学　　期 2021-2022（2）

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理第三次作业

LDA 模型与中文文本分类

院（系）名称	自动化科学与电气工程学院
专业名称	电子信息
学生姓名	蔡尽云
学号	ZY2103501
指导老师	秦曾昌

2022 年 5 月

1 问题描述

从给定的语料库中均匀抽取 200 个段落（每个段落大于 500 个词）， 每个段落的标签就是对应段落所属的小说。利用 LDA 模型对于文本建模，并把每个段落表示为主题分布后进行分类。验证与分析分类结果。

2 实验原理

2.1 Topic Model

主题模型（Topic Model）是以非监督学习的方式对文档的隐含语义结构(latent semantic structure)进行聚类(clustering)的统计模型。

主题模型（Topic Model）是一种常用的文本挖掘工具，用于发现文本主体中的隐藏语义结构。每个文档都应该对应着一个或多个的主题（topic），而每个主题都会有对应的词分布，通过主题，就可以得到每个文档的词分布。潜在 Dirichlet 分布（Latent Dirichlet Allocation, LDA）是 Topic model 的一种，用于将文档中的文本分类为特定的主题。

2.2 LDA 模型

LDA 是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是指我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

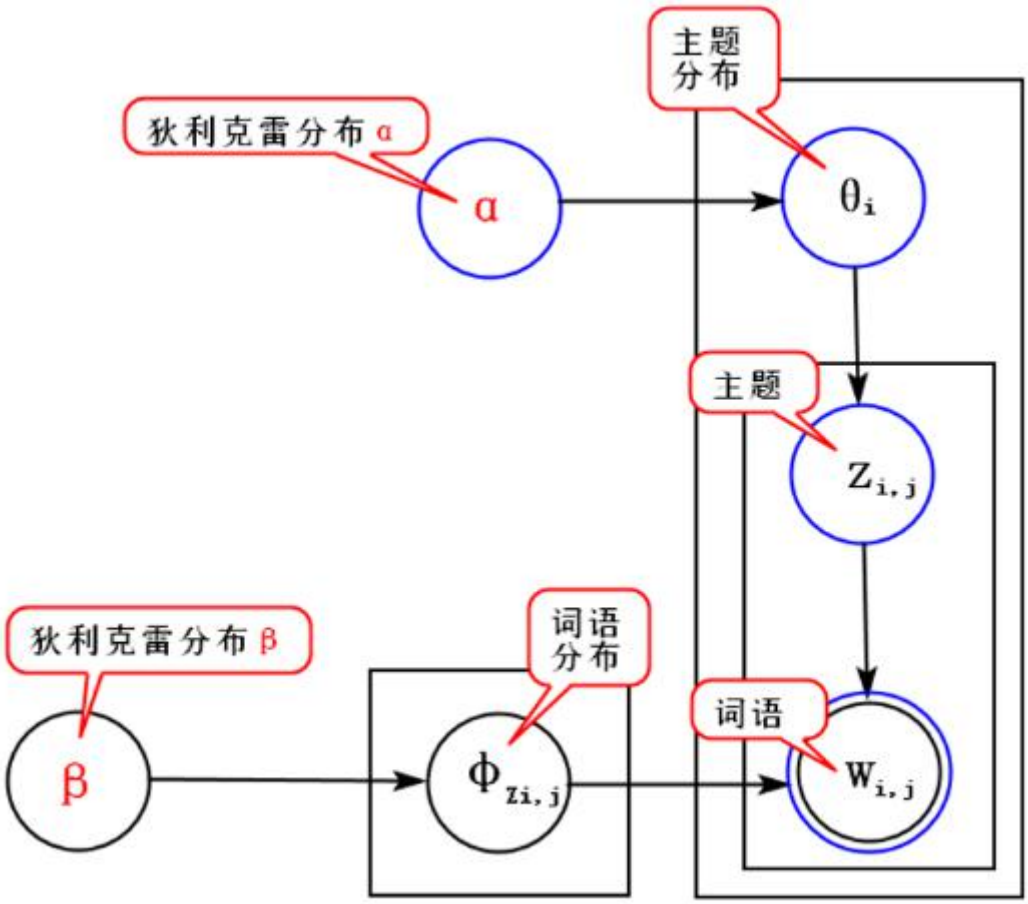
LDA 是一种非监督机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。它采用了词袋的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

利用 LDA 模型生成一篇文档的方式：

- 按照先验概率 $P(d_i)$选择一篇文档d_i。
- 从狄利克雷分布（即 Dirichlet 分布）α中取样生成文档d_i的主题分布θ_i，换言之，主题分布θ_i由超参数为α的 Dirichlet 分布生成。
- 从狄利克雷分布（即 Dirichlet 分布）β中取样生成主题$z_{i,j}$对应的词语分布$\varphi_{i,j}$，换言

