主题模型

Topic Models

王博 天津大学智能与计算学部 2019.12

目录

- 什么是主题模型?
- 三种主题模型： LSA，pLSA，LDA
- 主题模型的应用

什么是主题模型?

什么是主题模型?

我们经常需要判断两篇文档的相似度，怎么做呢?

文档 – 词汇 矩阵 （二值）

匹配词汇呀！ 每篇文章用其中的词汇分布来表示

文档

词汇	D1	D1	D1	D1	D1	D1
ANTHONY	1	1	0	0	0	1
BRUTUS	1	1	0	1	0	0
CAESAR	1	1	0	1	1	1
CALPURNIA	0	1	0	0	0	0
CLEOPATRA	1	0	0	0	0	0
MERCY	1	0	1	1	1	1
WORSER	1	0	1	1	1	0

除了是否出现，词汇出现的次数明显更能区分文章的内容

什么是主题模型?

我们经常需要判断两篇文档的相似度，怎么做呢?

文档 – 词汇 矩阵 （词频）

匹配词汇呀！ 每篇文章用其中的词汇分布来表示

文档

词汇	D1	D1	D1	D1	D1	D1
ANTHONY	157	73	0	0	0	1
BRUTUS	4	157	0	2	0	0
CAESAR	232	227	0	2	1	0
CALPURNIA	0	10	0	0	0	0
CLEOPATRA	57	0	0	0	0	0
MERCY	2	0	3	8	5	8
WORSER	2	0	1	1	1	5

像 “the”，“and”这样的词汇词频很高，却不能表达一篇文章的特点

什么是主题模型?

- **逆文本频率（term frequency-inverse document frequency, TF-IDF）**

$$\mathrm{TFIDF}_{ij} = \frac{\mathrm{tf}_{ij}}{\mathrm{tf}_{\bullet j}} \log \frac{\mathrm{df}}{\mathrm{df}_i}, \quad i = 1, 2, \cdots, m; \quad j = 1, 2, \cdots, n$$

- tf_{ij}：单词 w_i 出现在文本 d_j 中的频数
- tf_j：是文本 d_j 中出现的所有单词的频 数之和
- df_i：含有单词 w_i 的文本数
- df： 是文本集合D的全部文本数

文档 – 词汇 矩阵 （TF-IDF）

文档

词汇	D1	D1	D1	D1	D1	D1
ANTHONY	5.25	3.18	0.0	0.0	0.0	0.35
BRUTUS	1.21	6.10	0.0	1.0	0.0	0.0
CAESAR	8.59	2.54	0.0	1.51	0.25	0.0
CALPURNIA	0.0	1.54	0.0	0.0	0.0	0.0
CLEOPATRA	2.85	0.0	0.0	0.0	0.0	0.0
MERCY	1.51	0.0	1.90	0.12	5.25	0.88
WORSER	1.37	0.0	0.11	4.15	0.25	1.95

最能够代表一篇文章的词汇（TF-IDF值高的词汇），在这篇文章中出现的多（TF），但是在其他文章中出现的少（IDF）

什么是主题模型?

用词汇向量表达文章有什么问题?

词汇向量是一个“词袋”，丢失了词汇的语义结构。

文章的语义是词汇组成的模式，而不仅仅是词汇。

怎么表达文章的语义?

什么是主题模型?

一篇文章的语义非常丰富，想要准确全面的刻画比较困难

但是判断一篇文章“说了哪方面的事情”的则相对容易，这就是文章的“主题”

文章

主题?

语义?

