0. 写在前面

1 本课程总体结构

章节		教学内容
第一章 引言 (刘均, 2)		概念与研究背景；主要任务；挑战与研究方向；相关资源
第二章：自然语言的统计特性 (刘均, 1)		Zipf定律、Heaps定律、Benford 定律。
第三章：语言模型	词袋模型 (刘均, 3)	语言模型；词袋模型（BoW）；TF-IDF。 NLU任务：情感分析、文本聚类。
	概率语言模型 (李辰, 6)	概率语言模型；n-gram 模型；最大似然估计；平滑技术。 NLU任务：分词、语义关系抽取。
	主题模型 (刘均, 6)	生成模型；主题模型的图表示；LSA、PLSA、LDA；NMF等。 NLU任务：话题检测、推荐。
	神经网络语言模型 (李辰, 6)	分布式表示；C&W、CBOW、Skip-Gram、Glove等。 NLU任务：对话、实体消歧。
第四章：机器翻译	概述 (李辰, 1)	面临的挑战；发展历程；方法类别及特点；MT评估。
	统计机器翻译 (李辰, 3)	统计MT；Noisy Channel模型；IBM模型。
	神经网络机器翻译 与大语言模型 (刘均, 4)	RNN与LSTM简介；Encoder-Decoder框架；Attention模型；大语言模型。

- 这门课由于由两门老师授课，个人感觉结构比较混乱
- 由于时间紧任务重经费无，所以笔记还是按PPT内容和以上结构展开，即使有很多不合理的地方

2 考试有关事项

1. 基于矩阵分解的主题模型

1.0. 概述

1️⃣分布模型：$\text{Doc.} \xrightarrow[\text{非监督学习(聚类)}]{\text{主题分布}} \begin{cases} \textbf{Topic 1}(P_{T_1}) \xrightarrow[\text{非监督学习(聚类)}]{\text{词语分布}} \begin{cases} \text{Word}_{11}(P_{W_{11}}) \\ \text{Word}_{12}(P_{W_{11}}) \\ \vdots \\ \text{Word}_{1m}(P_{W_{1m}}) \end{cases} \\ \textbf{Topic 2}(P_{T_2}) \xrightarrow[\text{非监督学习(聚类)}]{\text{词语分布}} \begin{cases} \text{Word}_{21}(P_{W_{21}}) \\ \text{Word}_{22}(P_{W_{21}}) \\ \vdots \\ \text{Word}_{2m}(P_{W_{2m}}) \end{cases} \\ \vdots \\ \textbf{Topic n}(P_{T_n}) \xrightarrow[\text{非监督学习(聚类)}]{\text{词语分布}} \begin{cases} \text{Word}_{n1}(P_{W_{n1}}) \\ \text{Word}_{n2}(P_{W_{n1}}) \\ \vdots \\ \text{Word}_{nm}(P_{W_{nm}}) \end{cases} \end{cases}$

1. 主题分布：每篇文档由若干主题按一定比例构成
2. 词语分布：每个主题包含一组特定的词语，每个词具有不同的出现概率

2️⃣概率模型

1. 公式：$p(w|\text{Doc}) = \sum_{i=1}^{n} p\left(w|T_i\right) \cdot p\left(T_i|\text{Doc}\right)$
2. 含义：将文档的内容视为不同主题的组合→由每主题的词语概率预测文档中词语的分布

