北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理

学 院 名 称	自动化科学与电气工程学院
专 业 名 称	自动化
学 生 姓 名	张一凡
指 导 教 师	秦曾昌

2023 年 4 月

深度学习与自然语言处理第三次大作业-LDA 算法

一、 实验内容

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。

二、 实验数据

金庸小说集

三、 实验算法设计

LDA 算法原理：

LDA（Linear Discriminant Analysis）是一种非监督机器学习技术，主要用于文本主题建模。它通过将文档中的每个词视为从主题中选择的概率来识别文档集中的潜在主题。LDA 采用了词袋的方法，将文档表示为词频向量，转化为易于建模的数字信息。LDA 的核心思想是寻找最佳的投影方法，将高维样本投影到特征空间，使不同类别间的数据“距离”最大，同一类别内的数据“距离”最小。

四、 实验结果及分析

要求从语料库中均匀抽取 200 个段落作为样本，每个段落包含的词数不少于 500 个，并将每个段落的标签设为所属的小说名称。语料库共有 16 篇小说，每篇小说选取 13 个区间，每个区间包含的词数相等。因此，每篇小说共有 13 个段落可供选择，每个段落的词数为每个区间的前 500 个词。总共共有 208 个段落可供选择，需要从中均匀抽取 200 个段落作为样本。

首先进行初始化，每篇文章的每个词语随机赋予一个初始的 Topic 值，然后分别统计每篇文章的总词数、每篇文章的词频、每个 Topic 的总词数、每个 Topic 的词频；
再计算每个 topic 被选中的概率，然后进行迭代，训练模型。

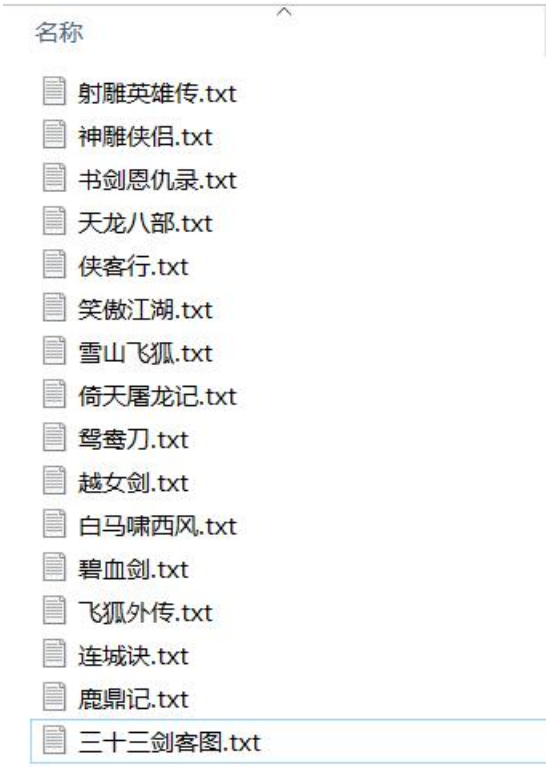
仍然在对应的 16 部小说中选择段落作为测试集，在之前训练集中选取的是 208 个段落中的第 0-500 个单词，因此在测试集中选取第 501-1000 单词形成测试段落，最终形成待分类文章。

在对 15 篇待分类文章进行预处理后，每个词被赋予一个随机的初始 Topic。每篇文章的总词数和词频被记录下来。模型训练已经得到了每个 Topic 的总词数和词频，不再需要统计，作为已知量用于测试数据。接下来，需要使用欧式距离的方式区分每篇待分类文章来自哪一本小说，即比较待分类文章与已知小说对于各个 Topic 的概率向量之间的距离，找出距离最近的小说标签。

实验结果如下：

```
['射雕英雄传.txt', '神雕侠侣.txt', '书剑恩仇录.txt', '天龙八部.txt', '侠客行.txt', '笑傲江湖.txt', '雪山飞狐.txt',
 '倚天屠龙记.txt', '鸳鸯刀.txt', '越女剑.txt', '白马啸西风.txt', '碧血剑.txt', '飞狐外传.txt', '连城诀.txt', '鹿
鼎记.txt', '三十三剑客图.txt']
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
```

与我们在金庸小说集的文件夹中存储这 16 本小说的顺序一致，说明我们分类正确。



五、实验总结分析

本次实验采用 LDA 模型对金庸小说集的主题进行分类，取得了良好的效果。本次任务进一步增强了我对自然语言处理的认识。然而，在实践过程中，发现代码运行时间较长。为了改进这一点，可以考虑对 Topic 进行简化，只选取关键词进行处理，这将显著降低运行时间。