为文章匹配提供文章表达

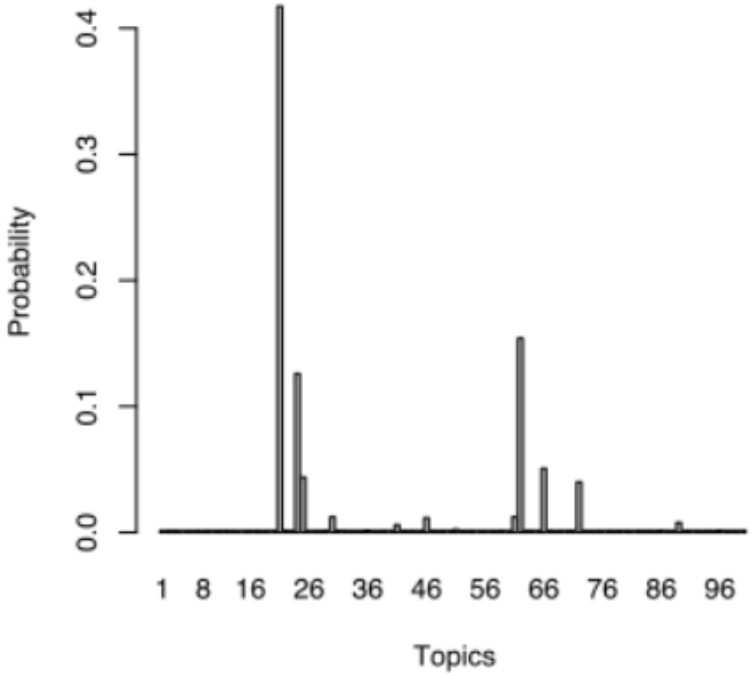
信息检索

为文章中的语言分析提供背景

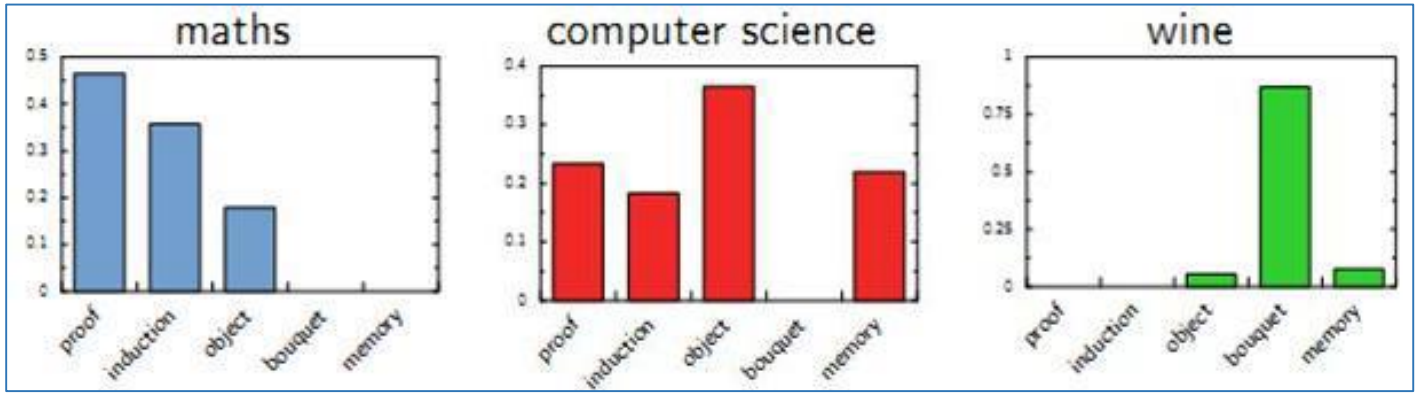
语义消歧

什么是主题模型?