之，词语分布$\varphi_{i,j}$由超参数为β的 Dirichlet 分布生成。

- 从词语的多项式分布$\varphi_{i,j}$中采样最终生成词语$\omega_{i,j}$。



3 实验设计

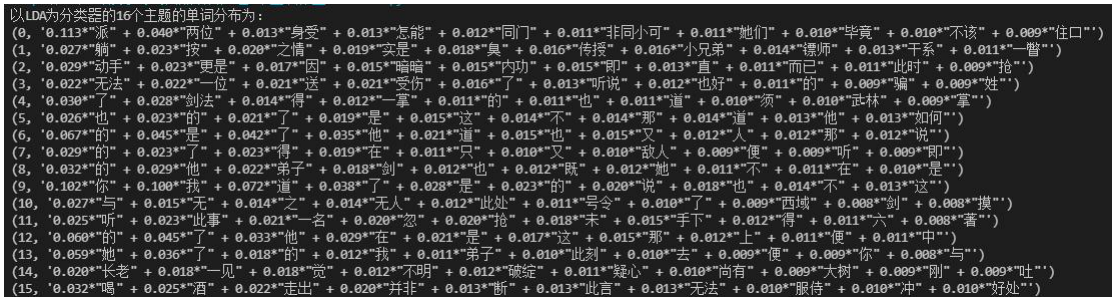
本次实验以第一次作业所提供的 16 本金庸武侠小说作为数据集，利用 LDA 进行文本分类。

1、数据预处理：与第一次作业相同，删去所有的隐藏符号，删除所有的非中文字符，不考虑上下文关系的前提下删去所有标点符号。以 jieba 库对中文语料进行分词。得到训练集和测试集。

2、运用 LDA 进行训练并验证：使用 gensim 中的 corpora 和其中自带的 LDA 模型来进行训练。

4 结果分析与总结

1、以 LDA 为分类器的 16 个主题的单词分布为：



(0, '0.113*"派" + 0.040*"两位" + 0.013*"身受" + 0.013*"怎能" + 0.012*"同门" + 0.011*"非同小可" + 0.011*"她们" + 0.010*"毕竟" + 0.010*"不该" + 0.009*"住口"')

(1, '0.027*"躺" + 0.023*"按" + 0.020*"之情" + 0.019*"实是" + 0.018*"臭" + 0.016*"传授" + 0.016*"小兄弟" + 0.014*"镖师" + 0.013*"干系" + 0.011*"一瞥"')

(2, '0.029*"动手" + 0.023*"更是" + 0.017*"因" + 0.015*"暗暗" + 0.015*"内功" + 0.015*"即" + 0.013*"直" + 0.011*"而已" + 0.011*"此时" + 0.009*"抢"')

(3, '0.022*"无法" + 0.022*"一位" + 0.021*"送" + 0.021*"受伤" + 0.016*"了" + 0.013*"听说" + 0.012*"也好" + 0.011*"的" + 0.009*"骗" + 0.009*"姓"')

(4, '0.030*"了" + 0.028*"剑法" + 0.014*"得" + 0.012*"一掌" + 0.011*"的" + 0.011*"也" + 0.011*"道" + 0.010*"须" + 0.010*"武林" + 0.009*"掌"')

(5, '0.026*"也" + 0.023*"的" + 0.021*"了" + 0.019*"是" + 0.015*"这" + 0.014*"不" + 0.014*"那" + 0.014*"道" + 0.013*"他" + 0.013*"如何"')

(6, '0.067*"的" + 0.045*"是" + 0.042*"了" + 0.035*"他" + 0.021*"道" + 0.015*"也" + 0.015*"又" + 0.012*"人" + 0.012*"那" + 0.012*"说"')

(7, '0.029*"的" + 0.023*"了" + 0.023*"得" + 0.019*"在" + 0.011*"只" + 0.010*"又" + 0.010*"敌人" + 0.009*"便" + 0.009*"听" + 0.009*"即"')

(8, '0.032*"的" + 0.029*"他" + 0.022*"弟子" + 0.018*"剑" + 0.012*"也" + 0.012*"既" + 0.012*"她" + 0.011*"不" + 0.011*"在" + 0.010*"是"')

(9, '0.102*"你" + 0.100*"我" + 0.072*"道" + 0.038*"了" + 0.028*"是" + 0.023*"的" + 0.020*"说" + 0.018*"也" + 0.014*"不" + 0.013*"这"')