1.1. LSA(SVD)模型

1️⃣奇异值分解

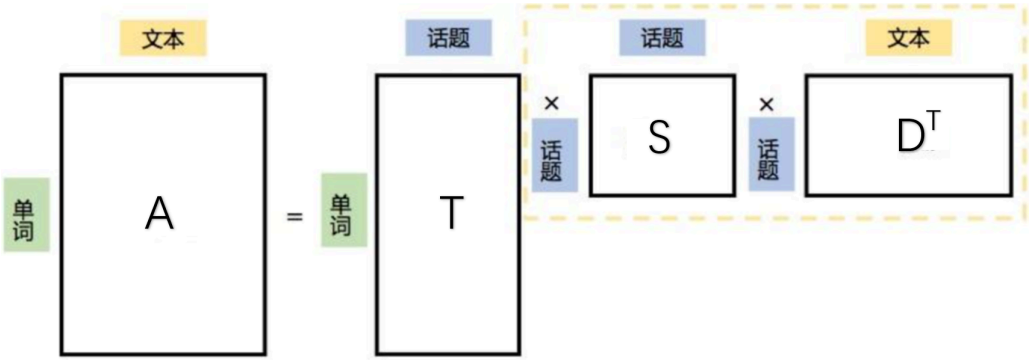
1. 含义：对任意$A_{m\times n}$可将其分解为三个矩阵$A_{m\times n}=U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$

矩阵类型	描述
左奇异矩阵 $U_{m\times m}$	为正交矩阵即$U_{m\times m}U^T_{m\times m}=I_{m\times m}$
奇异值矩阵 $\Sigma_{m\times n}$	为对角矩阵(对角为是奇异值)，如 $\begin{bmatrix} \alpha_1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \alpha_2 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \alpha_3 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_m & \cdots & 0 \end{bmatrix}_{m\times n}$
右奇异矩阵 $V_{n\times n}$	为正交矩阵即$V_{n\times n}V^T_{n\times n}=I_{n\times n}$

2. Eckart–Young–Mirsky定理：$A_k = U_k\Sigma_k V_k^T$奇异值的截断
   - U_k 和 V_k 分别是 U 和 V 的前 k 列
   - Σ_k 是奇异值矩阵 Σ 中前 k 个最大的奇异值组成的 $k\times k$ 子矩阵

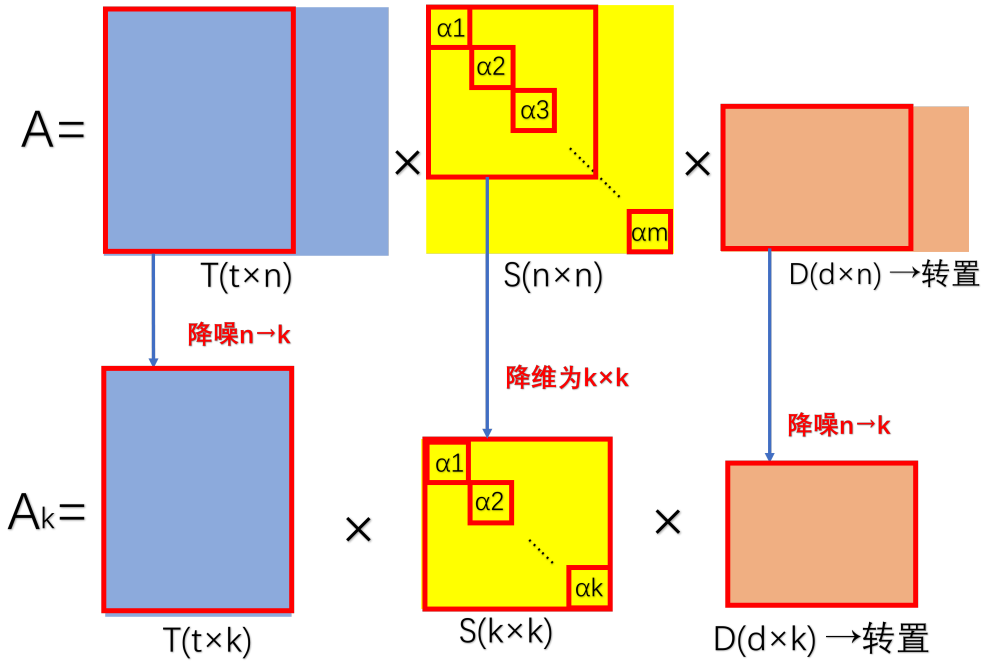
2 LSA模型步骤：原始Word-Doc矩阵 $\xrightarrow[\text{近似}]{\text{奇异分解}}$ 其近似的低阶矩阵