- 一篇文章可以包含多个主题
- 每个主题涉及多个词汇



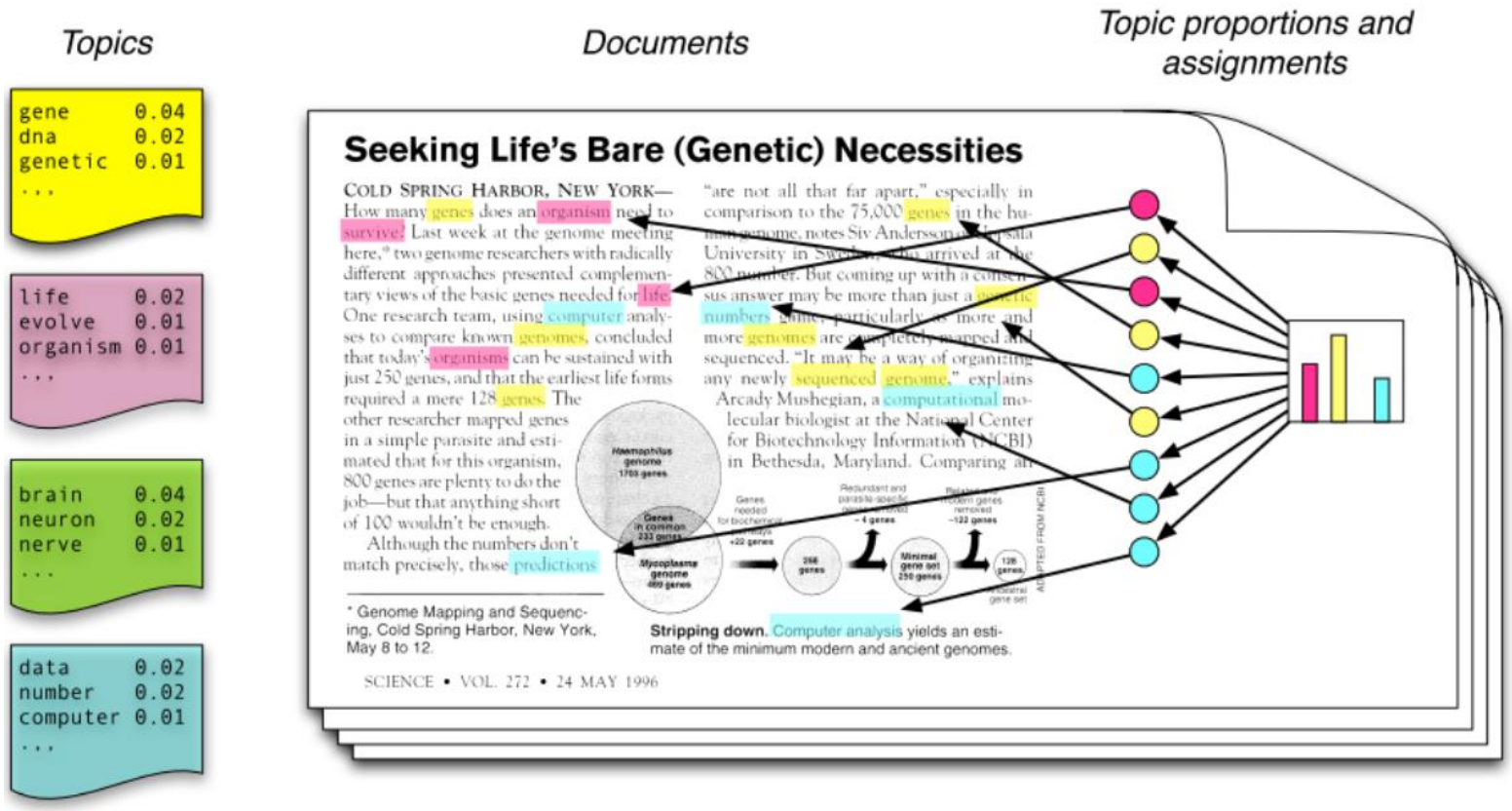
一个文档中的主题分布



每个主题中的词汇分布

什么是主题模型?

