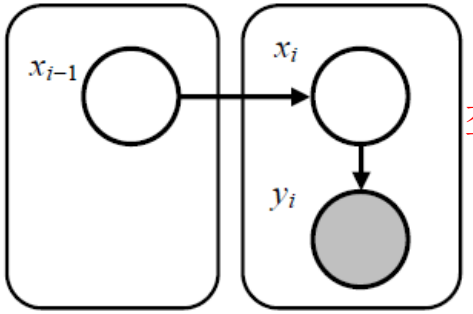


盘式记法

盘式记法(plate notation)是一种常用的图模型的简化记法。在盘式模型中，用一个框(称为盘)圈住图模型中重复的部分，并在框内标注重复的次数。盘式记法能够为我们表示和分析许多概率模型提供很大的方便，但它也有一定的局限性。例如，它无法表示盘内变量不同拷贝间的相关性，而这种相关性广泛出现于动态贝叶斯网络中。

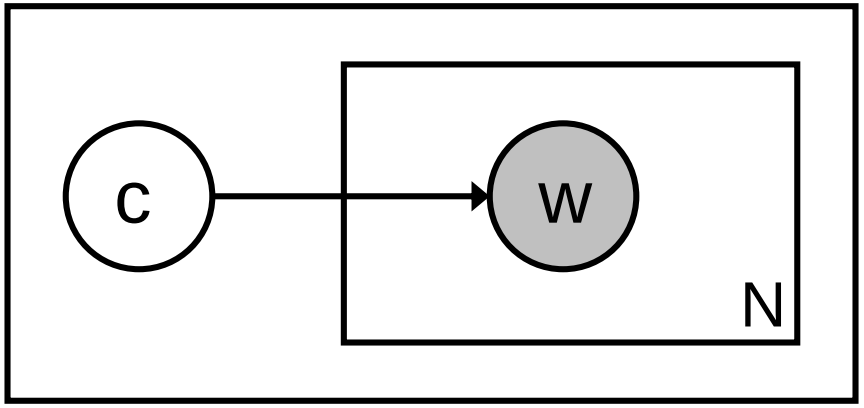


展开的概率图



盘式记法

Naïve Bayes model（盘式记法）



模型参数：使后验概率p(c|d)最大的参数集

参数集：

$p(c_i)$,i=1,2,…,类别总数

$p(w_k|c)$, j=1,2,…,词项总数

$$c^* = \arg\max_c \; p(c \mid w) \propto p(c)\, p(w \mid c) \quad = p(c) \prod_{n=1}^{N} p(w_n \mid c)$$

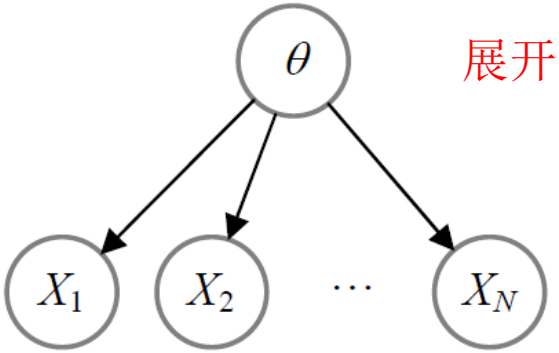
Object class decision

Prior prob. of the object classes

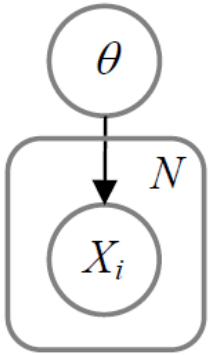
Image likelihood given the class

小结：什么是概率图模型

- Graphical Model（概率图模型）
  - 是一类用图形模式表达基于概率相关关系的模型的总称。
- 概率图的表示方法



展开的概率图



盘式记法

- 概率图求解→优化
  - 概率图对应参数的求解（如朴素贝叶斯分类器中的参数）

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c)$$

概率图模型的表示 Representation

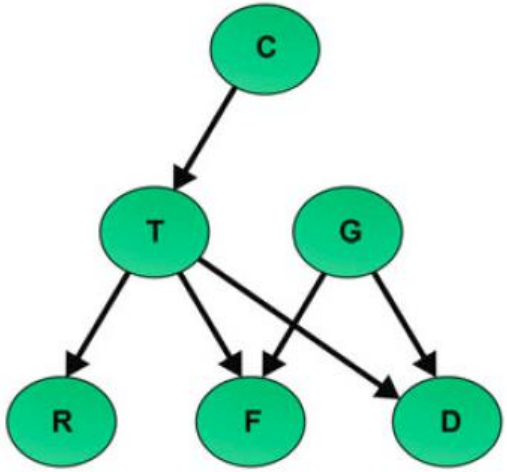
结构：G(V,E)
参数：CPTs

A **Bayesian network (BN)** represents the joint distribution of a set of n (discrete) variables, $X_1, X_2, \ldots, X_n$, as a directed acyclic graph (DAG) and a set of conditional probability tables (CPTs). Each node, that corresponds to a variable, has an associated CPT that contains the probability of each state of the variable given its parents in the graph. The structure of the network implies a set of conditional independence assertions, which give power to this representation.

A PGM is specified by two aspects: (i) a graph, $G(V, E)$, that defines the structure of the model; and (ii) a set of **local functions**, $f(Y_i)$, that define the parameters, where Y_i is a subset of X. The joint probability is obtained by the product of the local functions:

$$P(X_1, X_2, \ldots, X_N) = K \prod_{i=1}^{M} f(Y_i)$$

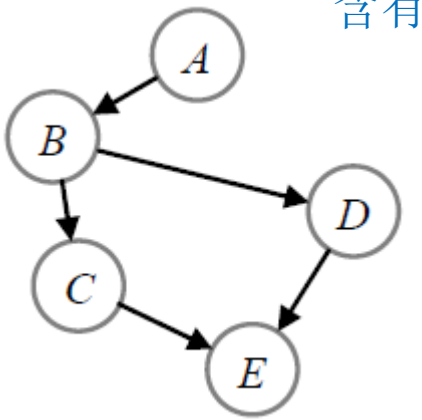
where K is a normalization constant. This representation in terms of a graph and a set of local functions (called ***potentials***) is the basis for inference and learning in PGMs.



Advances in Computer Vision and Pattern Recognition
Luis Enrique Sucar, **Probabilistic Graphical Models (Principles and Applications)**, 2015

概率图模型的推理
Inference

含有5个变量的贝叶斯网络及其表示的联合分布



$$P(A,B,C,D,E)=P(A)P(B|A)P(C|B)P(D|B)P(E|C,D)$$

如果观测到变量$E=e$，←*给定证据(Evidence)*
想要计算变量$C=c$的条件概率$P(c|e)$，←*推理(inference)*
则

$$P(c \mid e) = \frac{1}{Z}\sum_{a,b,d} P(a,b,c,d,e),$$

$$Z = \sum_{a,b,c,d} P(a,b,c,d,e).$$

精确推理
近似推理

概率图模型的学习 Learning

- 概率图结构已知，即为参数的学习（估计）
  - 常用的学习方法有两类：最大似然估计（MLE）、贝叶斯估计。前者视模型参数为定值，后者视其为随机变量。
  - MLE在数据完备的情况下，可将参数学习问题转化为充分统计量的计算问题，在数据不完备的情况下，采用EM算法，用迭代方式逐步最大化$p(x \mid \theta)$。
  - 贝叶斯估计在数据完备的情况下，根据误差准则不同，可以诱导出最大后验估计或者后验均值的估计方法，在数据不完备的情况下，可以将 θ 视为一种特殊的隐变量，从而问题归结为推理问题，可以采用变分贝叶斯方法近似求解。
- 概率图结构未知
  - 数据完备时，较好的方式是定义一个得分函数，评估结构与数据的匹配程度，然后搜索最大得分的结构。实际中需要根据奥克姆剃刀原理，选择可以拟合数据的最简单模型。如果预先假定结构是树模型（每个节点至多有一个父节点），则搜索可在多项式时间内完成，否则是NP-hard问题。
  - 数据不完备，需考虑structural EM算法。

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c)$$

小结：表示、推理、学习
Representation, Inference, Learning

- Representation
  - a graph ← 结构：G(V,E)
  - a set of local functions (called potentials) ←参数：CPTs
- Inference
  - answering different probabilistic queries based on the model and some evidence.
  - obtain the marginal or conditional probabilities of any subset of variables *Z* given any other subset *Y*.
- Learning
  - given a set of data values for *X* (that can be incomplete) estimate the structure (graph) and parameters (local functions) of the model.

概率图模型的常见类型

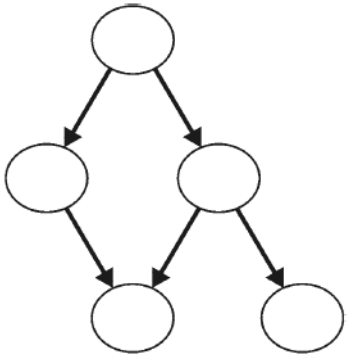
Directed Acyclic Graph
Undirected Graph

Table 1.3 Main types of probabilistic graphical models

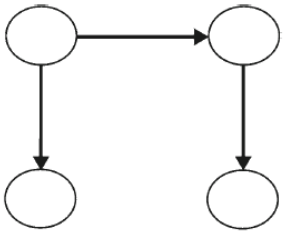
Type	**D**irected/**U**ndirected	**S**tatic/**D**ynamic	**P**rob./**D**ecisional
Bayesian classifiers	D/U	S	P
Markov chains	D	D	P
Hidden Markov models	D	D	P
Markov random fields	U	S	P
Bayesian networks	D	S	P
Dynamic Bayesian networks	D	D	P
Influence diagrams	D	S	D
Markov decision processes (MDPs)	D	D	D
Partially observable MDPs	D	D	D

Advances in Computer Vision and Pattern Recognition
Luis Enrique Sucar, Probabilistic Graphical Models (Principles and Applications)， 2015

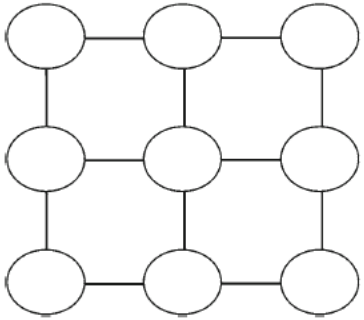
常见模型图示



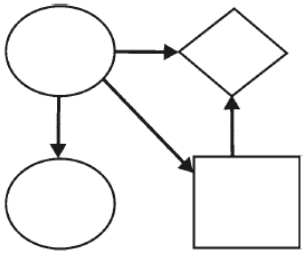
Bayesian Networks



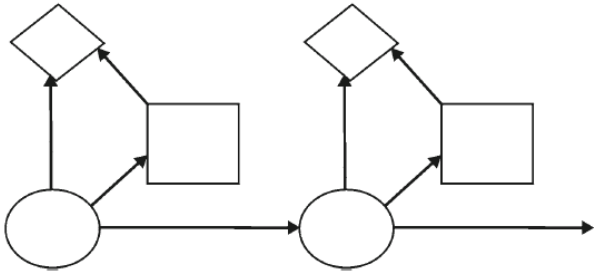
Hidden Markov Models



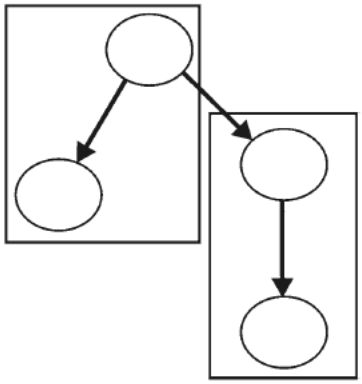
Markov Random Fields



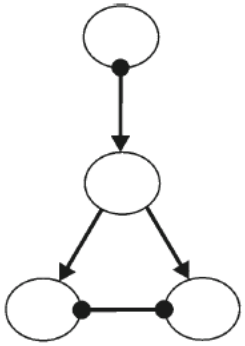
Decision Graphs



Markov Decision Processes



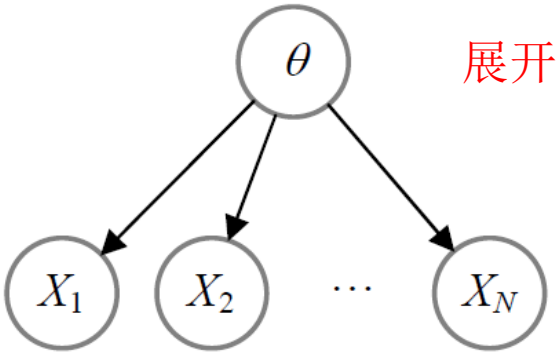
Relational Probabilistic Graphical Models



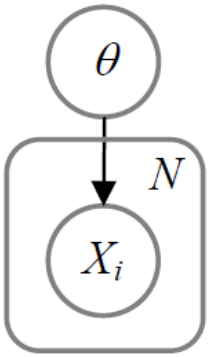
Graphical Causal Models

小结：什么是Graphical Model

- **Graphical Model（概率图模型）**
- **Probabilistic Graphical Models (PGMs)**



展开的概率图



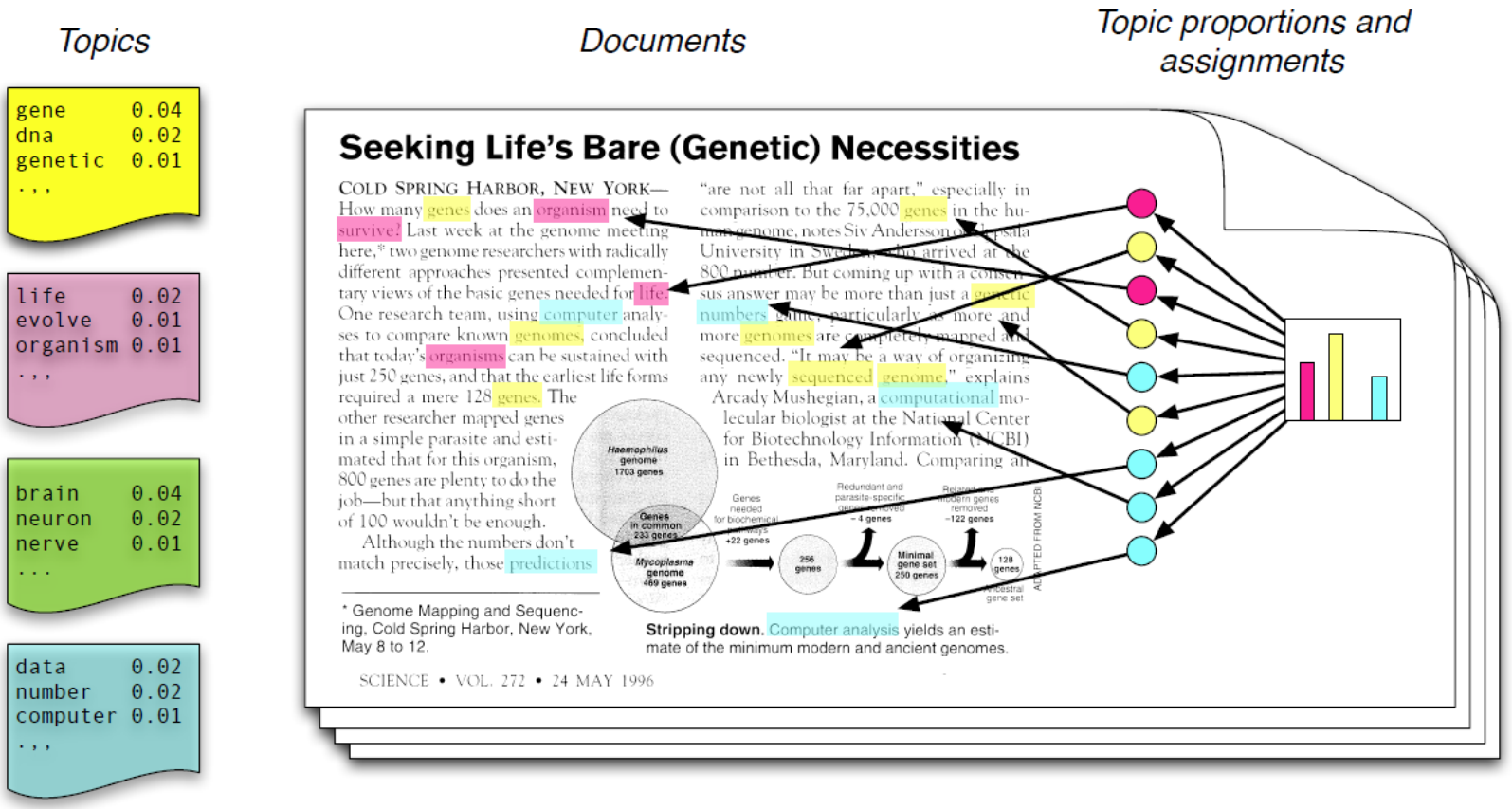
盘式记法

- 概率图理论共分为三个部分
  - Representation、Inference、Learning
- 概率图模型的常见类型
  - 贝叶斯网络采用有向无环图(Directed Acyclic Graph)
  - 马尔可夫随机场则采用无向图(Undirected Graph)

概率图及主题模型
Probabilistic Graphical Models / Topic Model

- 什么是Graphical Model
  - 定义、示例
  - Representation、Inference、Learning
- 主题模型与分类
  - LSA (Latent Semantic Analysis), 1990
  - pLSA (probabilistic Latent Semantic Analysis), 1999
  - LDA(Latent Dirichlet Allocation), 2003
  - Hierarchical Bayesian model
- 主题模型的R语言实现示例

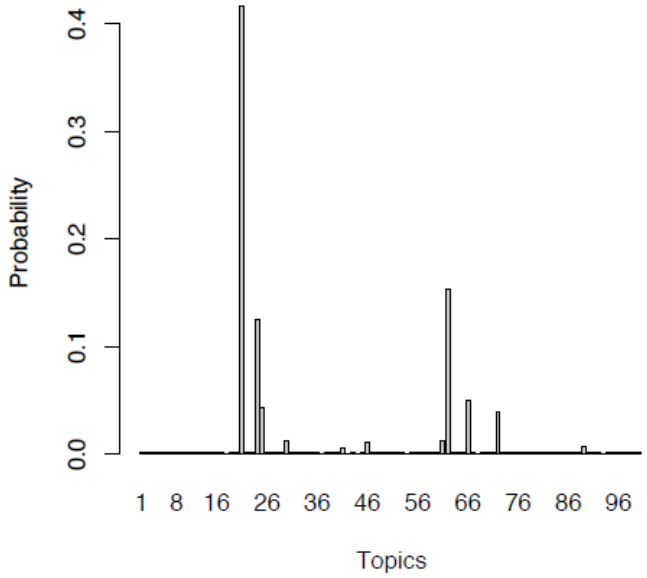
什么是主题模型？概念示意



We assume that some number of \topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this gure are illustrative - they are not fit from real data.