1. Word-Doc矩阵：
   - $A_{t\times d}=\begin{bmatrix} \text{Doc}_1\to\text{Word}_{11} & \text{Doc}_2\to\text{Word}_{12} & \cdots & \text{Doc}_n\to\text{Word}_{1d} \\ \text{Doc}_1\to\text{Word}_{21} & \text{Doc}_2\to\text{Word}_{22} & \cdots & \text{Doc}_n\to\text{Word}_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \text{Doc}_1\to\text{Word}_{t1} & \text{Doc}_2\to\text{Word}_{t2} & \cdots & \text{Doc}_n\to\text{Word}_{td} \end{bmatrix}$
   - $\text{Doc}_i\to\text{Word}_{ij}$可为词$\text{Word}_{ij}$的词频或者TF-IDF值
2. $A_{t\times d}$奇异分解：$A_{t\times d}=T_{t\times n}S_{n\times n}D^T_{d\times n}$



矩阵类型	描述
$S_{n\times n}$	奇异值按降序排列，代表重要的潜在语义的强度
$T_{t\times n}$	词汇矩阵，每列蕴含一个隐含概念(主题)
$D_{d\times n}$	文档矩阵，每列蕴含一个隐含概念(主题)

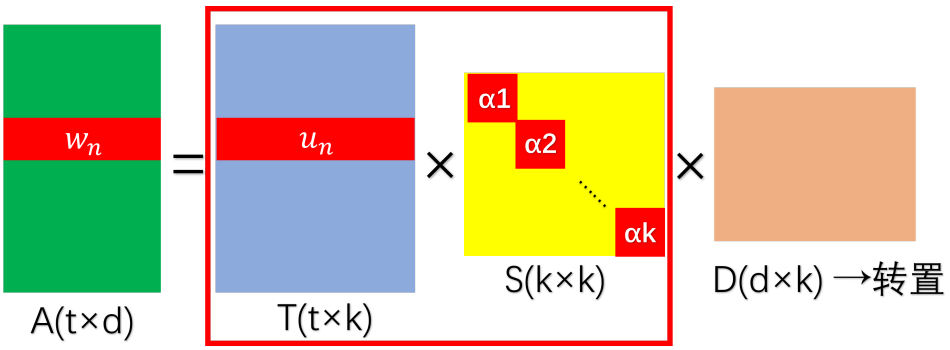
3. 低秩近似：$A\to A_k$



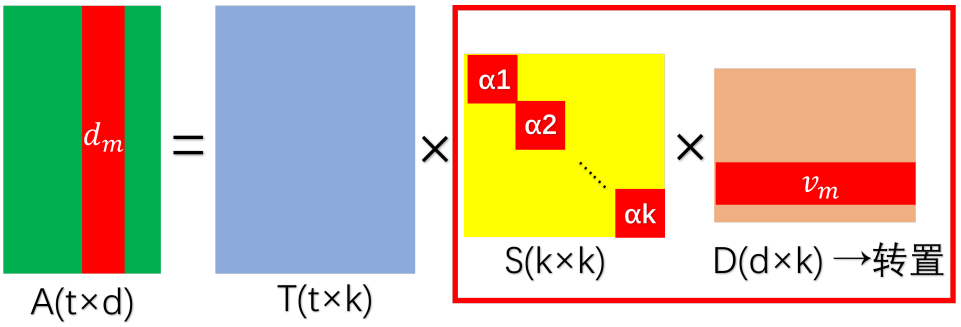
   - 降维：$S_{n\times n}\xrightarrow{\text{只保留前}k\text{个最大的奇异值}}S_{k\times k}$，其中$k$又称为预期主题数
   - 降噪：$A_{t\times d}=T_{t\times n}S_{n\times n}D^T_{d\times n}\xrightarrow{S_{n\times n}\text{降维}}A_{t\times d}=T_{t\times k}S_{k\times k}D^T_{d\times k}$，滤掉不重要的主题