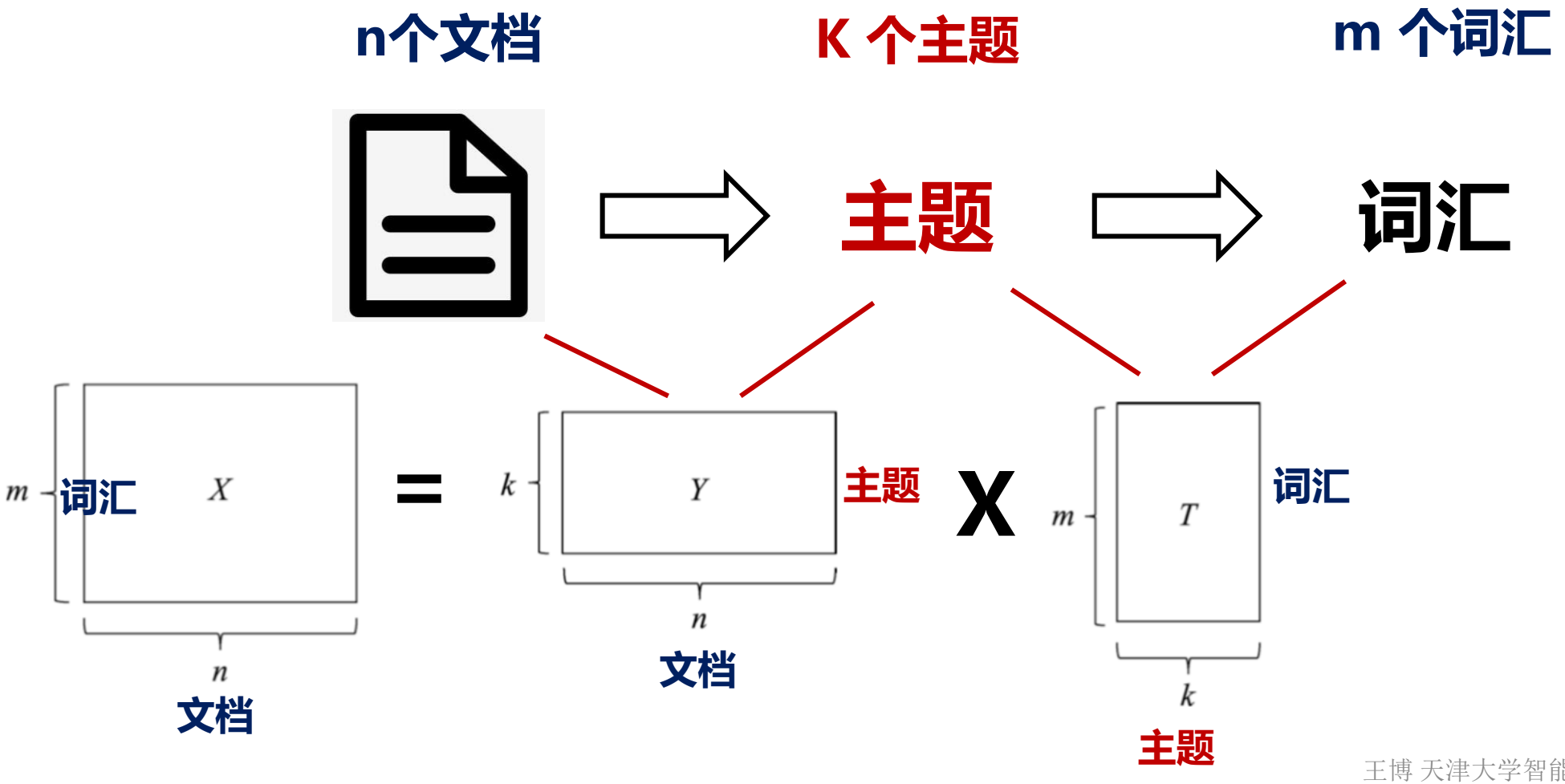
- 每个主题决定了文章中一部分内容
- 多个主题共同决定了文章整体



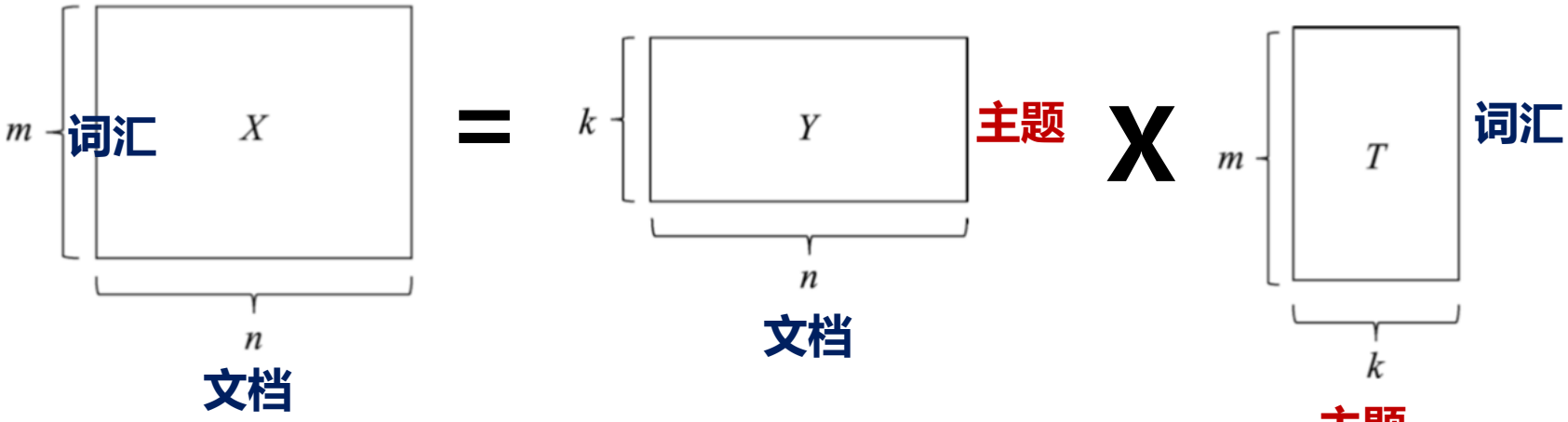
潜在语义分析 LSA

王博 天津大学智能与计算学部 2019.12

潜在语义分析 LSA（Latent Semantic Analysis)



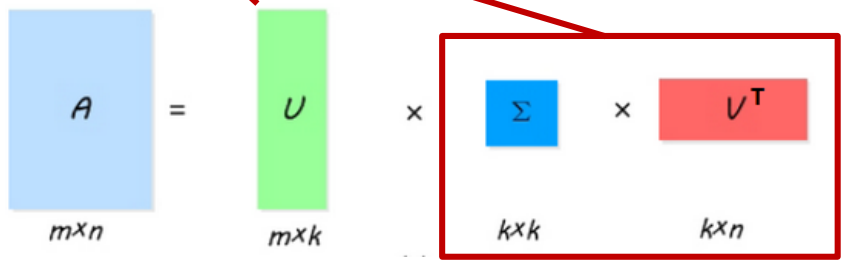
潜在语义分析 LSA（Latent Semantic Analysis）



$$X \approx TY$$

给定X，如何求T，Y?