Probabilistic topic models, DM Blei, Communications of the ACM, 2012 Retrieved: 2017.04.06 Google cited: 1622

什么是主题模型？例子



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

We fit a **100-topic** LDA model to 17,000 articles from the journal Science. At left is the inferred topic proportions for the example article（上页图所示文章）. At right are the top **15 most frequent words** from the most frequent topics found in this article.

Probabilistic topic models, DM Blei, Communications of the ACM, 2012 Retrieved: 2017.04.06 Google cited: 1622

概率图及主题模型
Probabilistic Graphical Models / Topic Model

- 什么是Graphical Model
- 主题模型与分类
  - LSA (Latent Semantic Analysis), 1990
  - pLSA (probabilistic Latent Semantic Analysis), 1999
  - LDA(Latent Dirichlet Allocation), 2003
  - Hierarchical Bayesian model
- 主题模型的R语言实现示例

LSA(Latent Semantic Analysis)
词项-文档矩阵的SVD分解，发现相关文档

c1:	*Human* machine *interface* for Lab ABC *computer* applications
c2:	A *survey of user* opinion of *computer system response time*
c3:	The *EPS user interface* management *system*
c4:	*System* and *human system* engineering testing of *EPS*
c5:	Relation of *user*-perceived *response time* to error measurement
m1:	The generation of random, binary, unordered *trees*
m2:	The intersection *graph* of paths in *trees*
m3:	*Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4:	*Graph minors: A survey*

文档集

原始的Term-Document矩阵

Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R.(1990).
Indexing By Latent Semantic Analysis. Journal of the American Society For Information Science, 41, 391-407. 10

LSA(Latent Semantic Analysis)
词项-文档矩阵的SVD分解，发现相关文档

$T_0 =$

0.22	−0.11	0.29	−0.41	−0.11	−0.34	0.52	−0.06	−0.41
0.20	−0.07	0.14	−0.55	0.28	0.50	−0.07	−0.01	−0.11
0.24	0.04	−0.16	−0.59	−0.11	−0.25	−0.30	0.06	0.49
0.40	0.06	−0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	−0.17	0.36	0.33	−0.16	−0.21	−0.17	0.03	0.27
0.27	0.11	−0.43	0.07	0.08	−0.17	0.28	−0.02	−0.05
0.27	0.11	−0.43	0.07	0.08	−0.17	0.28	−0.02	−0.05
0.30	−0.14	0.33	0.19	0.11	0.27	0.03	−0.02	−0.17
0.21	0.27	−0.18	−0.03	−0.54	0.08	−0.47	−0.04	−0.58
0.01	0.49	0.23	0.03	0.59	−0.39	−0.29	0.25	−0.23
0.04	0.62	0.22	0.00	−0.07	0.11	0.16	−0.68	0.23
0.03	0.45	0.14	−0.01	−0.30	0.28	0.34	0.68	0.18

$S_0 =$

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

保留S_0的最大两个奇异值

$$C = U\Sigma V^T$$

$$X = T_0 S_0 D_0$$

$D_0 =$

0.20	−0.06	0.11	−0.95	0.05	−0.08	0.18	−0.01	−0.06
0.61	0.17	−0.50	−0.03	−0.21	−0.26	−0.43	0.05	0.24
0.46	−0.03	0.21	0.04	0.38	0.72	−0.24	0.01	0.02
0.54	−0.23	0.57	0.27	−0.21	−0.37	0.26	−0.02	−0.08
0.28	0.11	−0.51	0.15	0.33	0.03	0.67	−0.06	−0.26
0.00	0.19	0.10	0.02	0.39	−0.30	−0.34	0.45	−0.62
0.01	0.44	0.19	0.02	0.35	−0.21	−0.15	−0.76	0.02
0.02	0.62	0.25	0.01	0.15	0.00	0.25	0.45	0.52
0.08	0.53	0.08	−0.03	−0.60	0.36	−0.04	−0.07	−0.45

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R.(1990). Indexing By Latent Semantic Analysis. Journal of the American Society For Information Science, 41, 391-407. 10

LSA(Latent Semantic Analysis)
词项-文档矩阵的SVD分解，发现相关文档

Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	0
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	0
minors	0	0	0	0	0	0	0	1	0

原始的Term-Document矩阵X

X中human-C2值为0，因为C2中并不包含human单词，但是$\hat{X}$中human-C2为0.40，表明human和C2有一定的关系，为什么呢？因为C2: A survey of user opinion of computer system response time中包含user单词，和human是近似词，故human-C2的值被提高了。

保留S_0的最大两个奇异值重构的$\hat{X}$

$\hat{X} =$

0.16	0.40	0.38	0.47	0.18	−0.05	−0.12	−0.16	−0.09
0.14	0.37	0.33	0.40	0.16	−0.03	−0.07	−0.10	−0.04
0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
0.45	1.23	1.05	1.27	0.56	−0.07	−0.15	−0.21	−0.05
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.22	0.55	0.51	0.63	0.24	−0.07	−0.14	−0.20	−0.11
0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
−0.06	0.23	−0.14	−0.27	0.14	0.24	0.55	0.77	0.66
−0.06	0.34	−0.15	−0.30	0.20	0.31	0.69	0.98	0.85
−0.04	0.25	−0.10	−0.21	0.15	0.22	0.50	0.71	0.62

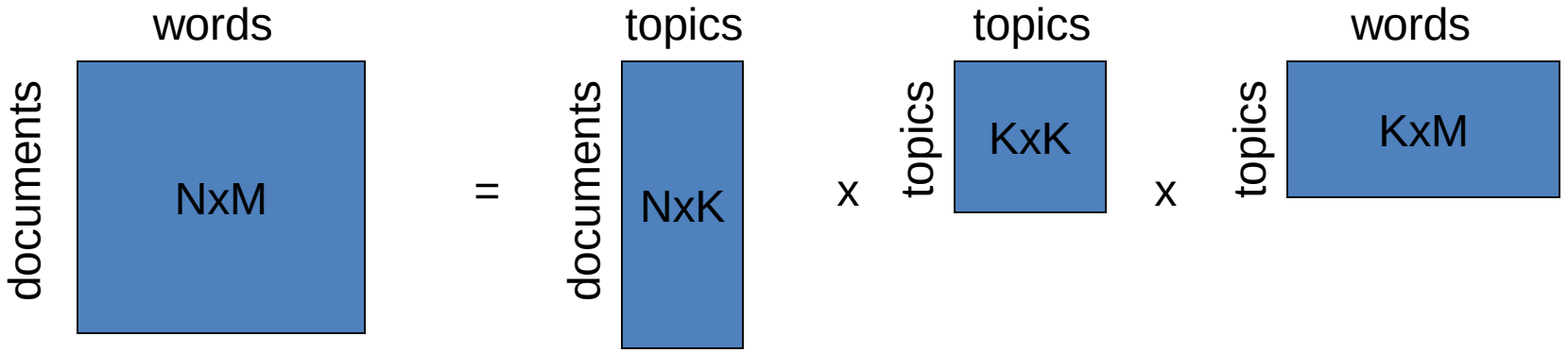
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R.(1990). Indexing By Latent Semantic Analysis. Journal of the American Society For Information Science, 41, 391-407. 10

小结：隐语义分析

1990: Latent Semantic Analysis (LSA)

- $D = \{d_1,\ldots,d_N\}$ N documents
- $W = \{w_1,\ldots,w_M\}$ M words
- $N_{ij} = \#(d_i,w_j)$ NxM co-occurrence term-document matrix

Singular Value Decomposition



概率图及主题模型
Probabilistic Graphical Models / Topic Model

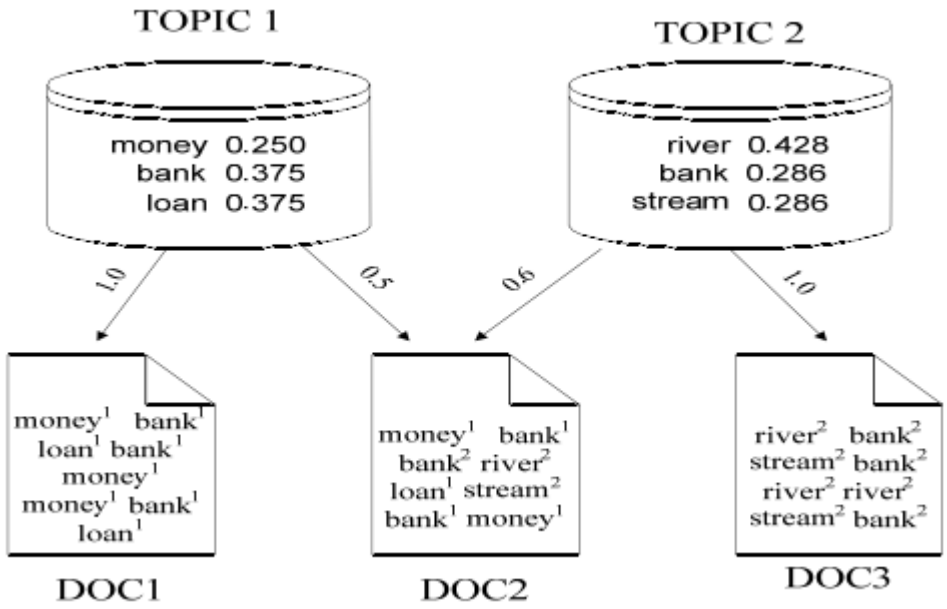
- 什么是Graphical Model
- 主题模型与分类
  - LSA (Latent Semantic Analysis), 1990
  - pLSA (probabilistic Latent Semantic Analysis), 1999
  - LDA(Latent Dirichlet Allocation), 2003
  - Hierarchical Bayesian model
- 主题模型的R语言实现示例

Probabilistic Latent Semantic Analysis， 1999
Thomas Hofmann, University of California, Berkeley, CA
2016.04 google cited: 1873

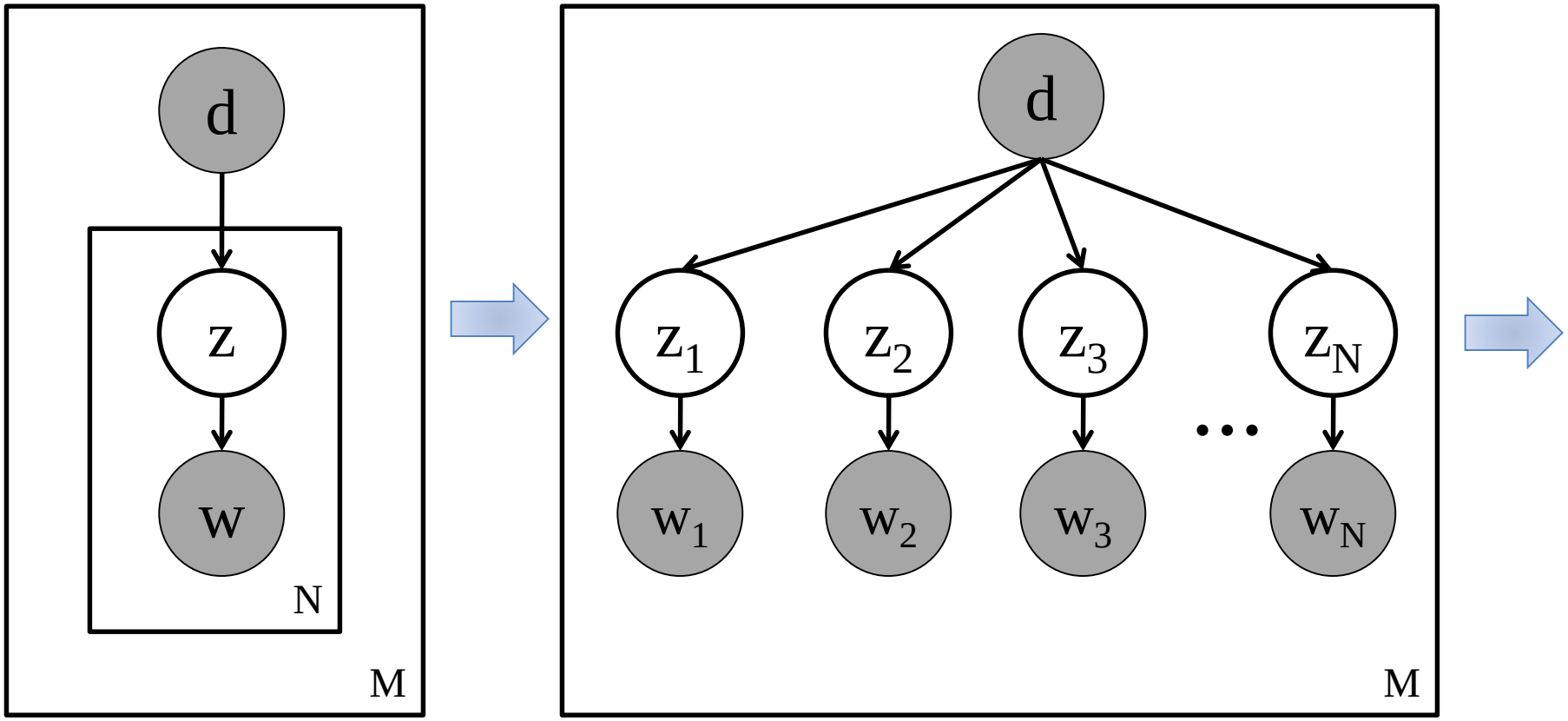
主题模型

- pLSA和LDA都是主题模型
  - In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents.