3 文档与词汇的表示

1. 词汇：$T_{t\times k}S_{k\times k}$的行向量，且$\hat{w}_n = u_n \times \mathbf{S}$



2. 文档：$D_{d\times k}S_{k\times k}$的行向量($S_{k\times k}D^T_{d\times k}$的列向量)，且$\hat{d}_m = \mathbf{S} \times v_m^T$



1.2. MNF建模

1 建模过程

1. 对$\mathbf{V}$寻找非负矩阵$\mathbf{HW}$使$\mathbf{V} \approx \mathbf{WH}$
2. 使得代价函数$\|V - WH\| = \sqrt{\sum_{i,j}(V_{i,j} - (WH)_{i,j})^2}$尽可能小

2 建模的意义

1. 非负：使分解结果更有意义
2. 示例：**(文档-单词)**$\xrightarrow{\text{NMF}}$**(文档-主题)**×**(主题-单词)**

2. 基于概率的主题模型

2.0. 概率模型概述

1 符号：其中K为话题数，$K \ll M$且为预先定义的超参数

集合	含义	随机变量
文本集$D=\{d_1, d_2, \dots d_N\}$	包含所有文本，N为文本总数	d(观测变量)
话题集$Z=\{z_1, z_2, \dots z_K\}$	包含所有可能的话题，K为预设话题总数	z(隐藏变量)
词汇集$W=\{w_1, w_2, \dots w_M\}$	所有可能的单词，M为单词总数	w(观测变量)

2️⃣三类分布：$P(d)$为可观测参数，如何估计$P(z|d)$和$P(w|z)$两参数派生了pLAS和LDA方法

分布	表示	含义
文档分布	$P(d) \sim$ 多项分布	生成文本d的概率
主题分布	$P(z\|d) \sim$ 多项分布	文本d生成话题z的概率，每个文本都有其主题分布
单词分布	$P(w\|z) \sim$ 多项分布	话题z生成单词w的概率，每个主题都有其单词分布

3️⃣观测表征

1. 观测数据：文本-单词共现矩阵，其中n(单词i, 文本j)表示单词i在文本j中出现的次数

共现矩阵T	**文d_1**	**文d_2**	...	**文d_N**
词w_1	n(词w_1, 文d_1)	n(词w_1, 文d_2)	...	n(词w_1, 文d_N)
词w_2	n(词w_2, 文d_1)	n(词w_2, 文d_2)	...	n(词w_2, 文d_N)
...	...	...	...	
词w_M	n(词w_M, 文d_1)	n(词w_M, 文d_2)	...	n(词w_M, 文d_N)

2. 生成概率：假设每个单词分布独立，则有$P(T)=\prod_{(w,d)} P(w,d)^{n(w,d)}$

4️⃣LDA与pLSA

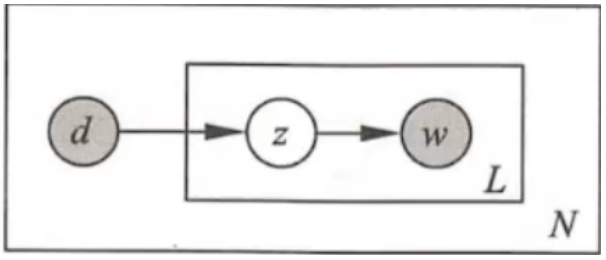
模型	思想	**对于两$P(z \mid d)$和$P(w \mid z)$待估参数**
pLSA	频率学	视作固定值(即均匀分布)，用最大似然估计解出来
LDA	贝叶斯	视作服从Dirichlet分布的随机变量，先验分布$\xrightarrow{\text{修正}}$最终分布