奇异值分解（SVD）



潜在语义分析 LSA

潜在语义分析 LSA（Latent Semantic Analysis)

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

Book	0.15	−0.27	0.04
Dads	0.24	0.38	−0.09
Dummies	0.13	−0.17	0.07
Estate	0.18	0.19	0.45
Guide	0.22	0.09	−0.46
Investing	0.74	−0.21	0.21
Market	0.18	−0.30	−0.28
Real	0.18	0.19	0.45
Rich	0.36	0.59	−0.34
Stock	0.25	−0.42	−0.28
Value	0.12	−0.14	0.23

*

3.91	0	0
0	2.61	0
0	0	2.00

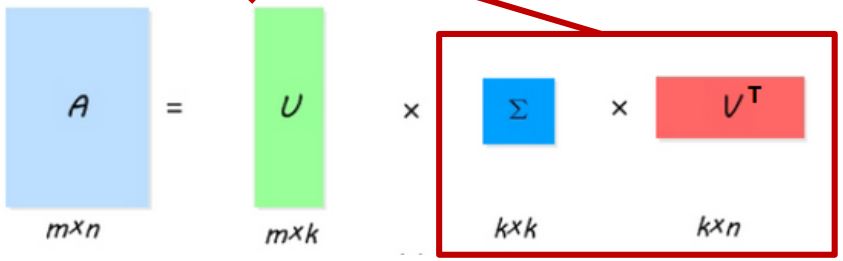
*

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.35	0.22	0.34	0.26	0.22	0.49	0.28	0.29	0.44
−0.32	−0.15	−0.46	−0.24	−0.14	0.55	0.07	−0.31	0.44
−0.41	0.14	−0.16	0.25	0.22	−0.51	0.55	0.00	0.34

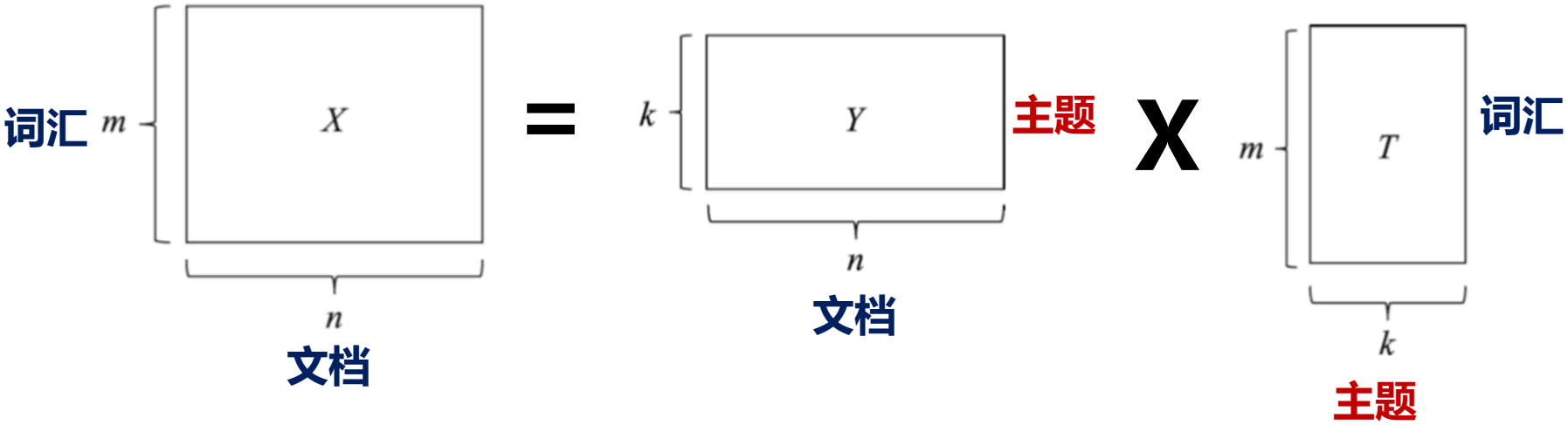
$$X \approx TY$$

给定X，如何求T，Y?

