

成 绩

****

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

1. **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。

1. **实验原理**
2. **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. 按照先验概率选择一篇文档；
2. 从超参数为的Dirichlet分布中取样生成文档的主题分布；
3. 从主题的多项式分布中取样生成文档的第个词的主题；
4. 从超参数为的Dirichlet分布中取样生成主题对应的词语分布；
5. 从词语的多项式分布中采样最终生成词语。
6. **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个词（），每个段落作为一个样本；
2. 指定主题数为，利用上述个训练样本训练LDA模型；
3. 利用训练好的LDA模型得到上述个训练样本的主题分布。由于主题数为，因此每个训练样本得到的主题分布为一个的向量；所有训练样本的主题分布则为一个的特征向量；
4. 利用上述训练样本的的特征向量以及对应的个标签训练一个线性SVM分类器；
5. 上述训练样本的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模型得到个测试样本的主题分布。同理，由于主题数为，因此每个测试样本得到的主题分布为一个的向量；所有测试样本的主题分布则为一个的特征向量；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。
7. **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