2.1 pLSA模型

1️⃣生成模型：

1. 定义：对生成概率$P(w,d)=P(d)\sum_z P(z|d)P(w|z)$形式的拆解

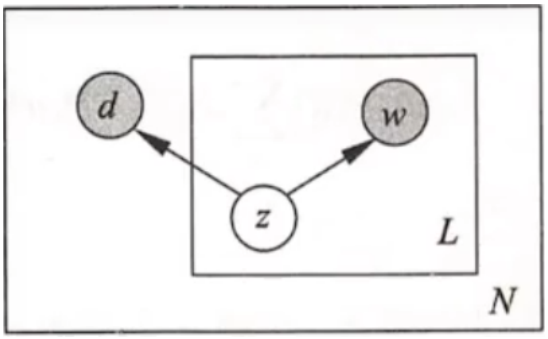
2. 概率依赖：文本→话题→单词



选择	描述	备注
文本	从D中，按$P(d)$选择文本$d$$\xrightarrow{\text{重复}N\text{次}}$生成$N$个文本	N为文本数量
话题	对每个文本，按$P(z\|d)$选择话题$z$$\xrightarrow{\text{重复}L\text{次}}$生成$L$个话题	L为文本(定/变)长
单词	对每个话题，按$P(w\|z)$选择一单词	N/A

2️⃣共现模型：

1. 定义：对生成概率$P(w,d)=\sum_{z\in Z}P(z)P(w|z)P(d|z)$形式的拆解
2. 概率依赖：话题→单词，话题→文本



选择	描述	备注
话题	从Z中，按$P(z)$选择话题$z$$\xrightarrow{\text{重复}L\text{次}}$生成$L$个话题	L为文本(定/变)长
单词	对每个话题，按$P(w\|z)$选择一单词	单词/文本的选择独立
文本	从D中，按$P(d\|z)$选择文本$d$$\xrightarrow{\text{重复}N\text{次}}$生成$N$个文本	N为文本数量

2.2. LDA模型简述

💢别看PPT了那就是一坨屎，以下内容来自维基百科

1️⃣LDA模型要素

1. 三种分布：

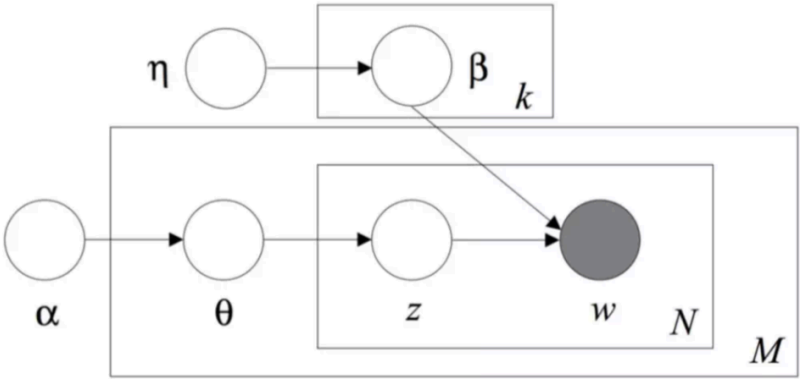
分布	维度	元素	隐藏/观测
主题分布Θ	文档数×主题数	$\theta_{i,j}$为文档i中主题j的比例	隐藏
词汇分布β	主题数×词汇数	$\beta_{i,j}$为主题i中词汇j的频次	隐藏
主题分布w	文档数×词汇数	$\mathrm{w}_{i,j}$为文档i中词汇j的频次	观测

2. 两种超参数：

超参数	描述	功能
α	文档集级参数，Dirichlet分布参数	生成文档的主题Θ
η	文档集级参数，Dirichlet分布参数	生成每个主题的β

2️⃣ LDA的生成：分布

$$p\left(w_i, z_i, \theta_i, \Phi \mid \alpha, \beta\right)=\prod_{j=1}^{N} p\left(\theta_i \mid \alpha\right) p\left(z_{i, j} \mid \theta_i\right) p(\Phi \mid \beta) p\left(w_{i, j} \mid \phi_{z_{i, j}}\right)$$