奇异值分解（SVD）



潜在语义分析 LSA（Latent Semantic Analysis)



LSA 的一个根本问题在于，尽管我们可以把Y的行或T的列看成是一个话题，但是由于每一列的值都可以看成是几乎没有限制的实数值（不是[0,1]概率），因此我们无法去进一步解释这些值到底是什么意思，也更无法从概率的角度来理解这个模型。

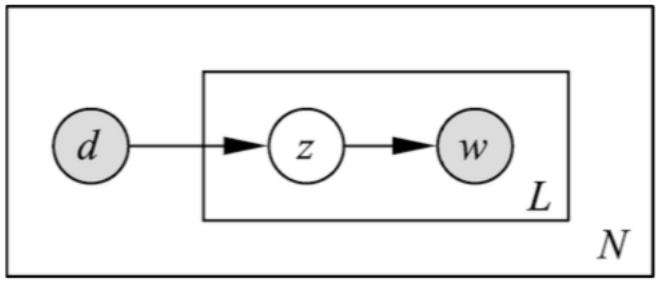
换言之，LSA中，矩阵X是有物理意义的，而Y和T是缺乏物理意义的。LSA只是形式上拟合了文档-主题-词汇的关系，但并没有真正表达这种关系

概率潜在语义分析 pLSA

王博 天津大学智能与计算学部 2019.12

概率潜在语义分析 pLSA（Latent Semantic Analysis)

⇒ **主题** ⇒ **词汇**



N是文档数量，L是某个文档中词汇数量

概率分布P(d)、条件概率分布P(z|d)、条件概率分布P(w|z)皆属于多项分布

P(d)：生成文本d的概率

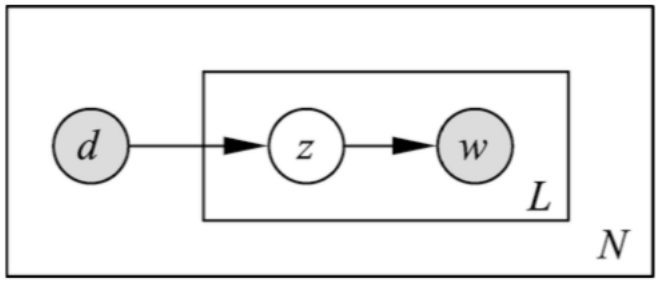
P(z|d)：文本d生成话题z的概率

P(w|z)：话题z生成单词w的概率

概率潜在语义分析 pLSA（Latent Semantic Analysis)

主题

词汇



N是文档数量，L是某个文档中词汇数量

生成模型通过以下步骤生成文本-单词共现数据：

（1）依据概率分布P(d)，从文本（指标）集合中随机选取一个文本d，共生成N 个文本；针对每个文本，执行以下操作

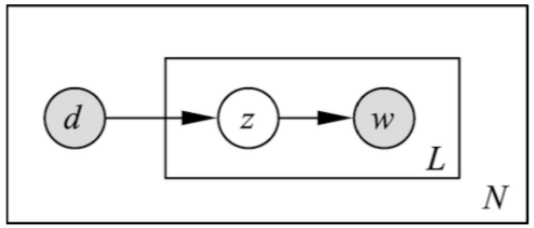
（2）在文本d给定条件下，依据条件概率分布P(z|d)，从话题集合随机选取一个话题z，共生成L个话题，这里L是文本长度

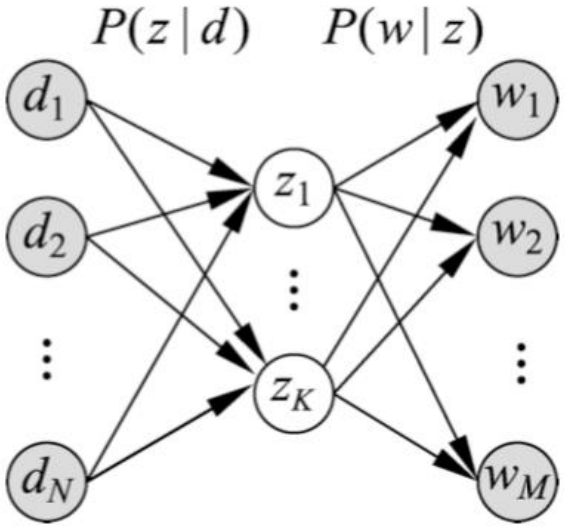
（3）在话题z给定条件下，依据条件概率分布P(wlz)，从单词集合中随机选取一个单词w

