学　　期 2023-2024（2）

北京航空航天大學
BEIHANG UNIVERSITY

NLP 第二次大作业

院（系）名称	自动化科学与电气工程学院
专业名称	控制工程
学生姓名	胡正皓
学　　号	ZY2303205

2024 年 5 月

一、研究背景

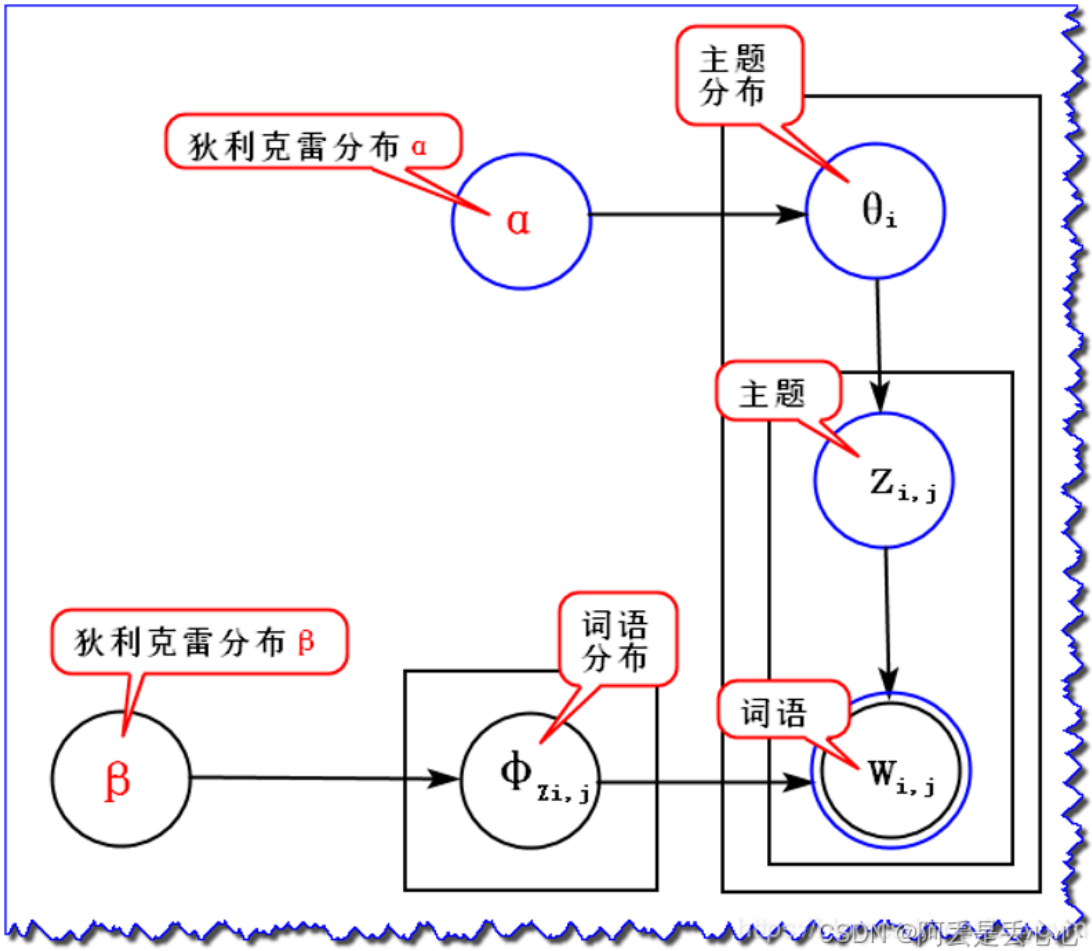
从下面链接给定的语料库中均匀抽取 1000 个段落作为数据集（每个段落可以有 K 个 token, K 可以取 20，100，500, 1000, 3000），每个段落的标签就是对应段落所属的小说。利用 LDA 模型在给定的语料库上进行文本建模，主题数量为 T，并把每个段落表示为主题分布后进行分类（分类器自由选择），分类结果使用 10 次交叉验证（i.e. 900 做训练，剩余 100 做测试循环十次）。实现和讨论如下的方面：（1）在设定不同的主题个数 T 的情况下，分类性能是否有变化？（2）以"词"和以"字"为基本单元下分类结果有什么差异？（3）不同的取值的 K 的短文本和长文本，主题模型性能上是否有差异？

二、LDA 模型

在 LDA 模型中，一篇文档生成的方式如下：

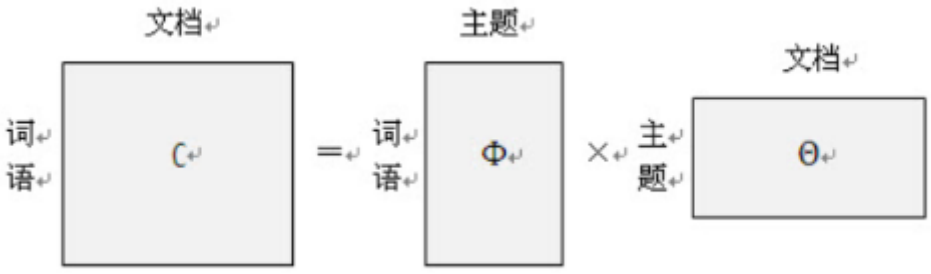
- 从狄利克雷分布 alpha 中取样生成文档 i 的主题分布 θi
- 从主题的多项式分布 θi 中取样生成文档 i 第 j 个词的主题 zij
- 从狄利克雷分布 beta 中取样生成主题 zij 对应的词语分布 φzij
- 从词语的多项式分布 φzij 中采样最终生成词语 wij

其中，类似 Beta 分布是二项式分布的共轭先验概率分布，而狄利克雷分布（Dirichlet 分布）是多项式分布的共轭先验概率分布。



如果我们要生成一篇文档，它里面的每个词语出现的概率为：

$$p(\text{词语}|\text{文档}) = \sum_{\text{主题}} p(\text{词语}|\text{主题}) \times p(\text{主题}|\text{文档})$$



看到文章推断其隐藏的主题分布，就是建模的目的。换言之，人类根据文档生成模型写成了各类文章，然后丢给了计算机，相当于计算机看到的是一篇篇已经写好的文章。现在计算机需要根据一篇篇文章中看到的一系列词归纳出当篇文章的主题，进而得出各个主题各自不同的出现概率：主题分布。

三、结果分析

在不同的 topic、token 以及字词分类下，模型的结果如表格所示:

	T	T=20	T=30	T=40	T=50	T=60
字	K=20	0.12	0.09	0.13	0.14	0.11
	K=100	0.27	0.29	0.35	0.33	0.37
	K=500	0.67	0.73	0.76	0.80	0.75
	K=1000	0.80	0.85	0.91	0.89	0.89
	K=3000	0.95	0.97	0.98	0.96	0.97
词	K=20	0.08	0.08	0.09	0.07	0.09
	K=100	0.14	0.10	0.10	0.09	0.13
	K=500	0.30	0.35	0.42	0.44	0.38
	K=1000	0.44	0.54	0.59	0.72	0.61
	K=3000	0.74	0.80	0.91	0.91	0.88

（1）可以看到，对于不同的主题个数，适当增加一定的主题个数有助于提高模型的准确度；

（2）对于字和词，字的效果要明显好于词分类的模型效果；

（3）对于不同的 k 的取值，可以看到 token 越多，模型的准确度也越高。