1. 第一部分：
    - 从先验Dirichlet分布α中抽样$\rightarrow$生成某一文档i的主题(多项式)分布θ_i
    - 从θ_i分布中抽样$\rightarrow$生成某一文档i的某一主题$z_{i,j}$
2. 第二部分：
    - 从先验Dirichlet分布η中抽样$\rightarrow$生成主题$z_{i,j}$的词语分布$\beta_{z_{i,j}}$
    - 从$\beta_{z_{i,j}}$分布中抽样$\rightarrow$生成词语$w_{i,j}$

4️⃣ LDA的求解(训练)：我也不信考试会考这B玩意儿

1. EM算法(Old-Fashioned)
2. Gibbs采，MCMC(Markov Chain Monte Carlo)算法

3.2.3. 番外: pLSA的EM求解

0️⃣ 总体思路

1. 极大似然估计：找到时$P(T)$最大的参数
2. EM算法：直接最大化对数似然函数非常困难，从而通过EM迭代的方式实现

1️⃣ 极大似然函数

1. 似然函数推导
    - 给定共现数据$\mathbf{T}=\{n(w_i,d_j)\}\rightarrow$要让$P(T)=\prod_{i,j}P(w_i,d_j)^{n(w_i,d_j)}$最大
    - 取对数+引入隐含变量：

$$\log P(T)=\sum_{i=1}^{M}\sum_{j=1}^{N}n(w_i,d_j)\times\log P(w_i,d_j)$$

$$\Big\Downarrow \text{引入隐含变量:}\ p(d_j)\sum_z p(z_k|d_j)p(w_i|z_k)$$

$$\log P(T)=\sum_{i=1}^{M}\sum_{j=1}^{N}\left(n(w_i,d_j)\times\left(\log p(d_j)+\log\left(\sum_z p(z_k|d_j)p(w_i|z_k)\right)\right)\right)$$

2. 似然函数分析：
   - 已知值：$n(w_i,d_j)$在$\mathbf{T}$向量中就有之，$p(d_j)$可由真实大量文本集得到
   - 参数值：$\log\left(\sum_z p(z_k|d_j)p(w_i|z_k)\right)$，其中$\log\sum$形式适合用EM算法解决

2 极大似然函数的下界

1. Jensen不等式

情况	$E(f(x))$与$f(E(x))$
$f(x)$为凸函数	$E(f(x))\geq f(E(x))$
$f(x)$为凹函数	$E(f(x))\leq f(E(x))$
$x=C$	$E(f(x))=f(E(x))$

2. $\log\left(\sum_z p(z_k|d_j)p(w_i|z_k)\right)$的处理：构建方差+应用Jensen不等式
   - $\sum_z p(z_k|d_j)p(w_i|z_k)\xrightarrow{z\text{的分布}Q(z)}\sum_z Q(z)\frac{p(z_k|d_j)p(w_i|z_k)}{Q(z)}\xrightarrow{X=\frac{p(z_k|d_j)p(w_i|z_k)}{Q(z)}}E(X)$
   - 原始$=\log(E(X))\xrightarrow[\text{Jensen不等式}]{\log(x)\text{为凹函数}}$原式
     $\geq E(\log(X))=\sum_z\left(\log\frac{p(z_k|d_j)p(w_i|z_k)}{Q(z)}\right)Q(z)$
3. 下界与极大似然：提升下界$\sum_z\left(\log\frac{p(z_k|d_j)p(w_i|z_k)}{Q(z)}\right)Q(z)$的最大值$\rightarrow$最大化似然函数

3 EM算法：详细步骤就不写了，我不信考试会考这B玩意儿

1. E步：确定Q函数$\rightarrow$表示当前参数估计下完全数据(观测数据+隐含变量)的对数似然的期望
   - 此处$Q=Q(z)=p(z_k|w_i,d_j)$
2. M步：迭代Q函数，不断更新参数$\rightarrow$使当前参数估计靠近最优值
   - 此处需要更新的参数为文档-主题分布$P(z|d)$，主题-词汇分布$P(w|z)$
   - 最终使$\sum_z\left(\log\frac{p(z_k|d_j)p(w_i|z_k)}{Q(z)}\right)Q(z)$最大，从而使$P(T)$最大