概率潜在语义分析 pLSA（Latent Semantic Analysis)

主题

词汇





- 文档d和词汇w都是已知的观察值。
- 主题z是隐变量。
- 要想求Z，关键要知道参数P(z|d)和P(w|z)

EM算法

概率潜在语义分析 pLSA

输入：设单词集合为 $W=\{w_1,w_2,\cdots,w_M\}$，文本集合为 $D=\{d_1,d_2,\cdots,d_N\}$，话题集合为 $Z=\{z_1,z_2,\cdots,z_K\}$，共现数据 $\{n(w_i,d_j)\},i=1,2,\cdots,M,j=1,2,\cdots,N$；

输出：$P(w_i|z_k)$ 和 $P(z_k|d_j)$。

（1）设置参数 $P(w_i|z_k)$ 和 $P(z_k|d_j)$ 的初始值。

（2）迭代执行以下 E 步，M 步，直到收敛为止。

E 步：

$$P(z_k|w_i,d_j)=\frac{P(w_i|z_k)P(z_k|d_j)}{\sum_{k=1}^{K}P(w_i|z_k)P(z_k|d_j)}$$

M 步：

$$P(w_i|z_k)=\frac{\sum_{j=1}^{N}n(w_i,d_j)P(z_k|w_i,d_j)}{\sum_{m=1}^{M}\sum_{j=1}^{N}n(w_m,d_j)P(z_k|w_m,d_j)}$$

$$P(z_k|d_j)=\frac{\sum_{i=1}^{M}n(w_i,d_j)P(z_k|w_i,d_j)}{n(d_j)}$$

潜在狄利克雷分配 LDA

王博 天津大学智能与计算学部 2019.12

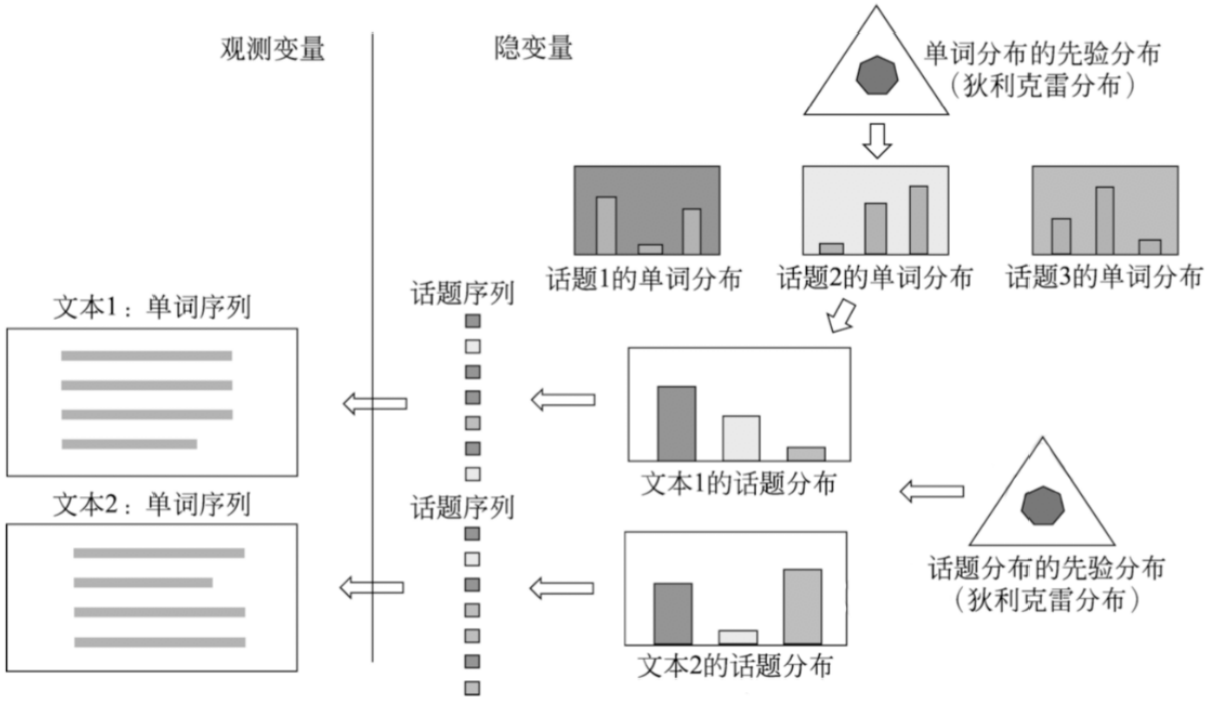
潜在狄利克雷分配 LDA (latent Dirichlet allocation)

