成 绩 _____

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理第 3 次作业

LDA 主题模型进行文本分类

院（系）名称	自动化科学与电气工程学院
专 业 名 称	自动化
学 号	ZY2103809
姓 名	王海腾
指 导 教 师	秦曾昌

2022 年 5 月 6 日

一、任务描述

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。

二、实验原理

1. LDA(Latent Dirichlet Allocation)模型

LDA(Latent Dirichlet Allocation)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，我们认为一篇文章的每一个词都是通过“文章以一定的概率选择了某一主题，并从这个主题中以一定的概率选择某一词语” 这个过程得到。LDA 模型假设生成某个文档的过程如下：

1）按照先验概率$p(d_i)$选择一篇文档d_i；

2）从超参数为α的 Dirichlet 分布中取样生成文档d_i的主题分布θ_i；

3）从主题θ_i的多项式分布中取样生成文档d_i的第j个词的主题$z_{i,j}$；

4）从超参数为β的 Dirichlet 分布中取样生成主题$z_{i,j}$对应的词语分布$\phi_{z_{i,j}}$；

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

2. 利用 LDA 主题模型进行文本分类

本文采用以下步骤/思路对金庸的小说集进行文本分类：

1）从给定的 16 本金庸小说数据集中，随机、均匀地抽取k个段落，每个段落的标签为对应小说的小说名，对上述语料库进行 jieba 分词，并添加通用停用词和金庸小说专用停用词每个段落包含n个词（$n \geq 500$），每个段落作为一个样本；

2）指定主题数为d，利用上述k_1个训练样本训练 LDA 模型；

3）利用训练好的 LDA 模型得到上述k_1个训练样本的主题分布。由于主题数为d，因此每个训练样本得到的主题分布为一个$1 \times d$的向量；所有训练样本的主题分布则为一个$k_1 \times d$的特征向量；

4）利用上述训练样本的$k_1 \times d$的特征向量以及对应的k_1个标签训练一个线性 SVM 分类器；

5）上述训练样本的$k_1 \times d$的特征向量通过训练好的 SVM 分类器，得到训练样本的预测标签，与真实的标签进行比较，计算训练样本文本分类准确率；

6）利用训练好的 LDA 模型得到k_2个测试样本的主题分布。同理，由于主题数为d，因此每个测试样本得到的主题分布为一个$1 \times d$的向量；所有测试样本的主题分布则为一个$k_2 \times d$的特征向量；该特征向量通过训练好的 SVM 分类器，得到测试样本的预测标签，与真实的标签进行比较，计算测试样本文本分类准确率。

三、实验结果

本次实验测试了不同的段落（文档）数、每个段落的字数、不同主题数对文本分类准确率的影响。

（1）实验结果如下表所示。

实验序号	主题数	段落（文档）数	每段话字数	训练集文本分类准确率（%）	测试集文本分类准确率（%）
1	20	200	500	30.72	5.128
2	50	200	500	42.48	5.128
3	100	200	500	56.21	12.8
4	200	200	500	70.59	17.95
5	500	200	500	80.39	2.601
6	50	1000	500	38.21	26.13
7	50	200	5000	40.52	5.126
8	200	1000	500	55.36	32.66

从实验结果可以看出：

（1）对比 1、2、3、4 的实验结果，可以看出，增加 LDA 模型的主题数，可以有助于增加训练集和测试集文本分类准确率

（2）对比 4 和 5 的结果，可以看出，LDA 模型的主题数增加过多时，虽然训练集文本分类准确率提高，但是测试集文本分类准确率降低，这可能是由于过拟合导致。

（3）对比 2 和 6 的结果，增加抽取的段落（文档）数，可以显著提高测试集文本分类准确率，但训练集文本分类准确率有所降低，可能是因为训练样本（段落数）太少的时候，训练集上容易引起过拟合导致；

（4）对比 3 和 5，或 4 和 6 的结果，可以看出，增加抽取段落的每段话字数，可以显著增加训练集和测试集上的文本分类准确率；

（2）主题个数为 10 时的主题分布结果图

```
[(0,
  '0.003*"心中" + 0.002*"众人" + 0.002*"师父" + 0.002*"武功" + 0.002*"韦小宝" + '
  '0.002*"心想" + 0.002*"弟子" + 0.002*"李文秀" + 0.002*"陈家洛" + 0.002*"之中"'),
 (1,
  '0.003*"心中" + 0.003*"武功" + 0.002*"之中" + 0.002*"李文秀" + 0.002*"长剑" + '
  '0.002*"张无忌" + 0.002*"众人" + 0.002*"心想" + 0.001*"二人" + 0.001*"两个"'),
 (2,
  '0.003*"李文秀" + 0.003*"心中" + 0.003*"武功" + 0.002*"师父" + 0.002*"心想" + '
  '0.002*"之中" + 0.002*"袁承志" + 0.002*"苏普" + 0.002*"陈达海" + 0.002*"少女"'),
 (3,
  '0.003*"心中" + 0.003*"师父" + 0.002*"韦小宝" + 0.002*"袁承志" + 0.002*"二人" + '
  '0.002*"武功" + 0.002*"剑士" + 0.002*"范蠡" + 0.002*"之中" + 0.002*"胡斐"'),
 (4,
  '0.004*"范蠡" + 0.003*"心中" + 0.003*"武功" + 0.002*"师父" + 0.002*"心想" + 0.002*"之中" '
  '+ 0.002*"剑士" + 0.002*"阿青" + 0.002*"众人" + 0.002*"二人"'),
 (5,
  '0.003*"剑士" + 0.002*"心中" + 0.002*"范蠡" + 0.002*"李文秀" + 0.002*"二人" + '
  '0.002*"之中" + 0.002*"袁承志" + 0.002*"武功" + 0.002*"令狐冲" + 0.002*"少女"'),
 (6,
  '0.002*"师父" + 0.002*"武功" + 0.002*"李文秀" + 0.002*"心中" + 0.002*"心想" + '
  '0.002*"脸上" + 0.002*"之中" + 0.002*"众人" + 0.002*"手中" + 0.001*"眼见"'),
 (7,
  '0.004*"韦小宝" + 0.003*"武功" + 0.003*"袁承志" + 0.002*"之中" + 0.002*"心想" + '
  '0.002*"众人" + 0.002*"心中" + 0.002*"陈家洛" + 0.001*"小人" + 0.001*"勾践"'),
...
  '0.002*"众人" + 0.002*"爹爹" + 0.002*"武功" + 0.002*"师父" + 0.002*"二人"'),
 (9,
  '0.005*"韦小宝" + 0.003*"心中" + 0.003*"武功" + 0.002*"心想" + 0.002*"二人" + '
  '0.002*"师父" + 0.002*"姑娘" + 0.002*"众人" + 0.002*"之中" + 0.002*"胡斐"')]
```