基本思想：（1）文档是若干主题的混合分布；（2）每个主题又是一个关于单词的概率分布



pLSA (probabilistic Latent Semantic Analysis)

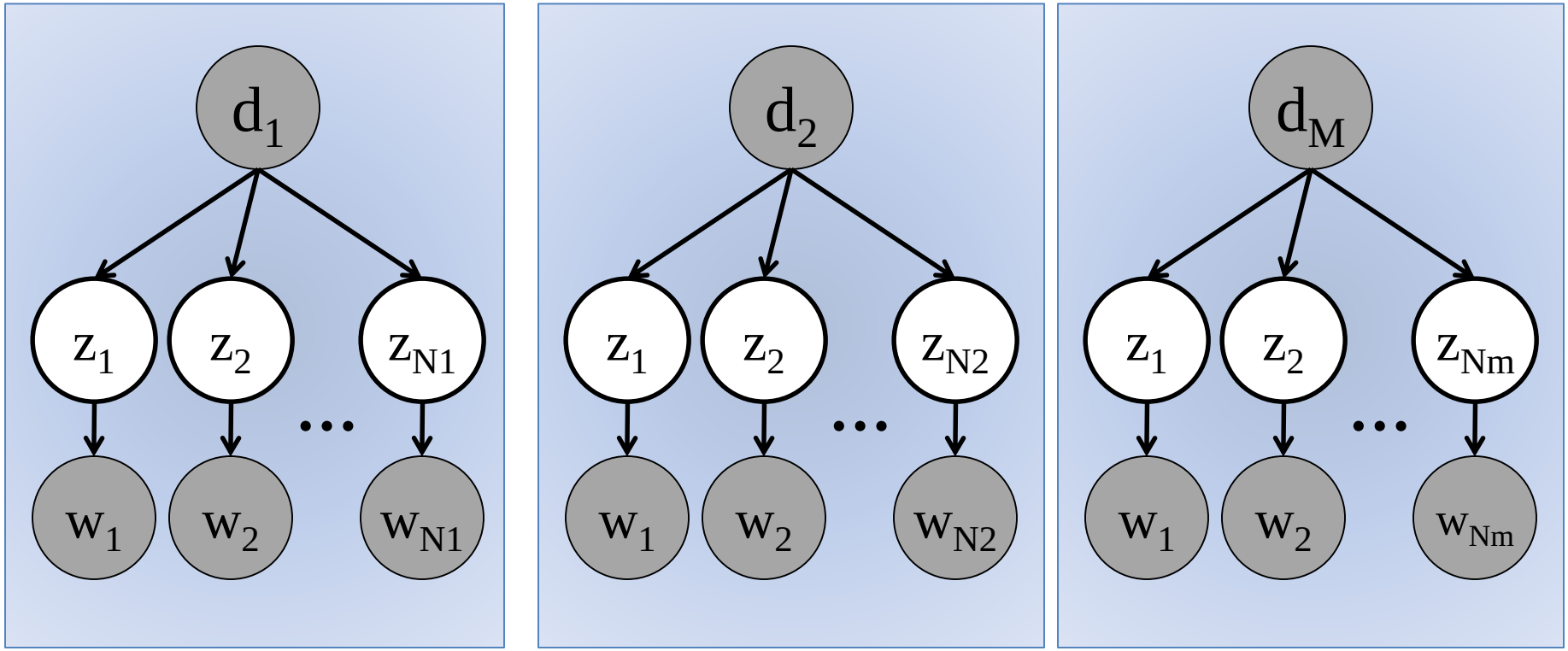


M：文档数目
N：文档d中的词项数目

$z_1, \ldots, z_N$ are variables. $z_i \epsilon [1,K]$.
K is the number of latent topics.

pLSA
probabilistic Latent Semantic Analysis

n(d,w)表示文档d中w出现的次数



p(w|z=1), p(w|z=2), p(w | z=N_M) are shared for all documents.

Likelihood: $\mathcal{L} = \prod_d \prod_w \left(\sum_z p(z \mid d) p(w \mid z) \right)^{n(d,w)}$

n(d,w)表示文档d中w出现的次数

Joint Probability vs Likelihood

- Joint probability

$$p(d,z,w) = \prod_d p(d) \prod_w \left(p(w \mid z) p(z \mid d) \right)^{n(d,w)}$$

- Likelihood (only for observed variables)

$$p(d,w) = \prod_d p(d) \prod_w \left(\sum_z p(w \mid z) p(z \mid d) \right)^{n(d,w)}$$

p(d) is assumed to be uniform

$$\mathcal{L} = p(w|d) = \prod_d \prod_w \left(\sum_z p(w \mid z) p(z \mid d) \right)^{n(d,w)}$$

pLSA – Objective Function

- pLSA tries to maximize the log likelihood:

$$\max_{p(z|d),p(w|z)} \quad \sum_d \sum_w n(d,w) \log \left(\sum_z p(z|d)p(w|z) \right)$$

$$s.t. \quad p(z|d) \geq 0, p(w|z) \geq 0, \quad \text{for any } w, z \text{ and } d$$

$$\sum_w p(w|z) = 1, \quad \text{for any } z$$

$$\sum_z p(z|d) = 1, \quad \text{for any } d$$

- Due to the summation over z inside log, we have to resort to EM.

Expectation–Maximization (EM) algorithm

- The EM algorithm is a method for ML learning of parameters in latent variable models.
- E-Step
  - 根据已经估计的参数计算隐藏变量的后验概率（即根据参数计算似然函数的期望）
- M-Step
  - 根据已经计算的后验概率更新参数（选择参数使似然最大化）

pLSA – EM Steps

$$\mathcal{L} = p(w|d) = \prod_d \prod_w \left(\sum_z p(w \mid z) p(z \mid d) \right)^{n(d,w)}$$

- The E-Step: 根据参数计算似然函数的期望

$$p(z|d,w) = \frac{p(w|z)p(z|d)p(d)}{p(d,w)} \propto p(w|z)p(z|d)$$

- The M-Step: 选择参数使似然最大化

$$p(z|d) \propto \sum_w n(d,w)q(z|d,w)$$

$$p(w|z) \propto \sum_d n(d,w)q(z|d,w)$$

pLSA vs LSA

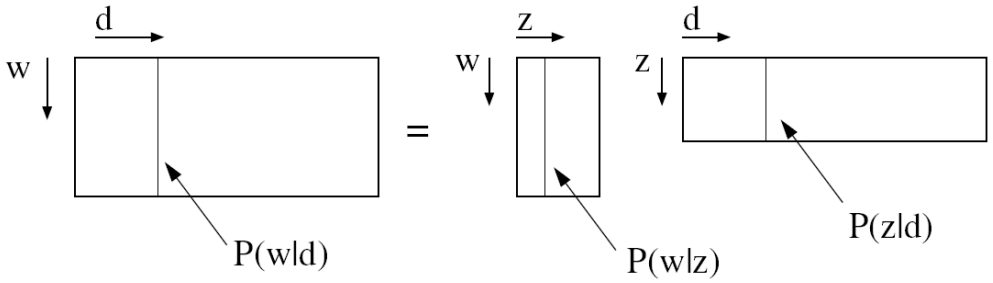
Each document can be decomposed as:

$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \quad j=1,2,\ldots, N_{d_i}$$

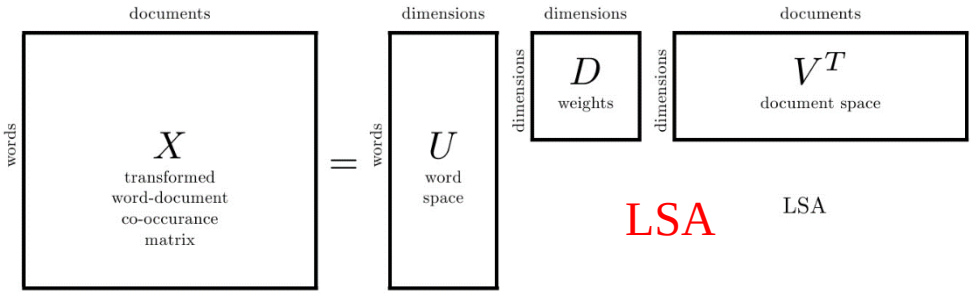
This is similar to the matrix decomposition.

$$p(w|d) = Z_{V\times k}\, p(z|d)$$

$$z^* = \arg\max_z p(z \mid d)$$



pLSA



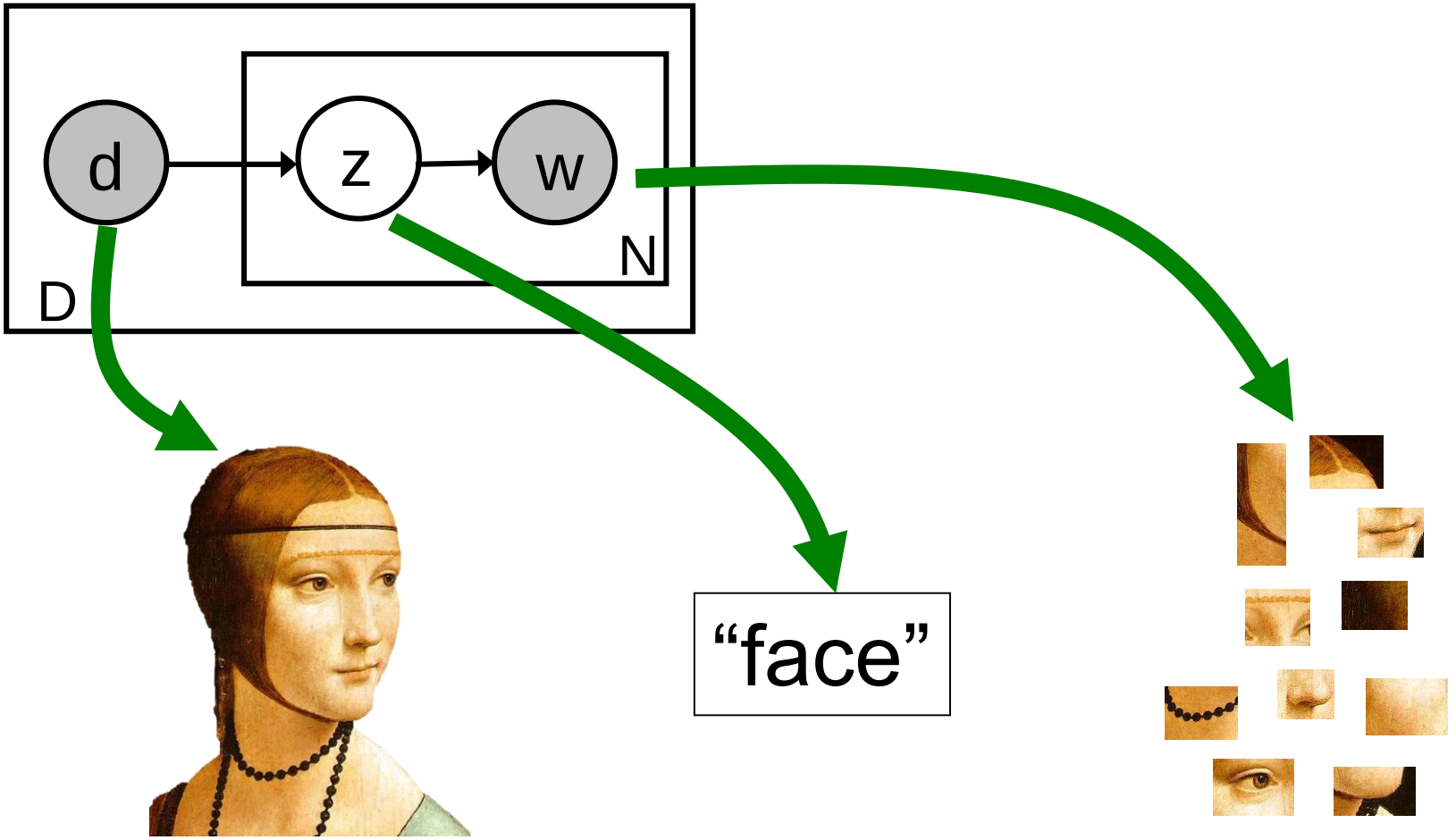
LSA

LSA

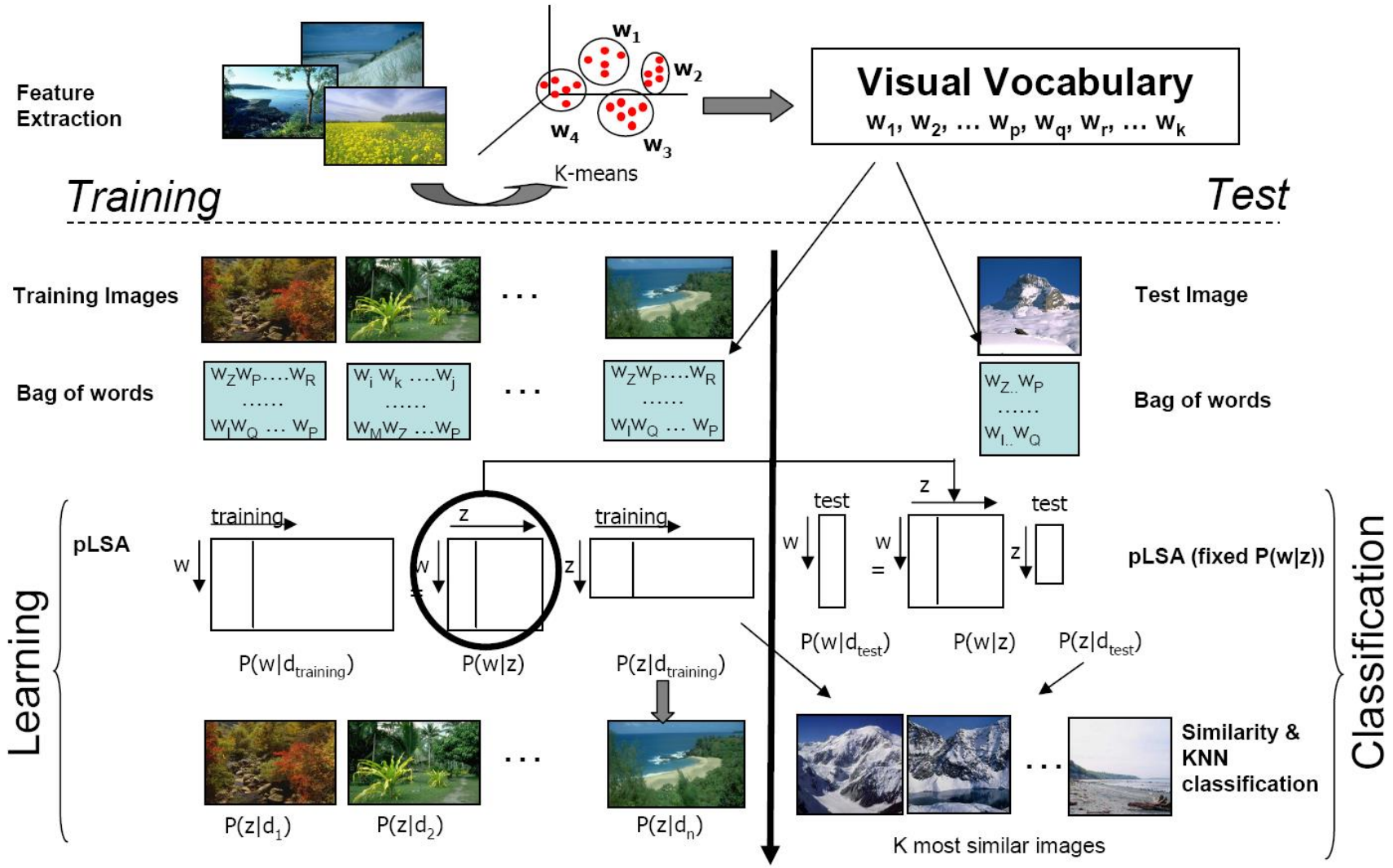
pLSA vs LSA

- LSA and pLSA perform dimensionality reduction
  - In LSA, by keeping only K singular values
  - In pLSA, by having K aspects
- The main difference is the way the approximation is done
- pLSA generates a model (aspect model) and maximizes its predictive power
- Selecting the proper value of K is heuristic in LSA
- Model selection in statistics can determine optimal K in pLSA

pLSA用于图像分类



pLSA应用：Scene Classification



Bosch, A. , Zisserman, A. and Munoz, X.
Scene Classification via pLSA， ECCV 2006, 2016.04 google cited: 742

Classification Result

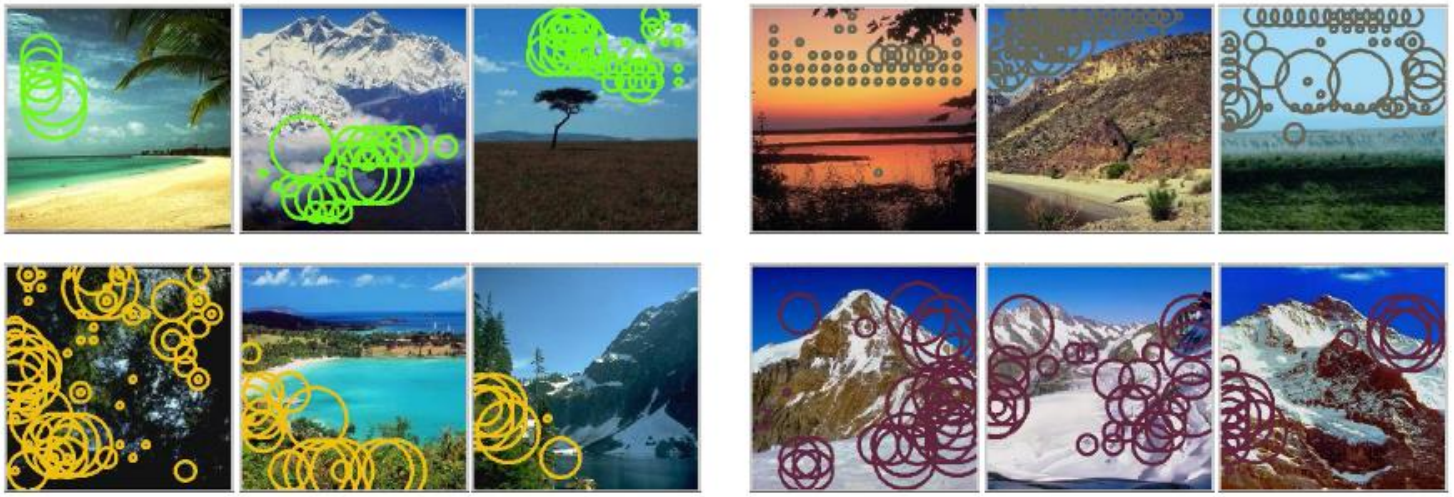


Fig. 5. Topics segmentation. Four topics (clouds – top left, sky – top right, vegetation – lower left, and snow/rocks in mountains – lower right) are shown. Only circular regions with a topic posterior $P(z|w,d)$ greater than 0.8 are shown.