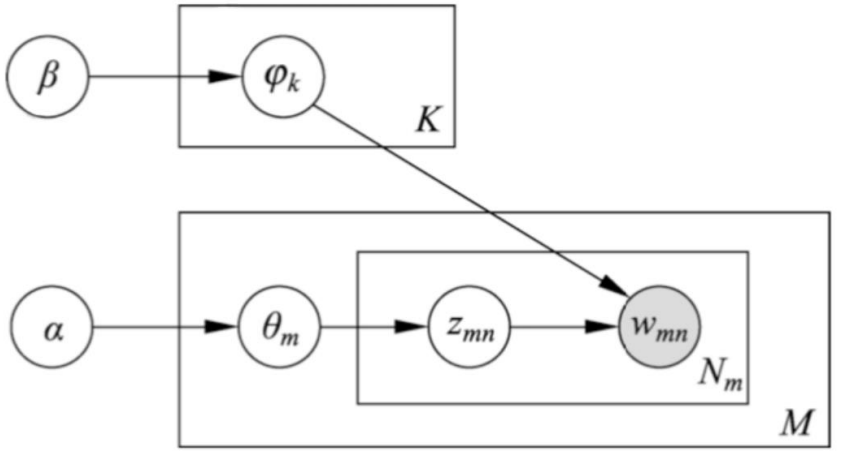
pLSA中单词和主题的先验分布都假设是均匀分布的，也就是假设我们对他们的先验分布一无所知。

这种假设使得pLSA比较容易出现过拟合。

文档生成话题和话题生成单词的过程是典型的多项分布，在贝叶斯学习中，狄利克雷分布常作为多项分布的先验分布使用

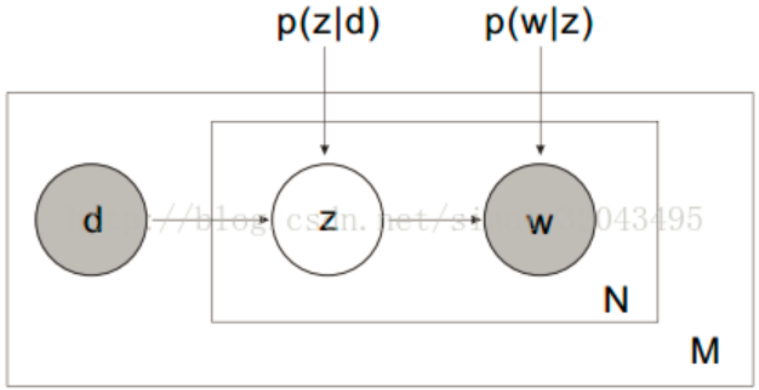
LDA将狄利克雷分布做为话题和单词生成的先验分布





潜在狄利克雷分配 LDA

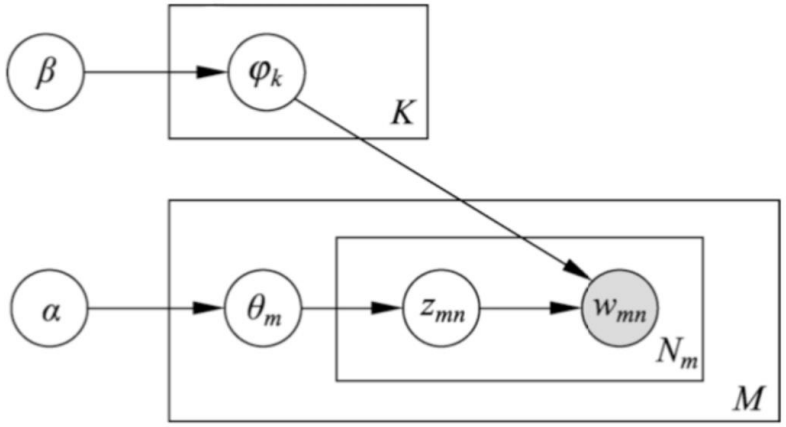
pLSA的生成模型



假设

- 每篇文档的文档-主题分布θ已知且确定
- 每篇文档的topic-word分布φ已知且确定

LDA的生成模型



- 文档的θ和φ均未知，由Dirichlet分布确定

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

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

主题模型的应用

- 文档降维
- 信息提取和搜索
- 语义分析
- 文档分类/聚类，文章摘要，社区挖掘
- 基于内容的图像聚类
- 推荐系统

Thanks !