
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 | SY2103106 |
| 姓 名 | 段晓玥 |
| 指 导 教 师 | 秦曾昌 |


2022 年 5 月 3 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 主题模型

LDA 是一种较流行的主题模型，它可以将文档集中每篇文章的主题以概率分布的形式给出，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$\alpha$的 Dirichlet 分布中取样生成文档$d_i$的主题分布$\theta_i$；

3）从主题$\theta_i$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$\beta$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，每个段落包含$n$个字（$n \geq 500$），每个段落作为一个样本；

2）将随机抽取的$k$个段落（即$k$个样本）中的 80%作为训练样本，剩余 20%作为测试样本。训练样本数为$k_1 = 80\%k$，测试样本数为$k_2 = 20\%k$；

3）指定主题数为$d$，利用上述$k_1$个训练样本训练 LDA 模型；

4）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；

5）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；

6）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；

7）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。

其中，上述步骤（1）~（3）为数据准备、预处理和训练 LDA 模型；步骤（4）（5）为训练线性 SVM 分类器；步骤（6）（7）为计算训练和测试样本的文本分类准确率。

## 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。此外，还考察了是否去除停用词对分类准确率的影响。没有什么实际含义的功能词，或用十分广泛但对这样的词搜索引擎无法保证能够给出真正相关的搜索结果、难以帮助缩小搜索范围的词。实验中停用词表由百度、哈工大等创造的停用词表给出。去除停用词有助于数据清洗得更干净、获得的文本更具有实际含义。

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 | 是否去除停用词 | 训练集文本分类准确率（%） | 测试集文本分类准确率（%） |
|---|---|---|---|---|---|---|
| 1 | 50 | 200 | 500 | 否 | 34.38 | 2.50 |
| 2 | 50 | 200 | 500 | 是 | 41.25 | 12.50 |
| 3 | 50 | 1000 | 500 | 否 | 23.25 | 15.50 |
| 4 | 50 | 1000 | 500 | 是 | 30.63 | 25.00 |
| 5 | 50 | 1000 | 5000 | 否 | 44.00 | 46.00 |
| 6 | 50 | 1000 | 5000 | 是 | 52.75 | 55.00 |
| 7 | 20 | 1000 | 5000 | 是 | 49.75 | 51.00 |
| 8 | 100 | 1000 | 5000 | 是 | 62.75 | 64.00 |

从实验结果可以看出：

① 对比 1 和 2，或 3 和 4，或 5 和 6 的结果，可以看出，去除无意义的停用词，可以增强样本中文本的实际含义，显著增强训练集和测试集的文本分类正确率；

② 对比 1 和 3，或 2 和 4 的结果，可以看出，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；

③ 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；

④ 对比 6、7、8 的结果，可以看出，LDA 建模时选取合适的主题数对有效提取特征具有重要的影响。在一定范围内，选取的主题数较多，可以建模更为精细的特征，提高训练集和测试集的文本分类准确率。