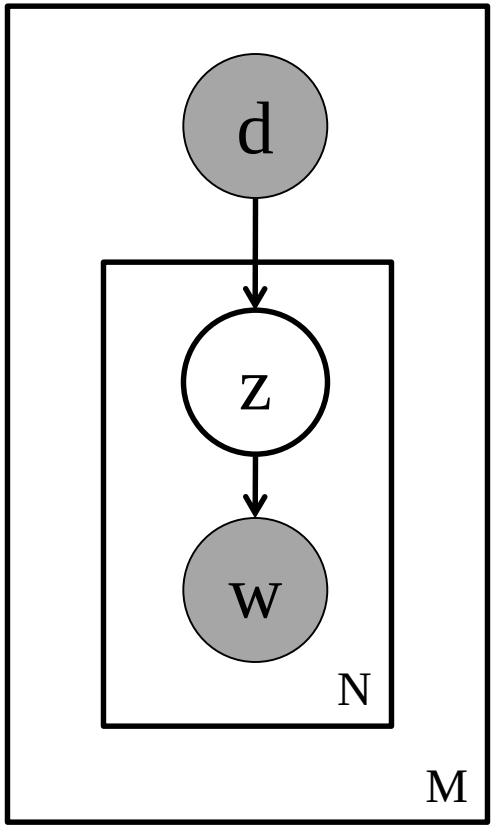
# img. (nt)	2000	1600	1024	512	256	128	32
Perf. $P(z\|d)$	86.9	86.7	84.6	79.5	75.3	68.2	58.7
Perf. BOW	83.1	82.6	80.4	72.8	60.2	52.0	47.3

Table 2. Comparison of $P(z|d)$ and BOW performance as the number of training images used in KNN is decreased. The classification task is into 8 categories from the OT dataset.

Bosch, A. , Zisserman, A. and Munoz, X.
Scene Classification via pLSA， ECCV 2006, 2016.04 google cited: 742

小结：pLSA

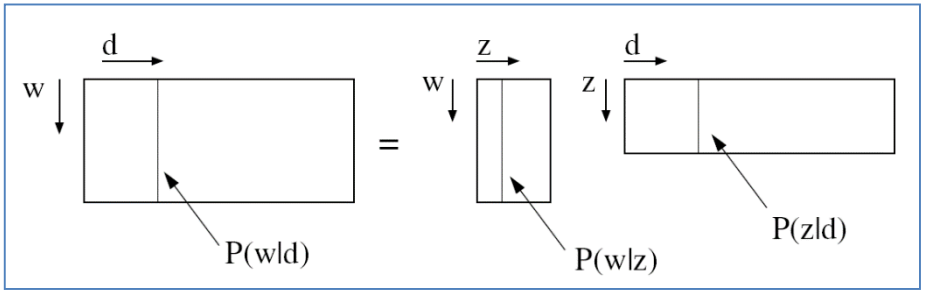
- topic model：（1）文档是若干主题的混合分布；（2）每个主题又是一个关于单词的概率分布



$$p(d,z,w) = \prod_d p(d) \prod_w \left(p(w \mid z) p(z \mid d) \right)^{n(d,w)}$$

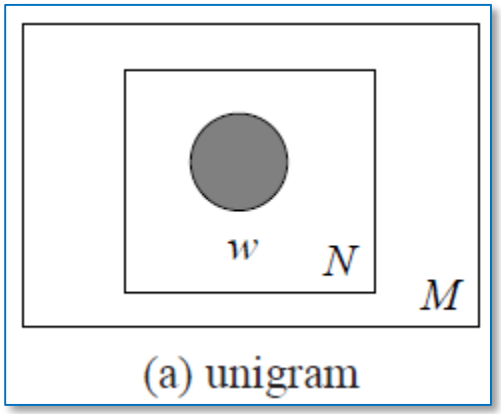
$$\mathcal{L} = p(w|d) = \prod_d \prod_w \left(\sum_z p(w \mid z) p(z \mid d) \right)^{n(d,w)}$$

pLSA



小结：如何生成M份包含N个单词的文档

3种文档生成模型： (a)unigram (b)mixture of unigrams (c) pLSA



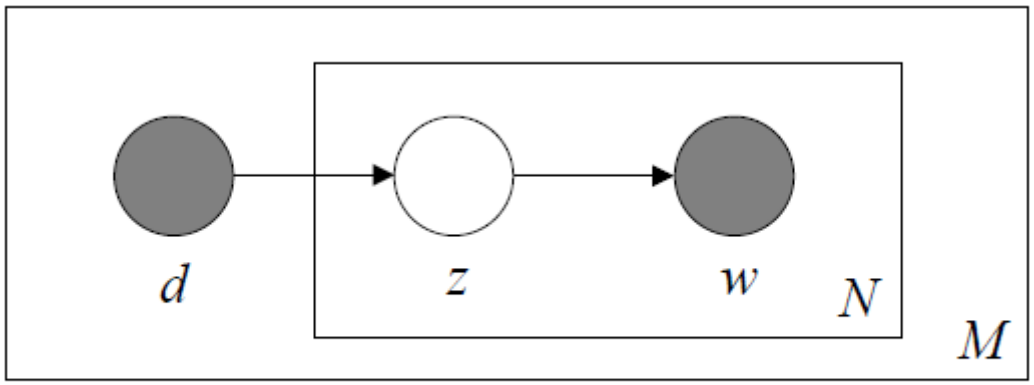
(a) unigram

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$

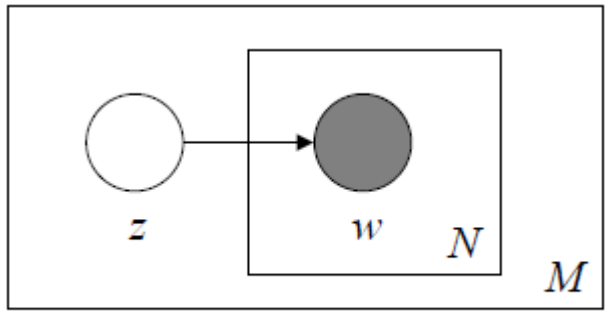
没有主题

一个文档只有一个主题

$$p(w) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n|z)$$



(c) pLSI/aspect model



(b) mixture of unigrams

$$p(d, w_n) = p(d) \sum_{z} p(w_n|z) p(z|d)$$

文档可以包含多个主题

概率图及主题模型
Probabilistic Graphical Models / Topic Model

- 什么是Graphical Model
  - 定义、示例
  - Representation、Inference、Learning
- 主题模型与分类
  - LSA (Latent Semantic Analysis), 1990
  - pLSA (probabilistic Latent Semantic Analysis), 1999
  - LDA(Latent Dirichlet Allocation), 2003
  - Hierarchical Bayesian model
- 主题模型的R语言实现示例

Latent dirichlet allocation
David M. Blei, Andrew Y. Ng, Michael I. Jordan
Journal of Machine Learning Research, 2003
2016.04 google cited: 14167

吴恩达（1976-，英文名：Andrew Ng），华裔美国人

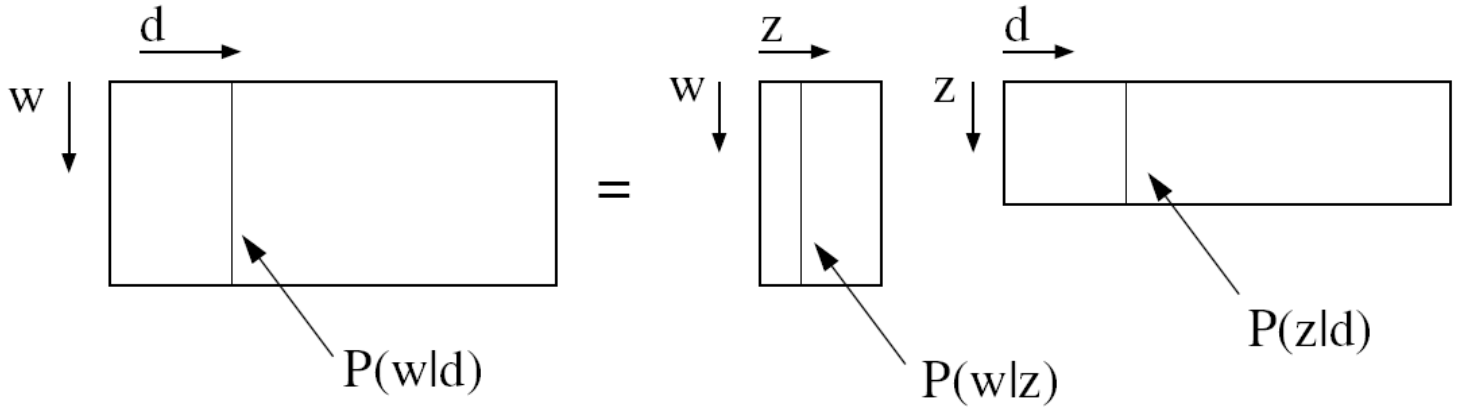
1976年生于英国，之后在香港和新加坡；
1992年毕业于新加坡莱佛士书院；
1997年获得卡内基梅隆大学计算机科学学士学位；
1998年获得麻省理工硕士学位；
2002年在加州大学伯克利分校获得博士学位；
2002年9月-斯坦福大学计算机科学系和电气工程系的副教授，斯坦福人工智能实验室的主任；
2011年1月–2012年6月，创办并领导Google深度学习项目；
2012年1月-今，Coursera联合创始人。
2014年5月16日，吴恩达加入百度，担任百度公司首席科学家，负责百度研究院的领导工作，尤其是Baidu Brain计划。

2017年03月22日，吴恩达在社交平台发布公开信，宣布自己将从百度离职，开启自己在人工智能领域的新篇章。

Latent dirichlet allocation
David M. Blei, Andrew Y. Ng, Michael I. Jordan
Journal of Machine Learning Research, 2003
2016.04 google cited: 14167

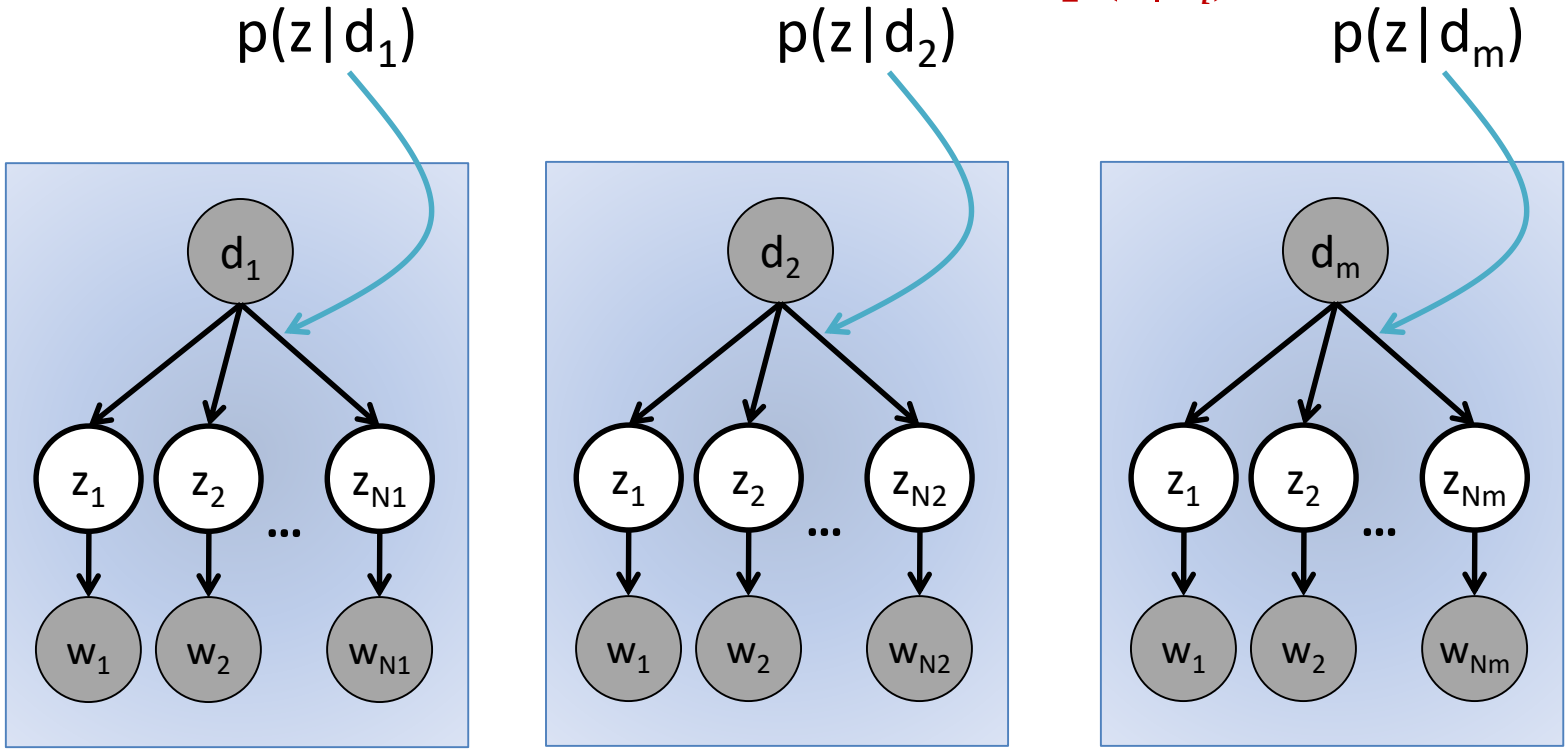
Problems in pLSA

- pLSA provides no probabilistic model at the document level. Each doc has its own topic mixture proportion.
- The number of parameters in the model grows linearly with M (the number of documents in the training set).



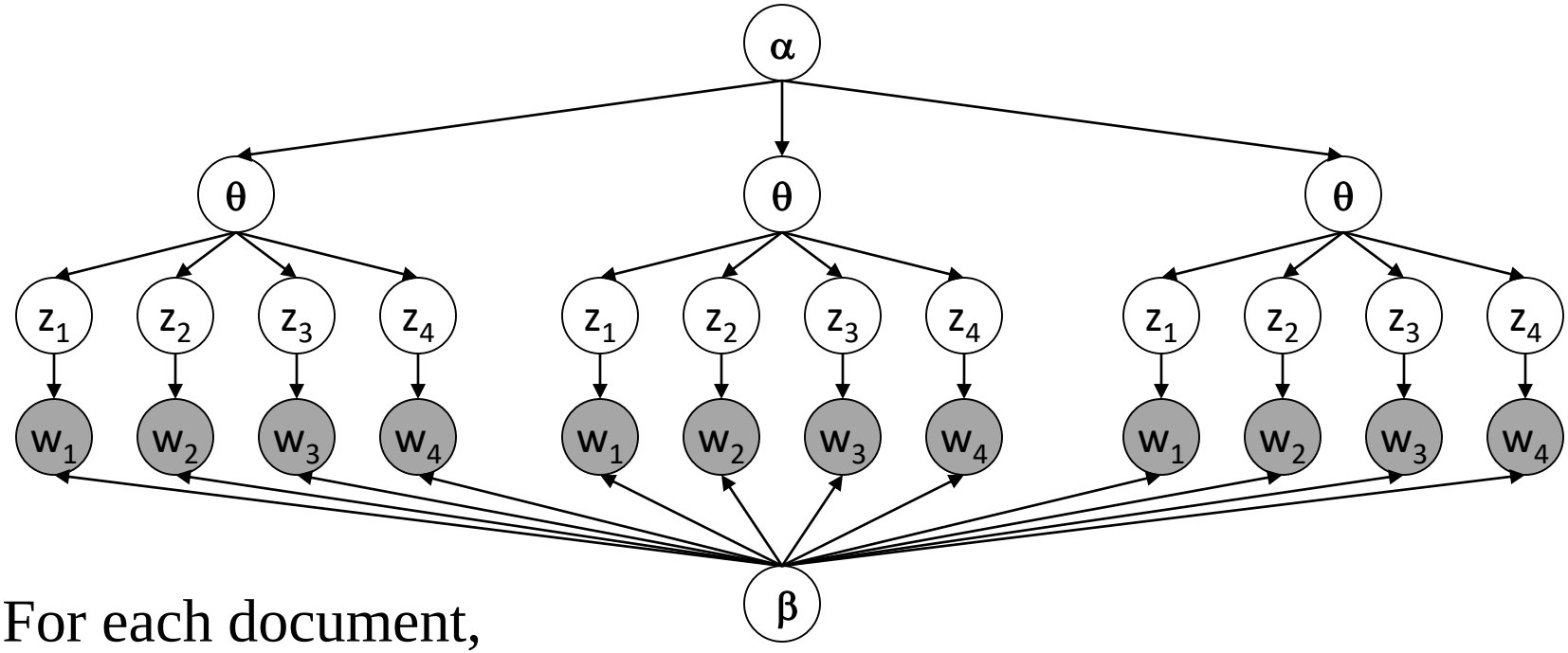
Problems in pLSA

- There is no constraint for distributions $p(z|d_i)$.