(10, '0.027*"与" + 0.015*"无" + 0.014*"之" + 0.014*"无人" + 0.012*"此处" + 0.011*"号令" + 0.010*"了" + 0.009*"西域" + 0.008*"剑" + 0.008*"摸"')

(11, '0.025*"听" + 0.023*"此事" + 0.021*"一名" + 0.020*"忽" + 0.020*"抢" + 0.018*"未" + 0.015*"手下" + 0.012*"得" + 0.011*"六" + 0.008*"著"')

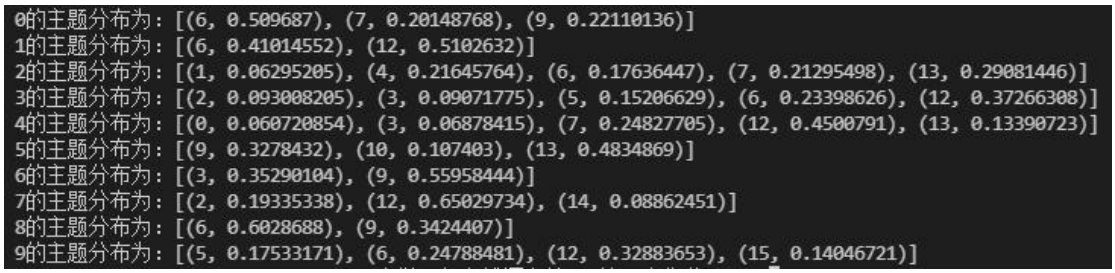
(12, '0.060*"的" + 0.045*"了" + 0.033*"他" + 0.029*"在" + 0.021*"是" + 0.017*"这" + 0.015*"那" + 0.012*"上" + 0.011*"便" + 0.011*"中"')

(13, '0.059*"她" + 0.036*"了" + 0.018*"的" + 0.012*"我" + 0.011*"弟子" + 0.010*"此刻" + 0.010*"去" + 0.009*"便" + 0.009*"你" + 0.008*"与"')

(14, '0.020*"长老" + 0.018*"一见" + 0.018*"觉" + 0.012*"不明" + 0.012*"破绽" + 0.011*"疑心" + 0.010*"尚有" + 0.009*"大树" + 0.009*"刚" + 0.009*"吐"')

(15, '0.032*"喝" + 0.025*"酒" + 0.022*"走出" + 0.020*"并非" + 0.013*"断" + 0.013*"此言" + 0.013*"无法" + 0.010*"服侍" + 0.010*"冲" + 0.010*"好处"')

2、随机选取一些段落，将其文本进行预处理后作为测试集，测试 LDA 模型对于文本的分类效果，得到不同测试段落的主题分布。本次实验选取 10 个段落进行测试，通过最后每个 topic 的概率确定段落的主题分布。



0 的主题分布为：[(6, 0.509687), (7, 0.20148768), (9, 0.22110136)]，属于第 1 个主题。

1 的主题分布为：[(6, 0.41014552), (12, 0.5102632)]，属于第 12 个主题。

2 的主题分布为：[(1, 0.06295205), (4, 0.21645764), (6, 0.17636447), (7, 0.21295498), (13, 0.29081446)]，属于第 13 个主题。

3 的主题分布为：[(2, 0.093008205), (3, 0.09071775), (5, 0.15206629), (6, 0.23398626), (12, 0.37266308)]，属于第 12 个主题。

4 的主题分布为：[(0, 0.060720854), (3, 0.06878415), (7, 0.24827705), (12, 0.4500791), (13, 0.13390723)]，属于第 12 个主题。

5 的主题分布为：[(9, 0.3278432), (10, 0.107403), (13, 0.4834869)]，属于第 13 个主题。

6 的主题分布为：[(3, 0.35290104), (9, 0.55958444)]，属于第 9 个主题。

7 的主题分布为：[(2, 0.19335338), (12, 0.65029734), (14, 0.08862451)]，属于第 12 个主题。

8 的主题分布为：[(6, 0.6028688), (9, 0.3424407)]，属于第 6 个主题。

9 的主题分布为：[(5, 0.17533171), (6, 0.24788481), (12, 0.32883653), (15, 0.14046721)]，属于第 12 个主题。

3、由实验结果可知，分类效果不是特别好，通过 16 个主题的单词分布可以看出，这些单词都是一些十分常见的单词，并不具有特殊性，而在此基础上还能有不错的效果也证实了 LDA 的建模有效性。

5 代码链接

见 https://github.com/ErrricCai/DL-NLP/tree/main/HW3

参考：

https://blog.csdn.net/weixin_42663984/article/details/116264233

https://blog.csdn.net/shzx_55733/article/details/116280982?spm=1001.2014.3001.5502

https://www.cnblogs.com/Luv-GEM/p/10881838.html