北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理
第三次大作业

基于 LDA 的中文文本分类

院（系）名称	自动化科学与电气工程学院
学 生 学 号	ZY2103810
学 生 姓 名	游虎杰

2022 年 5 月

一、问题描述

从每篇金庸小说中均匀抽取 200 个段落，利用 LDA 模型表示每篇小说的主题概率分布，并从每篇小说中随机抽取 60 个段落作为测试集，计算测试集中每个段落的主题分布，并依据测试集中段落主题分布与每篇小说主题分布的误差平方和，选取误差平方和最小的一篇小说作为该段落所属小说的判别结果。由于金庸小说文本存在分段错误，导致部分小说每个段落的字数非常少（如《白马啸西风》），因此限定每个段落字数不少于十个

二、问题表达

LDA（Latent Dirichlet Allocation）模型由 Blei, David M.、Ng, Andrew Y.、Jordan 于 2003 年提出，用来推测文档的主题分布。它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题分布后，便可以根据主题分布进行主题聚类或文本分类。LDA 模型是话题模型(topic model)的典型代表，通过概率模型建立了从主题到文档中每个词的关系，从利用贝叶斯概率，能通过文档的词推导出文档的主题。

本次实验以第一次实验所提供的金庸先生的 16 本武侠小说作为数据集，利用 LDA 进行文本分类，小说如下：

1、白马啸西风

2、碧血剑

3、飞狐外传

4、连城诀

5、鹿鼎记

6、三十三剑客图

7、射雕英雄传

8、神雕侠侣

9、书剑恩仇录

10、天龙八部

11、侠客行

12、笑傲江湖

13、雪山飞狐

14、倚天屠龙记

15、鸳鸯刀

16、越女剑

三、具体算法实现

1. LDA 模型训练

选取主题数为 35，从每本小说中分别抽取 500 个段落，将每本小说的段落分别合并为一个大段落，计算每本小说对应大段落的主题概率分布，以此作为该小说的主题概率分布。

2. 模型测试验证

随机从每本小说中抽取 60 个段落作为测试集，计算每个段落的主题概率分布，并计算该段落主题概率分布与每本小说主题概率分布的误差平方和，选取误差平方和最小的一本小说，作为该段落的来源判别结果，并计算总的判别准确率。

四、运行结果

每本小说的主题概率分布如图 1 所示：

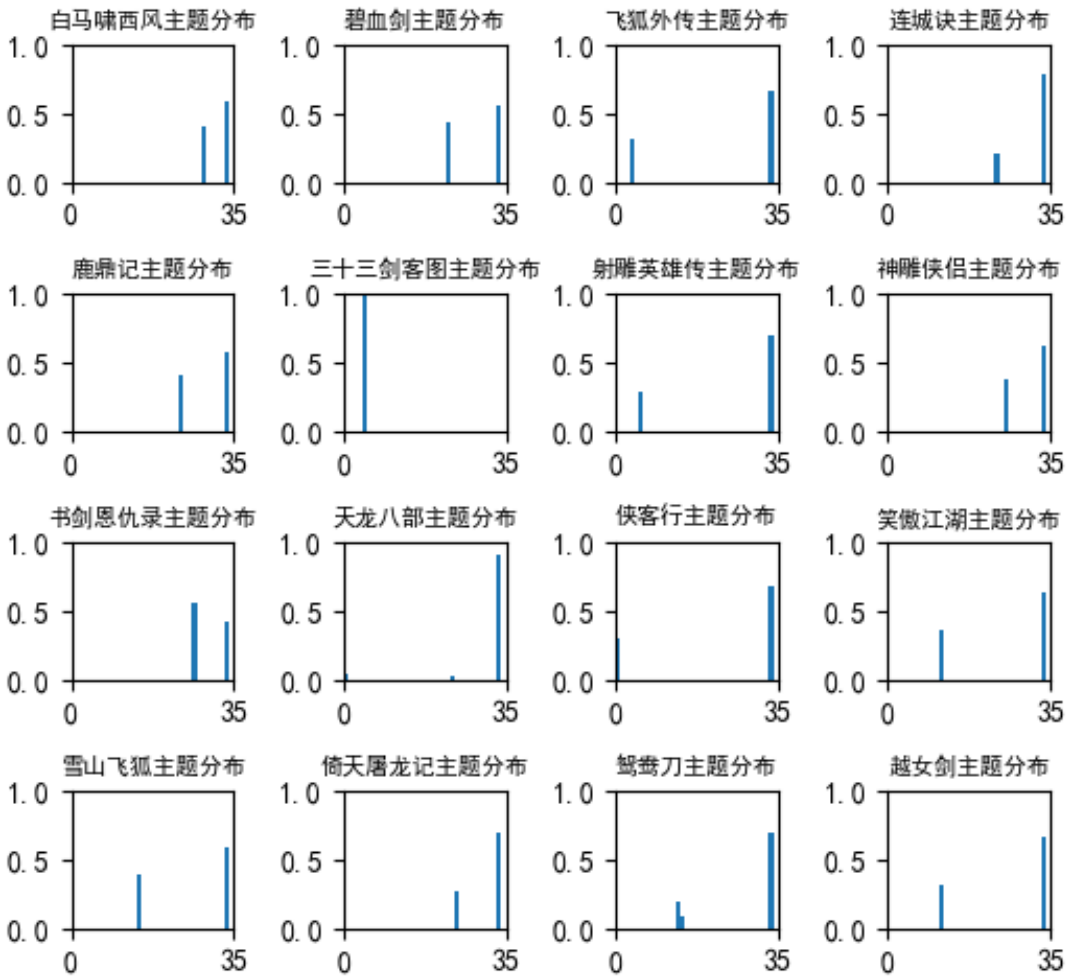

图 1 各小说主题概率分布

测试集的测试准确率为 0.4125

由实验结果可知，对于某一本小说，其主题概率分布主要集中于两到三个主题，一般情况下，不同小说间的主题概率分布情况相差较大，因此主题分布可以作为不同小说之间一项有辨识度的特征。然而，验证集的分类准确率并不高，通过分析总结，原因如下：首先，训练集与测试集之间存在差异；其次，LDA 模型训练过程可能存在过拟合；再者，少部分小说的主题概率分布非常接近；最后，部分段落的分词可能比较普通，难以反应其主要特征。

五、个人总结和体会

通过这次作业，我对 LDA 模型的理解有了进一步的加深，对文本的 LDA 建模过程有了更为熟练的掌握。同时，在编写代码的过程中，我对 python 的应用水平有了进一步

提升。

六、作业代码

https://github.com/youlll/DP_NLP3.git