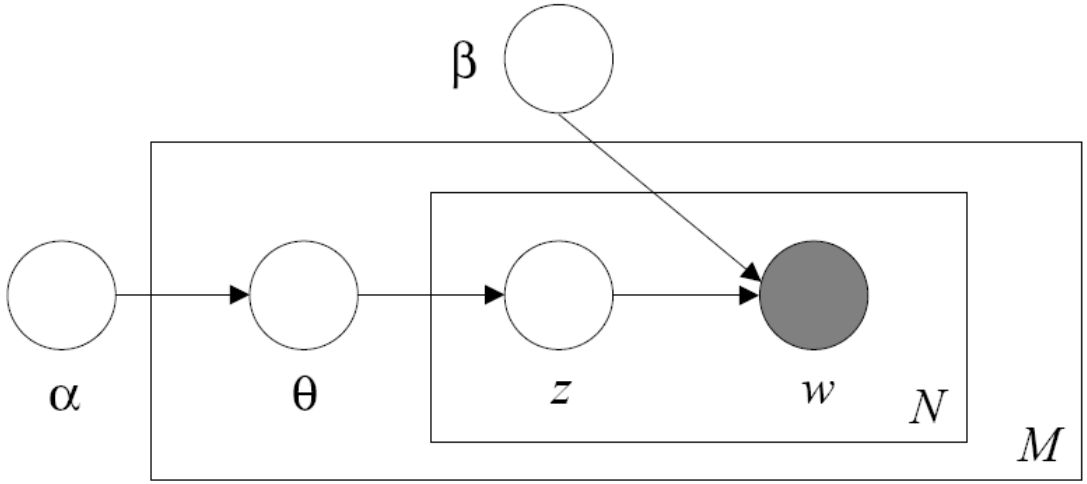
- Easy to lead to serious problems with over-fitting.

The LDA Model



- For each document,
- Choose θ~**Dirichlet(α)**
- For each of the N words w_n:
  - Choose a topic z_n~ **Multinomial(θ)**
  - Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

The LDA Model
文档是如何生成的?

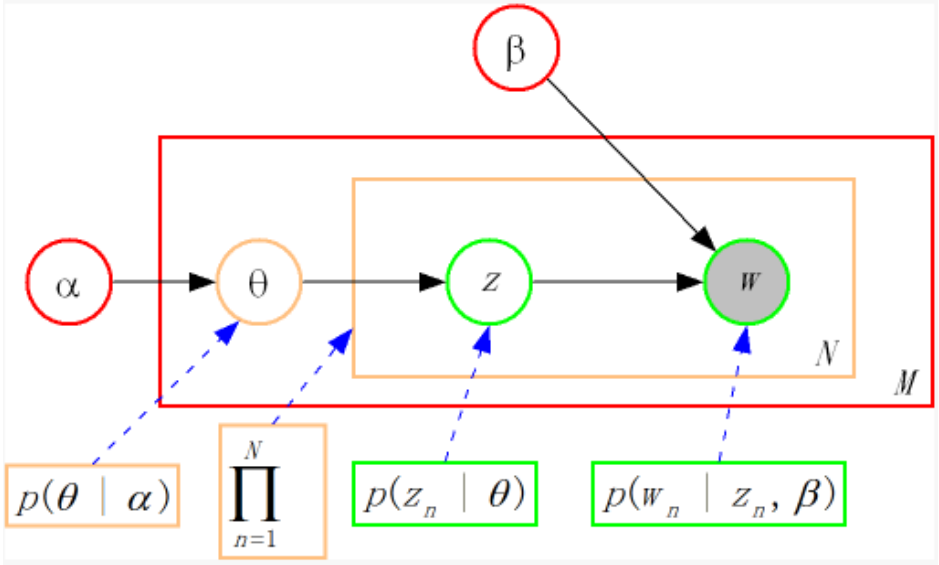


$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

- For each document,
- Choose $\theta \sim p(\theta)$, Dirichlet(α)
- For each of the N words w_n:
  - Choose a topic $z_n \sim p(z|\theta)$, Multinomial(θ)
  - Choose a word $w_n \sim p(w|z)$, from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

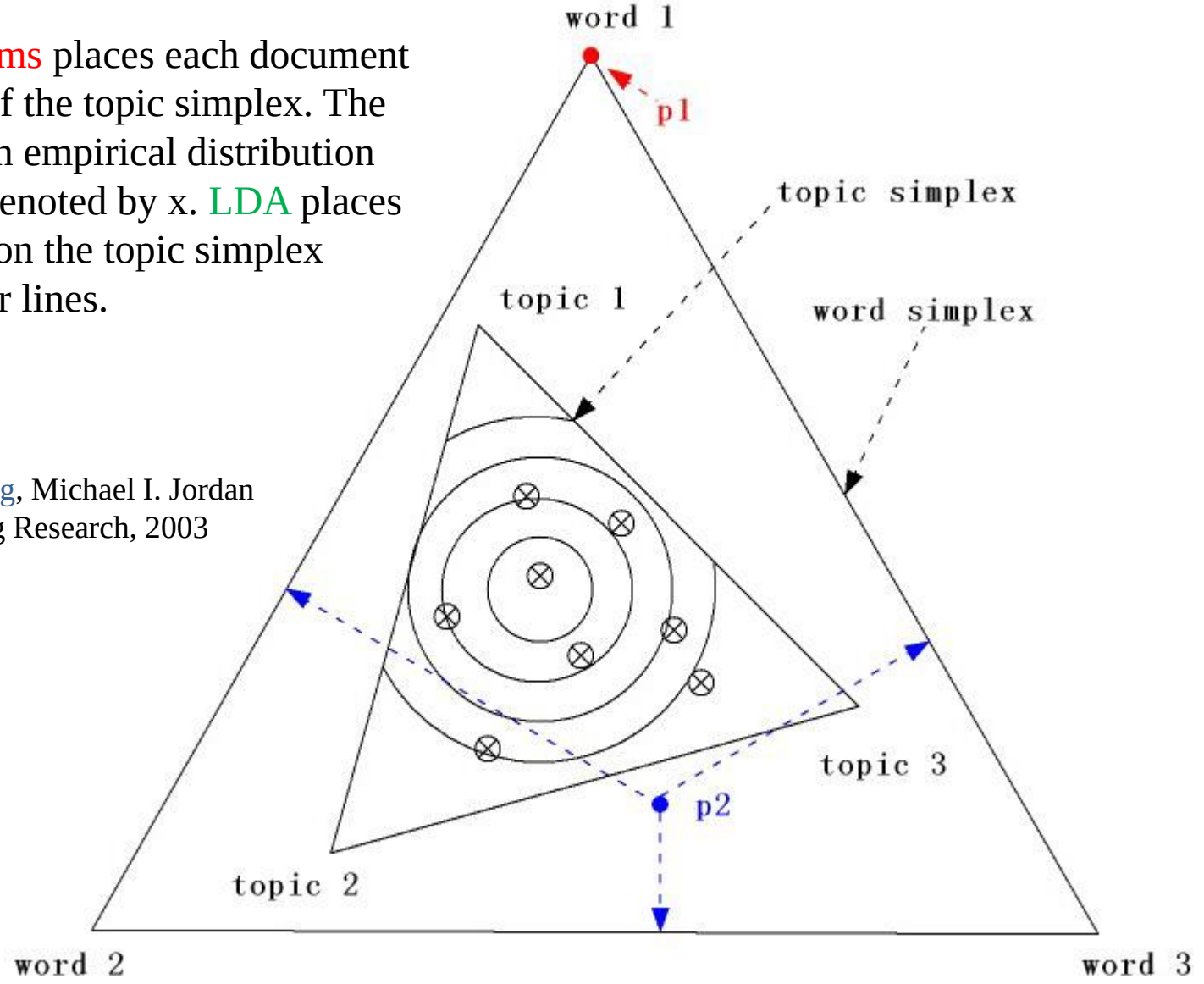
LDA的文档生成模型

1. corpus-level (红色)：α和β是语料级别的参数，对于每个文档都是一样的，因此在generate过程中只需要sample一次。
2.document-level (橙色)：θ是文档级别的参数，意即每个文档的θ参数是不一样的，也就是说每个文档产生topic z的概率是不同的，所以对于每个文档都要sample一次θ。
3. word-level (绿色)：最后z和w都是文档级别的变量，z由参数θ产生，之后再由z和β共同产生w，一个w对应一个z。



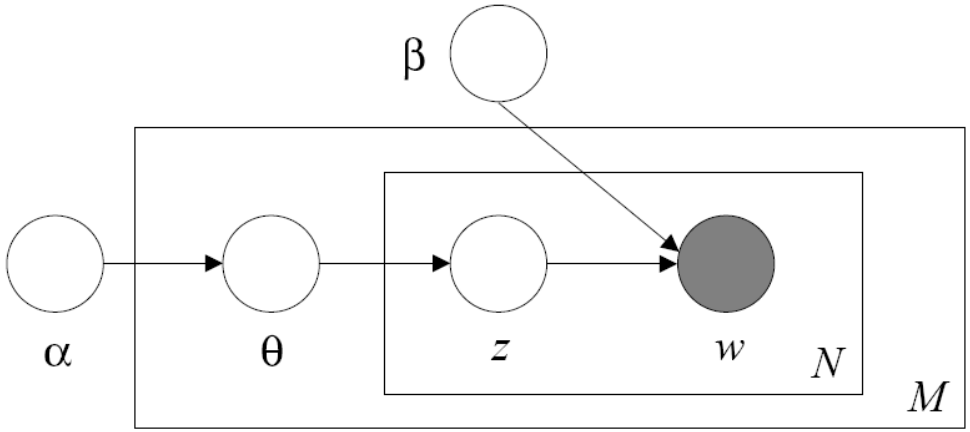
几何学解释

The mixture of unigrams places each document at one of the corners of the topic simplex. The pLSI model induces an empirical distribution on the topic simplex denoted by x. LDA places a smooth distribution on the topic simplex denoted by the contour lines.



Latent dirichlet allocation
David M. Blei, Andrew Y. Ng, Michael I. Jordan
Journal of Machine Learning Research, 2003
2016.04 google cited: 14167

Joint Probability



- Given parameter α and β

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

where

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

Likelihood

- Joint Probability

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

- Marginal distribution of a document

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta$$

- Likelihood over all the documents

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d$$

CVPR2015的LDA分析

CVPR 2015 papers
(in nicer format than this)
maintained by @karpathy

NEW: This year I also embedded the (1,2-gram) tfidf vectors of all papers with t-sne and placed them in an interface where you can navigate them visually. I'm not sure if it's useful but it's really cool.

Below every paper are TOP 100 most-occuring words in that paper and their color is based on LDA topic model with k = 7.
(It looks like 0 = datasets?, 1 = deep learning, 2 = videos , 3 = 3D Computer Vision , 4 = optimization?, 5 = low-level Computer Vision?, 6 = descriptors?)

Toggle LDA topics to sort by: TOPIC0 TOPIC1 TOPIC2 TOPIC3 TOPIC4 TOPIC5 TOPIC6

Going Deeper With Convolutions
Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, Rabinovich, Andrew

[pdf] [rank by tf-idf similarity to this]

[learning, work, vision, visual] [conv, inception, network, deep, convolutional, layer, architecture, neural, googlenet, maxpool, increase, ilsvrc, imagenet, size, performance, table, pooling, output, depthconcat, larger, suggests, challenge, training, higher, increasing, ensemble, max, top, pool, trained, accuracy, connected, improved, compared, previous, increased, stage, expensive, dropout, auxiliary, validation, design, highly] [detection, object, bounding, box, based, average] [model, approach, current, single, localization, error, inference, structure, depth, allows] [number, sparse, computational, data, linear, order, main, reduction, optimal, matrix, problem, construction] [computer, figure, quality, result] [image, large, well, dense, external, rate, region]

Propagated Image Filtering
Rick Chang, Jen-Hao, Frank Wang, Yu-Chiang

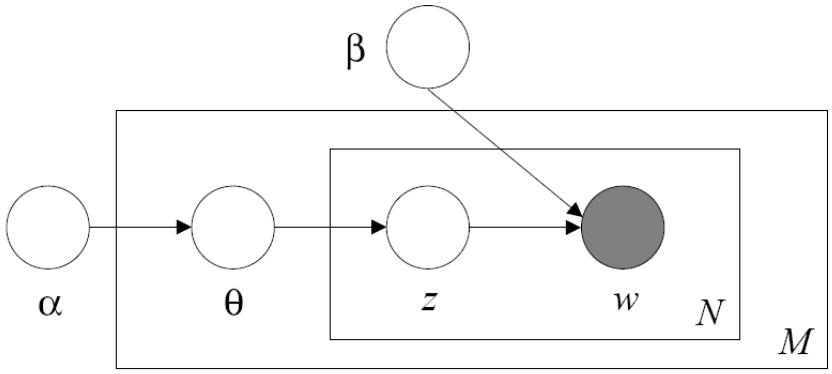
[pdf] [rank by tf-idf similarity to this]

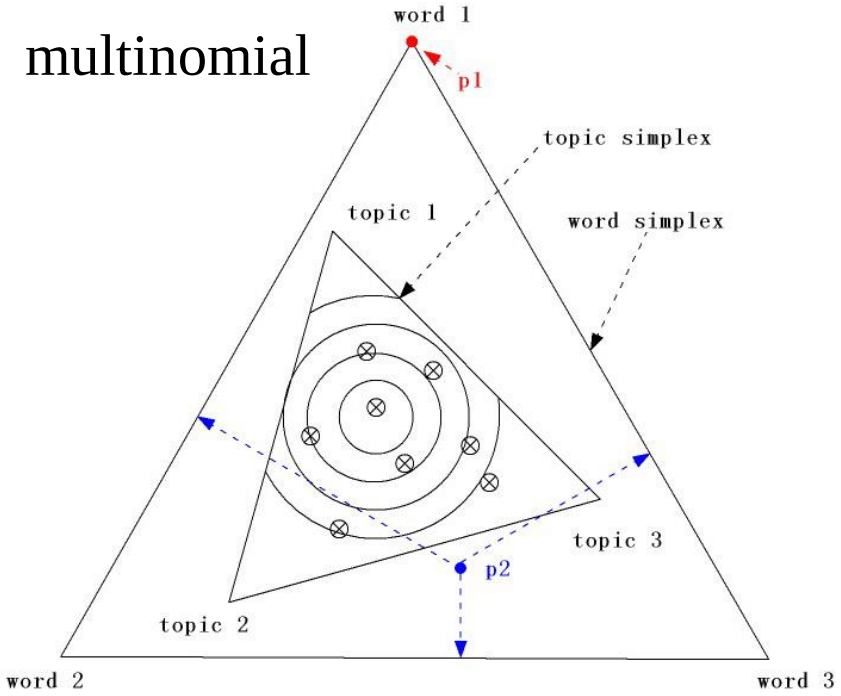
http://www-cs-faculty.stanford.edu/people/karpathy/cvpr2015papers/

小结：LDA

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

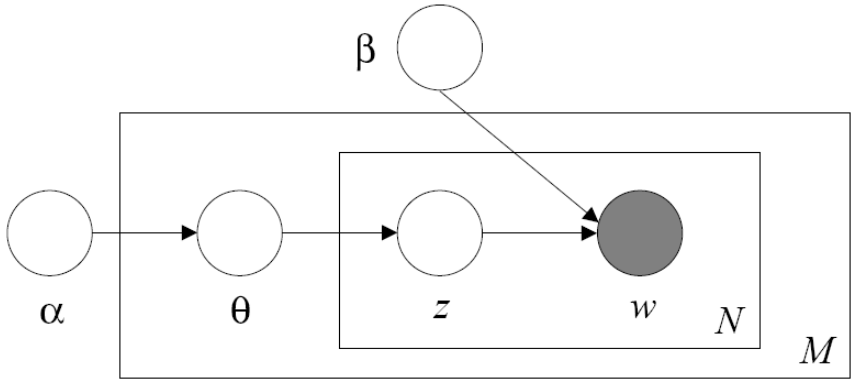
- *latent Dirichlet allocation* (LDA)
  - A generative probabilistic
- LDA is a three-level hierarchical Bayesian model.
- $\theta \sim p(\theta)$, Dirichlet(α)
- topic $z_n \sim p(z|\theta)$, Multinomial(θ)
- word $w_n \sim p(w|z)$, from $p(w_n|z_n,\beta)$, multinomial





LDA的局限性

- “ bag of words” 的假设
- 主题的词项分布是不随时间变化的
- 主题的数目是已知并固定的
- 忽略主题之间的相关性



- For each document,
- Choose $\theta \sim p(\theta)$, **Dirichlet**(α)
- For each of the N words w_n:
  - Choose a topic $z_n \sim p(z|\theta)$, **Multinomial**(θ)
  - Choose a word $w_n \sim p(w|z)$, from $p(w_n|z_n,\beta)$, a **multinomial** probability conditioned on the topic z_n.

LDA改进

David Blei

关注

Professor of Statistics and Computer Science, Columbia University
Machine Learning, Statistics, Probabilistic topic models, Bayesian nonparametrics, Approximate posterior inference
在 columbia.edu 的电子邮件经过验证 - 首页

http://www.cs.columbia.edu/~blei/

标题 1–20	引用次数	发表年份
Latent dirichlet allocation DM Blei, AY Ng, MI Jordan Journal of machine Learning research 3 (Jan), 993-1022	18062	2003
Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. YW Teh, MI Jordan, MJ Beal, DM Blei NIPS, 1385-1392	2847	2004
Supervised topic models JD Mcauliffe, DM Blei Advances in neural information processing systems, 121-128	1722	2008
Probabilistic topic models DM Blei Communications of the ACM 55 (4), 77-84	1713	2012
Matching words and pictures K Barnard, P Duygulu, D Forsyth, N Freitas, DM Blei, MI Jordan Journal of machine learning research 3 (Feb), 1107-1135	1680	2003
Dynamic topic models DM Blei, JD Lafferty Proceedings of the 23rd international conference on Machine learning, 113-120	1412	2006
Correlated topic models D Blei, J Lafferty Advances in neural information processing systems 18, 147	1355	2006

←语料库中的主题随时间变化

←考虑主题间的相关性，Dirichlet → a log-normal

https://scholar.google.com.hk/citations?user=8OYE6iEAAAAJ&hl=zh-CN Retrieved:2017-04-06

概率图及主题模型
Probabilistic Graphical Models / Topic Model

- 什么是Graphical Model
  - 定义、示例
  - Representation、Inference、Learning
- 主题模型与分类
  - LSA (Latent Semantic Analysis), 1990
  - pLSA (probabilistic Latent Semantic Analysis), 1999
  - LDA(Latent Dirichlet Allocation), 2003
  - Hierarchical Bayesian model
- 主题模型的R语言实现示例

A bayesian hierarchical model for learning natural scene categories
Li Fei-Fei, Pietro Perona, CVPR 2005, 2016.04 google cited: 2942

Li Fei-Fei

Professor of Computer Science, Stanford University
Artificial Intelligence, Machine Learning, Computer Vision,
在 cs.stanford.edu 的电子邮件经过验证 - 首页

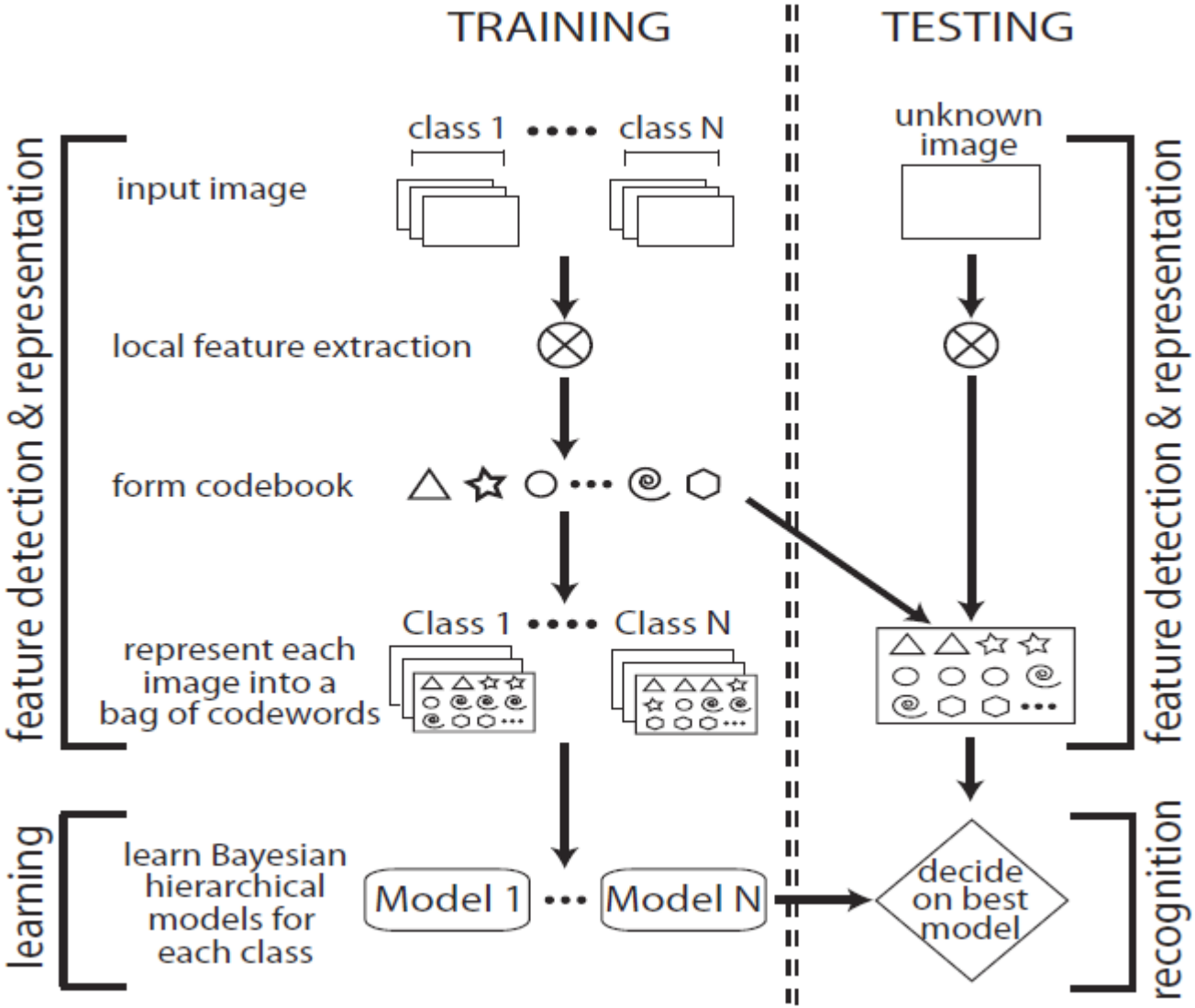
https://scholar.google.com/citations?user=rDfyQnIAAAAJ&hl=zh-CN retrieved: 20170407

2016年11月，谷歌宣布李飞飞加入其云团队

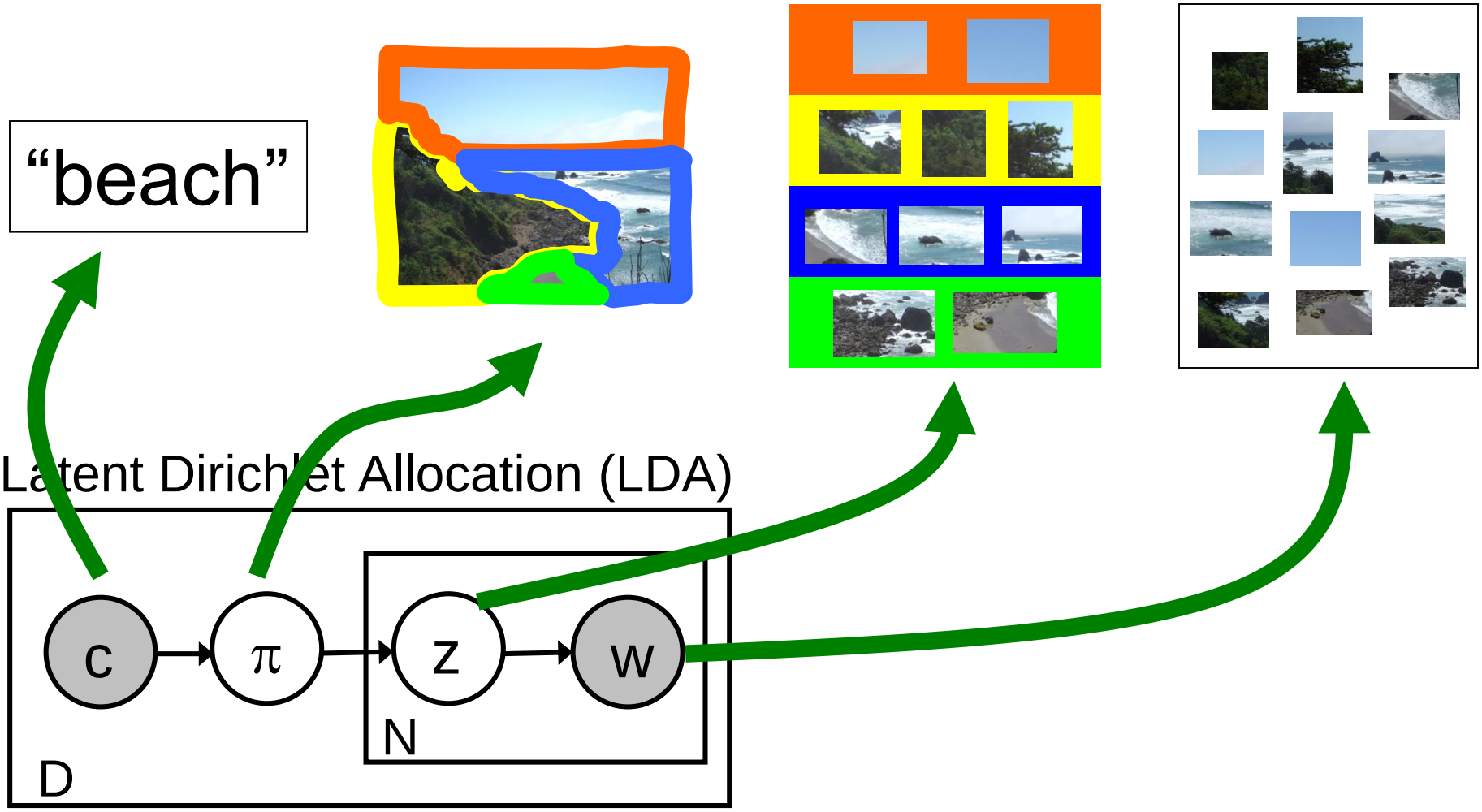
标题 1–20		
Imagenet: A large-scale hierarchical image database J Deng, W Dong, R Socher, LJ Li, K Li, L Fei-Fei Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on ...	3793	2009
A bayesian hierarchical model for learning natural scene categories L Fei-Fei, P Perona Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer ...	3361	2005
Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories L Fei-Fei, R Fergus, P Perona Computer vision and Image understanding 106 (1), 59-70	[illegible]	[illegible]
Imagenet large scale visual recognition challenge O Russakovsky, J Deng, H Su, J Krause, S Satheesh, S Ma, Z Huang, ... International Journal of Computer Vision 115 (3), 211-252	[illegible]	2015
Unsupervised learning of human action categories using spatial-temporal words JC Niebles, H Wang, L Fei-Fei International Journal of Computer Vision 79 (3), 299-318	1649	[illegible]
Large-scale video classification with convolutional neural networks A Karpathy, G Toderici, S Shetty, T Leung, R Sukthankar, L Fei-Fei Proceedings of the IEEE conference on Computer Vision and Pattern ...	929	[illegible]

生于北京，长在四川，16岁随父母移居美国。现为斯坦福大学计算机系终身教授，人工智能实验室与视觉实验室主任。李飞飞教授主要研究方向为机器学习、计算机视觉、认知计算神经学，侧重大数据分析为主。1999年获普林斯顿大学本科学位，2005年获加州理工学院电子工程博士学位。2009年她加入斯坦福大学任助理教授，并于2012年担任副教授（终生教授），此前分别就职于普林斯顿大学（2007-2009）、伊利诺伊大学香槟分校（2005-2006）。李飞飞教授为TED 2015大会演讲嘉宾……

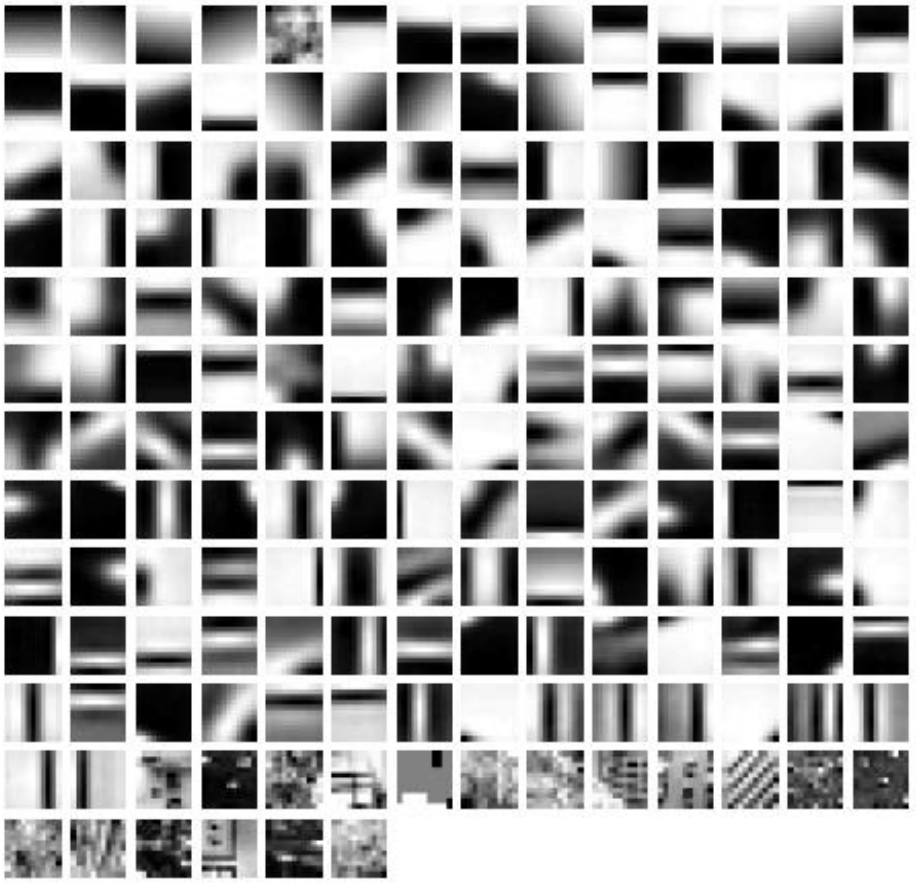
从文本分类→图像分类


TRAINING
TESTING
class 1
class N
unknown image
input image
local feature extraction
form codebook
Class 1
Class N
represent each image into a bag of codewords
learn Bayesian hierarchical models for each class
Model 1
Model N
decide on best model
feature detection & representation
feature detection & representation
learning
recognition

Hierarchical Bayesian text models



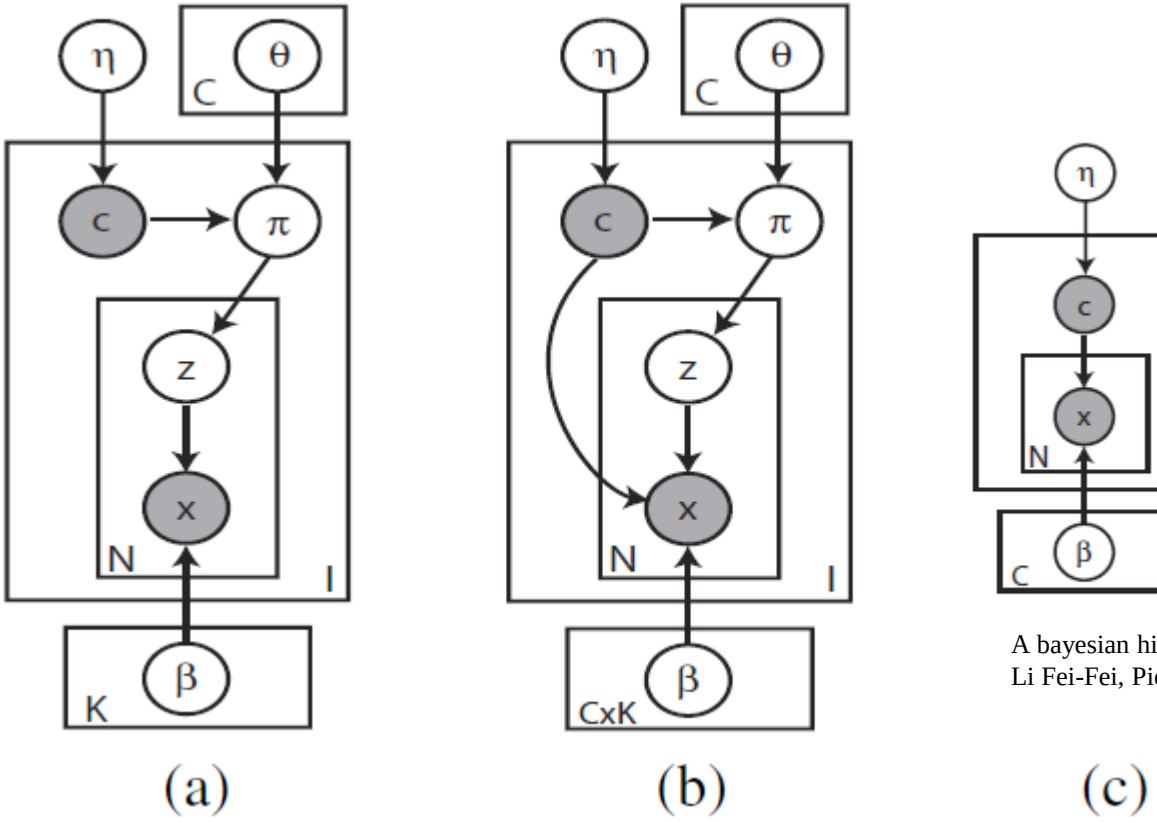
Codebook



A codebook obtained from 650 training examples from all 13 categories (50 images from each category). Image patches are detected by a sliding grid and random sampling of scales. The codewords are sorted in descending order according to the size of its membership. Interestingly most of the codewords appear to represent simple orientations and illumination patterns, similar to the ones that the early human visual system responds to.

Theme Model for scene categorization

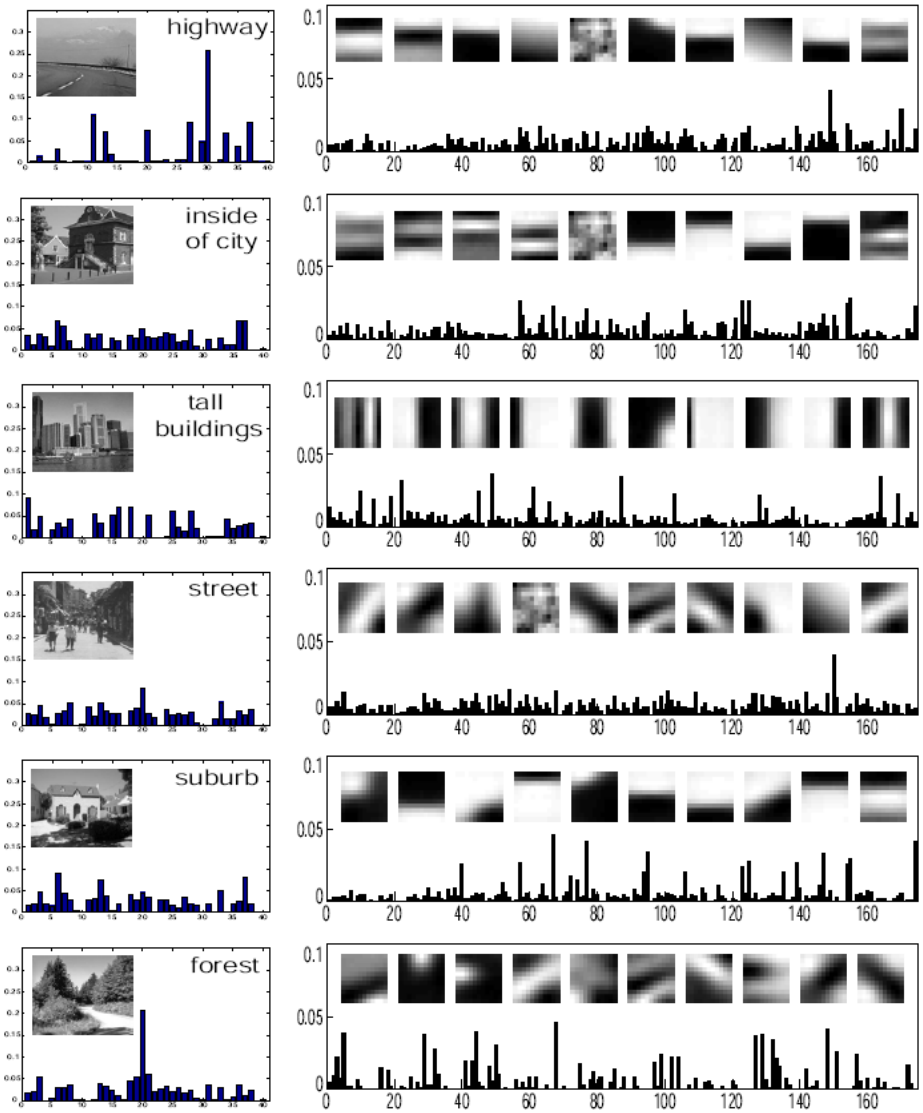
$c \sim p(c|\eta)$, multinomial(η)
$\pi \sim p(\pi|c, \theta)$, Dir ($\theta$)
$z_n \sim$ multinomial (π)
$x_n \sim p(x_n|z_n, \beta)$,



A bayesian hierarchical model for learning natural scene categories
Li Fei-Fei, Pietro Perona, CVPR 2005, 2016.04 google cited: 1942

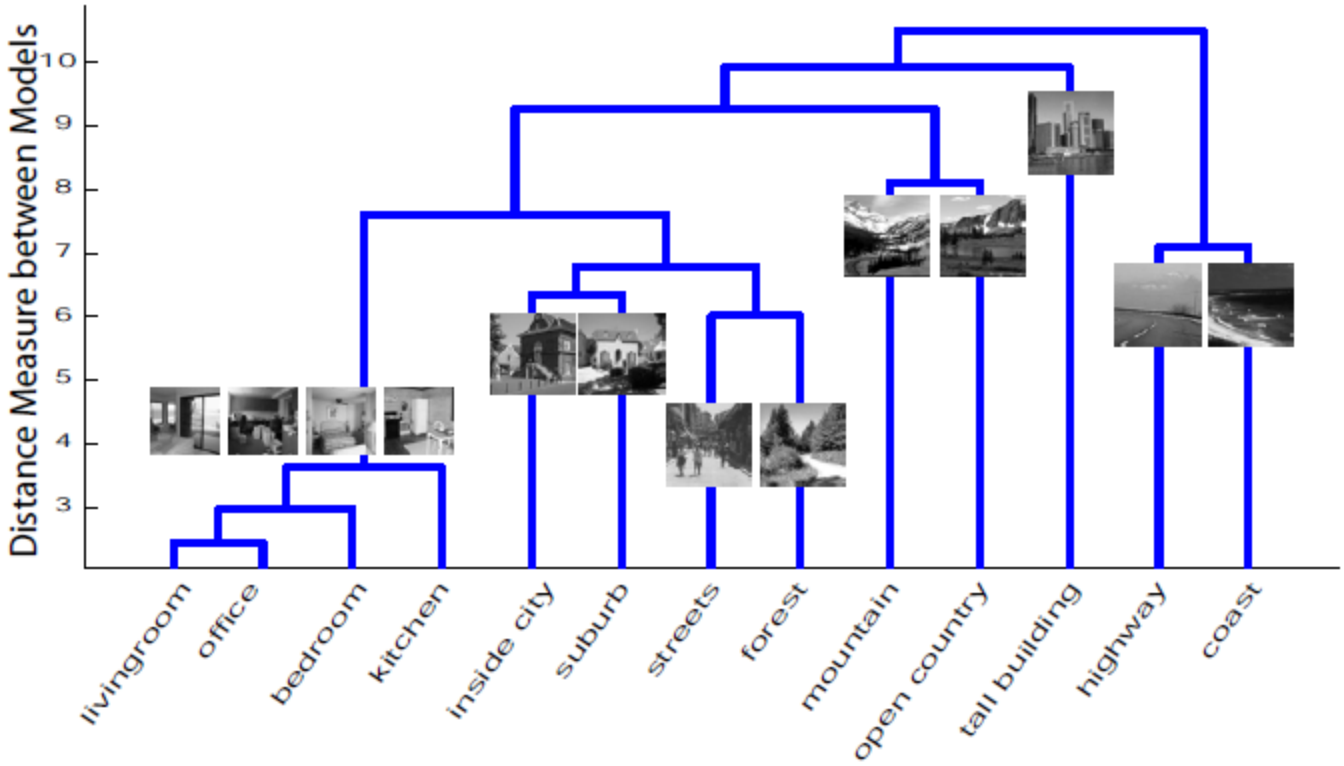
(a) Theme Model 1 for scene categorization that shares both the intermediate level themes as well as feature level codewords. **(b)** ThemeModel 2 for scene categorization that shares only the feature level codewords; **(c)** Traditional texton model

Topic Distribution in Different Categories



Internal structure of the models learnt for each category. Each row represents one category. The left panel shows the distribution of the 40 intermediate themes. The right panel shows the distribution of codewords as well as the appearance of 10 codewords selected from the top 20 most likely codewords for this category model.

Topic Hierarchical Clustering



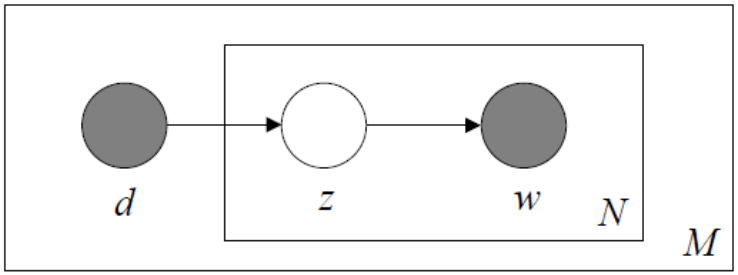
Dendrogram of the relationship of the 13 category models based on theme distribution. y-axis is the pseudo-Euclidean distance measure between models.

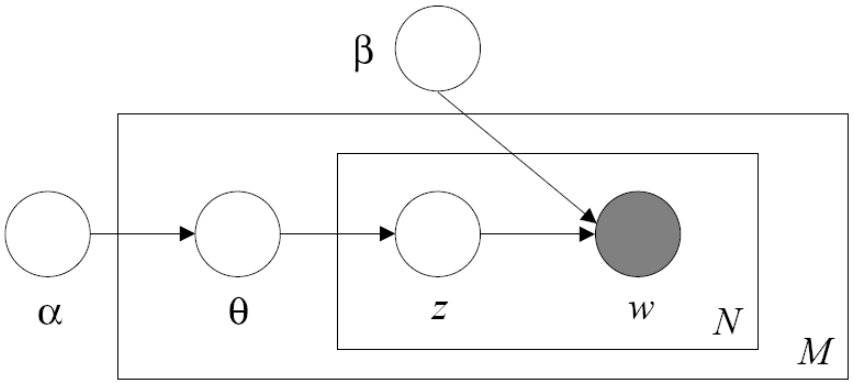
More Topic Models

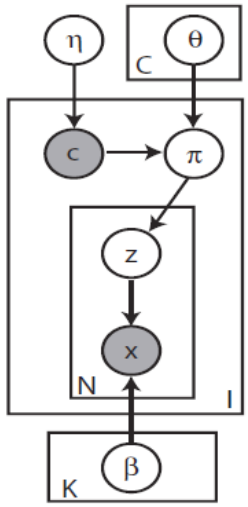
- Hierarchical Dirichlet Process, Journal of the American Statistical Association 2003
- Correlated Topic Model, NIPS 2005
- Dynamic topic models, ICML 2006
- Nonparametric Bayes pachinko allocation, UAI 2007
- Supervised LDA, NIPS 2007
- MedLDA – Maximum Margin Discrimant LDA, ICML 2009
- Online learning for latent dirichlet allocation, NIPS 2010
- Hierarchically supervised latent Dirichlet allocation, NIPS 2011
- A spectral algorithm for latent dirichlet allocation, NIPS 2012
- …
- TopicRNN: Combine RNN and Topic Model, ICLR 2017,
- Autoencoding Variational Inference For Topic Models, ICLR 2017
- Neural Relational Topic Models for Scientific Article Analysis, CIKM 2018

小结：主题模型与分类

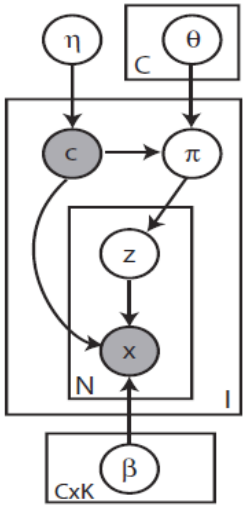
- LSA (Latent Semantic Analysis), 1990
- pLSA (probabilistic Latent Semantic Analysis), 1999
- LDA(Latent Dirichlet Allocation), 2003
- Hierarchical Bayesian model， 2009



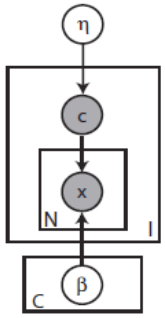




(a)



(b)



(c)

概率图及主题模型
Probabilistic Graphical Models / Topic Model

- 什么是Graphical Model
  - 定义、示例
  - Representation、Inference、Learning
- 主题模型与分类
  - LSA (Latent Semantic Analysis), 1990
  - pLSA (probabilistic Latent Semantic Analysis), 1999
  - LDA(Latent Dirichlet Allocation), 2003
  - Hierarchical Bayesian model
- 主题模型的R语言实现示例

主题模型的R语言实现示例

- 在R语言中，有两个包（package）提供了LDA模型：lda和topicmodels。
  - lda提供了基于Gibbs采样的经典LDA、MMSB（the mixed-membership stochastic blockmodel)、RTM（Relational Topic Model）和基于VEM（variational expectation-maximization）的sLDA (supervised LDA)、RTM.。
  - topicmodels基于包tm，提供LDA_VEM、LDA_Gibbs、CTM_VEM（correlated topics model）三种模型。
- 可视化包——LDAvis包

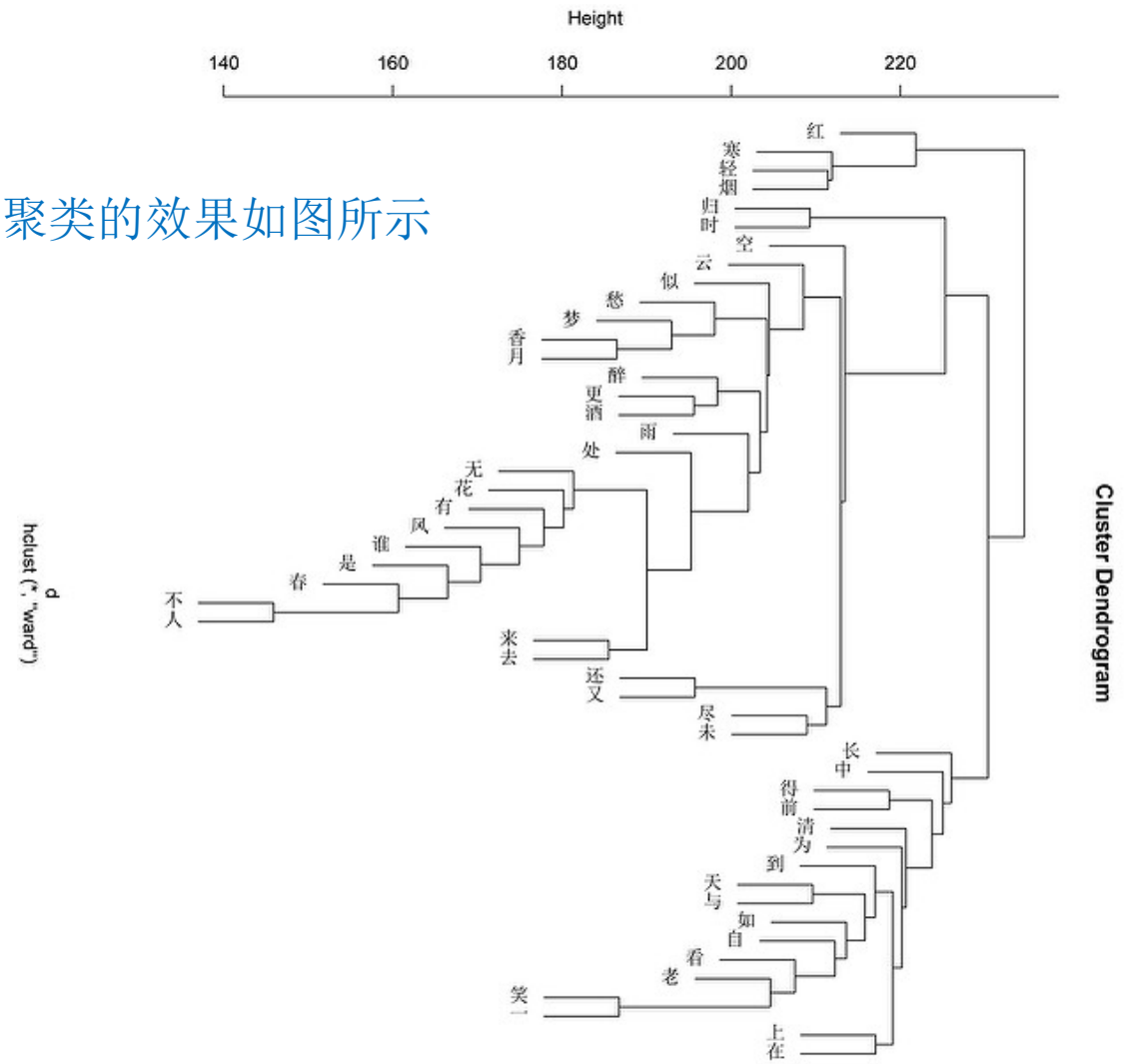
http://blog.csdn.net/sinat_26917383/article/details/51547298 retrieved: 2017.04.06

宋词的词频

东风夜放花千树：对宋词进行主题分析初探
http://chengjunwang.com/cn/2013/09/topic-modeling-of-song-peom/

根据词与词之间共现概率对词进行聚类

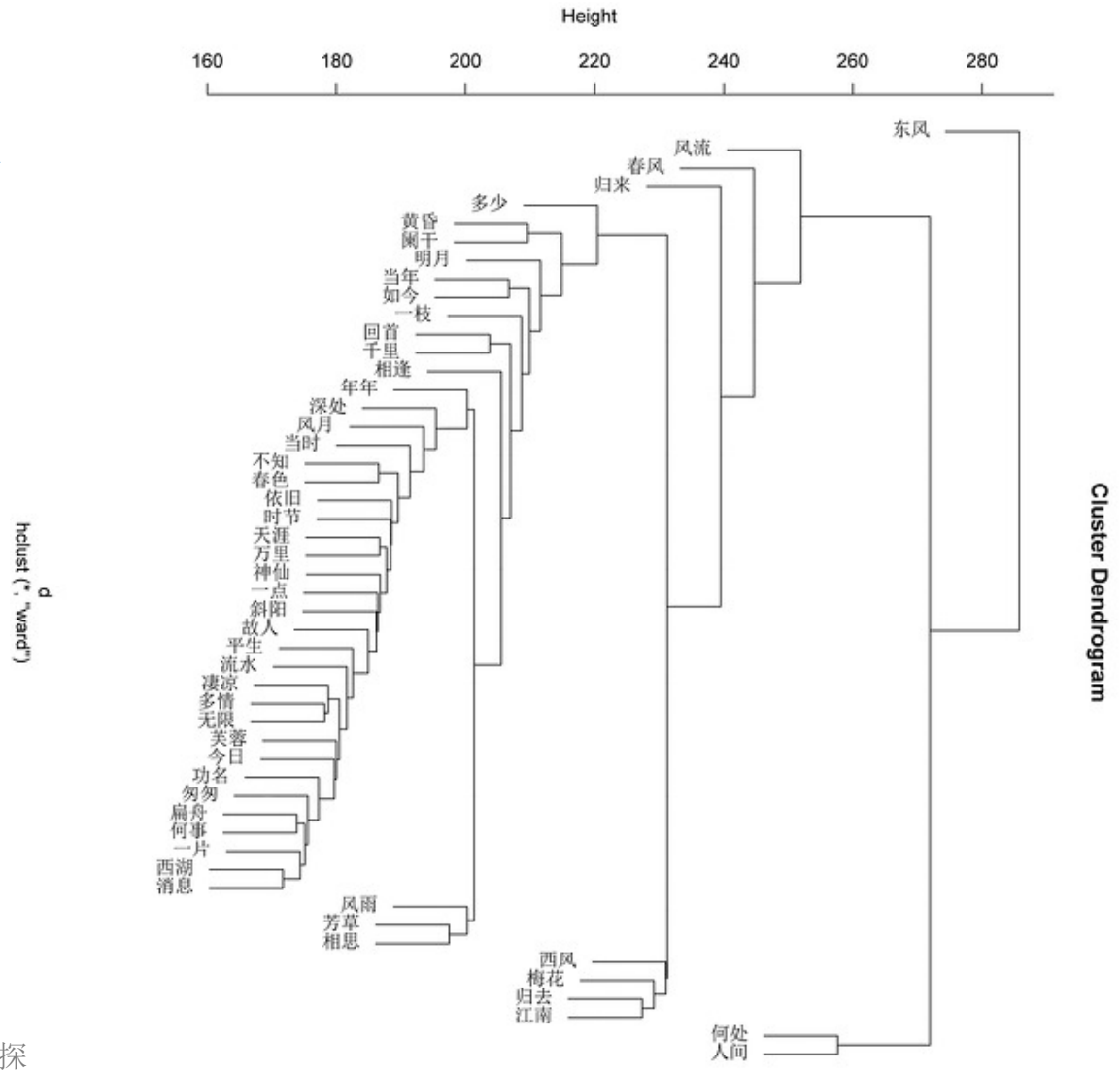
长度大于1词聚类的效果如图所示



东风夜放花千树：对宋词进行主题分析初探
http://chengjunwang.com/cn/2013/09/topic-modeling-of-song-peom/

根据词与词之间共现概率对词进行聚类

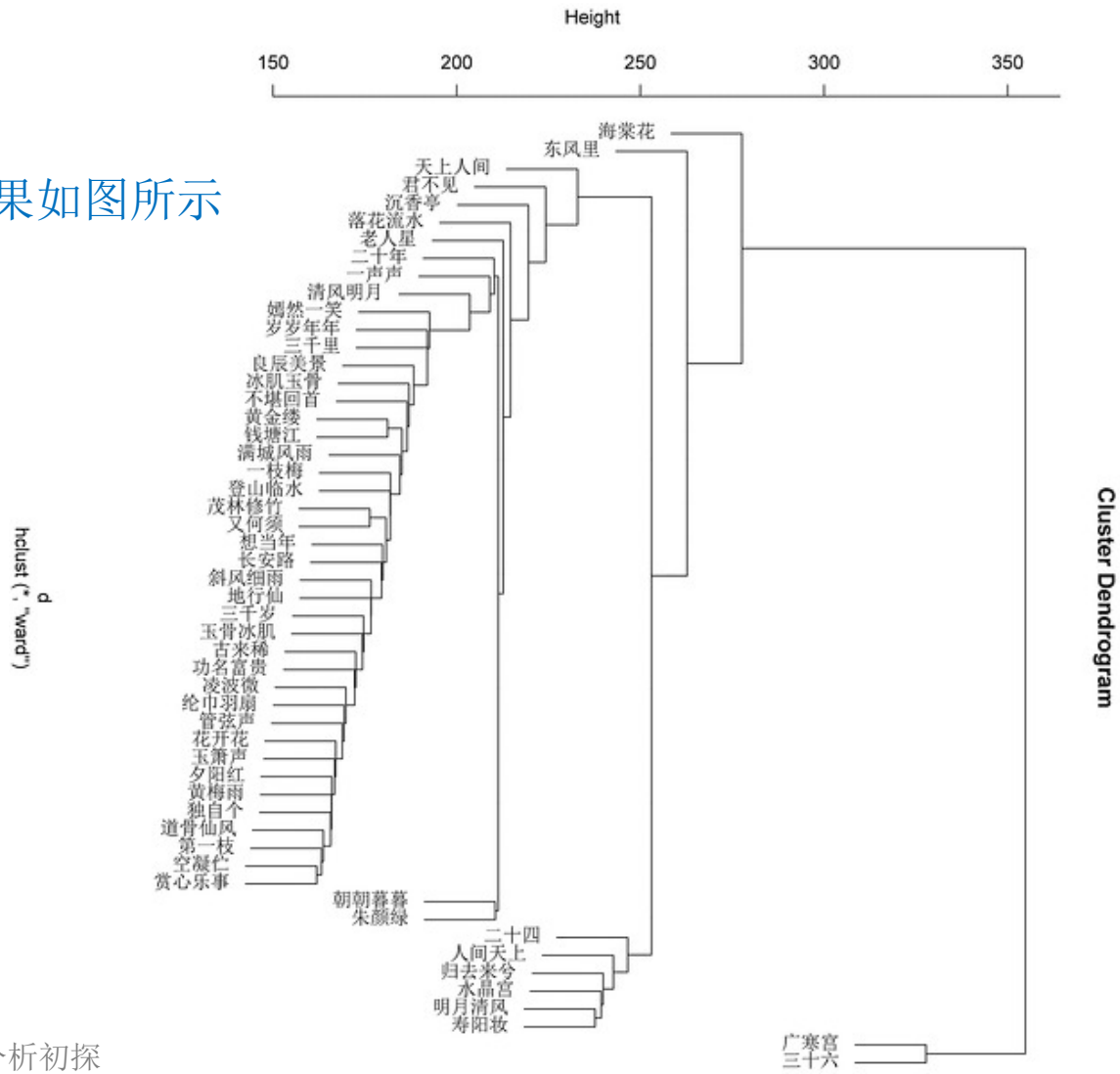
对长度大于2的词的聚类结果如图所示，可见宋词的确注重“风流倜傥”，连分类都和风向有关系。



东风夜放花千树：对宋词进行主题分析初探
http://chengjunwang.com/cn/2013/09/topic-modeling-of-song-peom/

根据词与词之间共现概率对词进行聚类

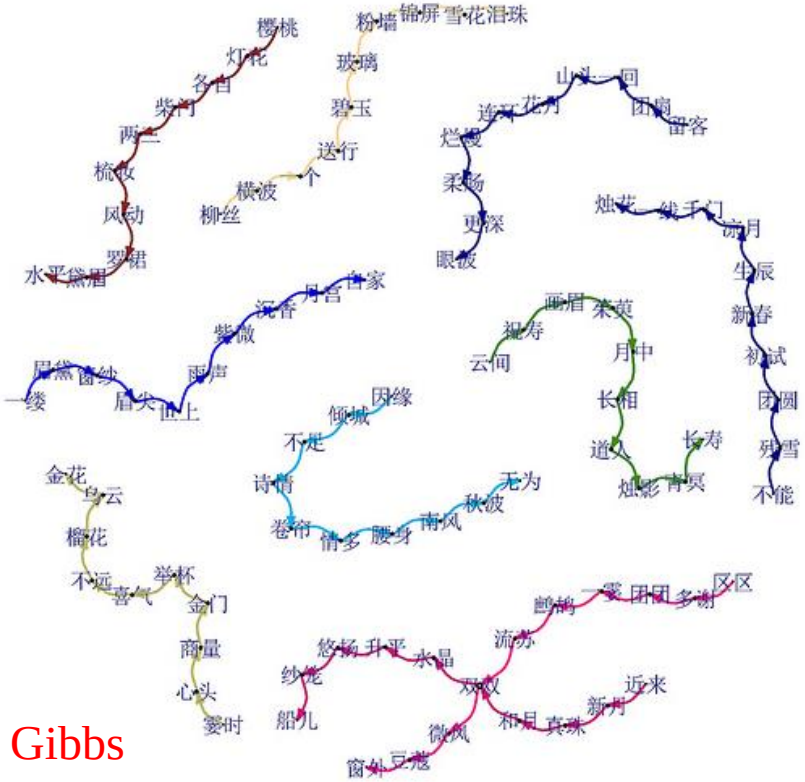
长度大于3的词聚类结果如图所示

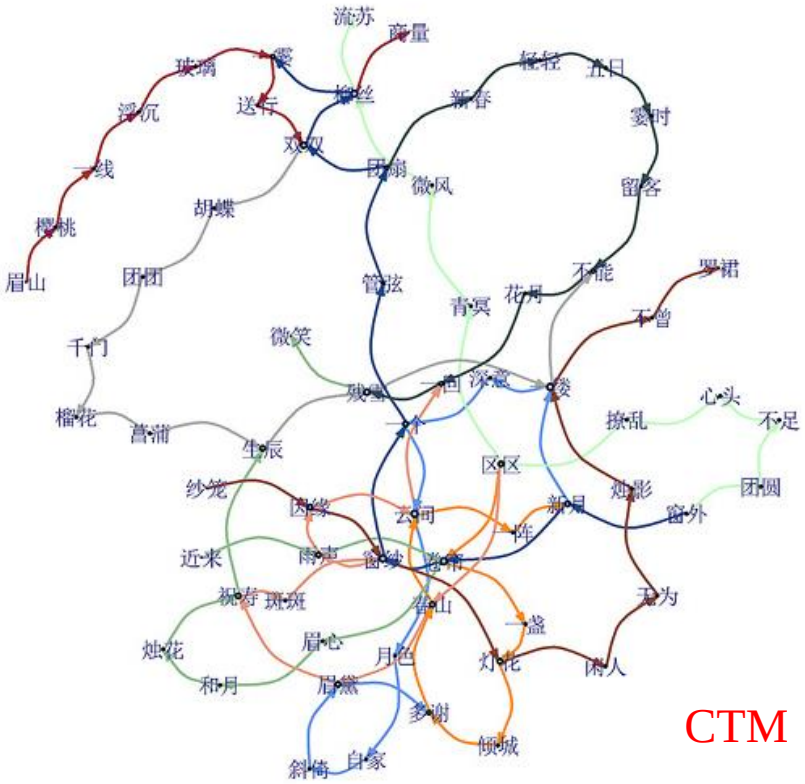


东风夜放花千树：对宋词进行主题分析初探
http://chengjunwang.com/cn/2013/09/topic-modeling-of-song-peom/

主题网络图

topicmodels这个R包是由Bettina Grun和 Johannes Kepler两个人贡献的，目前支持VEM(variational expectation-maximization), VEM (fixed alpha)，Gibbs和CTM(correlated topics model)四种主题模型，关于其详细介绍，可以阅读他们的论文，关于主题模型的更多背景知识可以阅读Blei的相关文章。

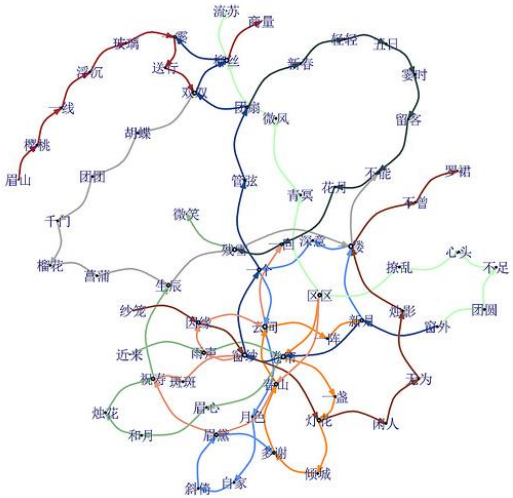


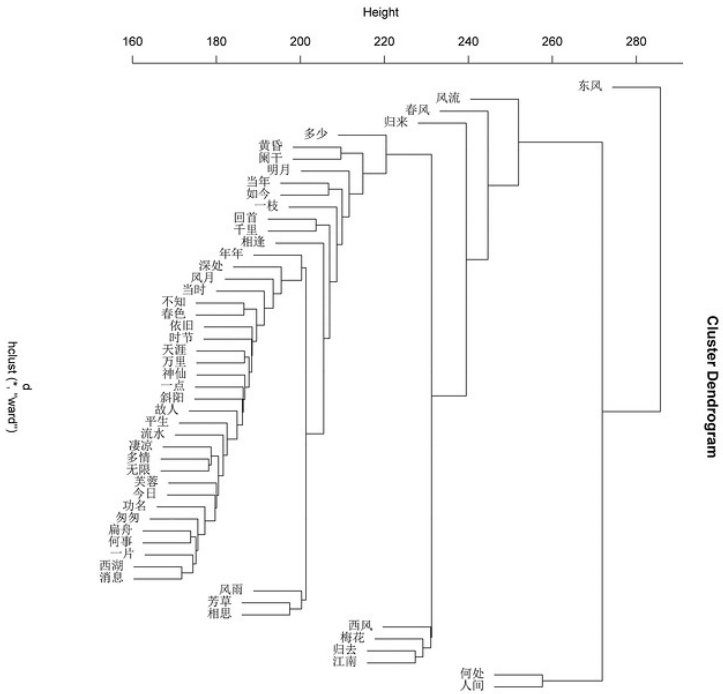


东风夜放花千树：对宋词进行主题分析初探
http://chengjunwang.com/cn/2013/09/topic-modeling-of-song-peom/

小结：主题模型的R语言实现示例

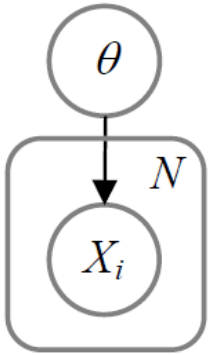
- R语言的已有package
  - lda
  - topicmodels
  - 可视化包：LDAvis





概率图及主题模型
Probabilistic Graphical Models / Topic Model

- 什么是Graphical Model
  - 定义、示例
  - Representation、Inference、Learning
- 主题模型与分类
  - LSA (Latent Semantic Analysis), 1990
  - pLSA (probabilistic Latent Semantic Analysis), 1999
  - LDA(Latent Dirichlet Allocation), 2003
  - Hierarchical Bayesian model
- 主题模型的R语言实现示例



盘式记法