**NLP作业2：LDA文本建模分类**

顾韬

1312855584@qq.com

**Abstract**

LDA(Latent Dirichlet Allocation)潜在狄利克雷分布。LDA主题模型是一种文档生成模型，是一种非监督机器学习技术。对于一篇文档，LDA认为其有多个主题的，而每个主题又对应着不同的词。一篇文档的构造过程，首先是以一定的概率选择某个主题，然后再在这个主题下以一定的概率选出某一个词，这样就生成了这篇文档的第一个词。不断重复这个过程，就生成了整篇文章。LDA的使用是上述文档生成过程的逆过程，即根据一篇得到的文档，去寻找出这篇文档的主题，以及这些主题所对应的词。

**Introduction**

本次实验使用LDA模型对语料库进行文本建模分类，分类结果使用10次交叉验证。并且验证分析了以下三个问题：1）分类性能在不同主题数的情况的区别；2）以“词”和“字”为基本单元的分类效果的区别；3）不同tokens的文本对主题模型性能的影响。

**Principle**

在LDA模型中，一篇文档生成的方式如下：

1）从狄利克雷分布alpha中取样生成文档 的主题分布；

2）从主题的多项式分布中取样生成文档第 个词的主题；

3）从狄利克雷分布beta中取样生成主题对应的词语分布；

4）从词语的多项式分布中采样最终生成词语；

其中，类似Beta分布是二项式分布的共轭先验概率分布，而狄利克雷分布（Dirichlet分布）是多项式分布的共轭先验概率分布。

**Experiments**

**实验part1：语料库预处理**

语料库预处理：去除停词广告词标点符号等，相对第一个实验，优化了无意义词语的处理。

完成处理后读取并保存对应的数据，以字典形式保存word各char对应的数据。分别存储在word\_dict.pkl与char\_dict.pkl中。由于部分文档存在着不以’\n’结尾的情况，即存在每段的字符数均较少，所以分词与分字读取所有文件数据。

**实验part2：主实验**

为了均匀截取1000段数据，计算每部小说所需数据量（本次实验平均采样），不足1000段则随机抽取。建立书名与id的字典，将数据标签转为数字id。使用train\_test\_split进行数据分割，构建词频矩阵。使用如下代码进行文本建模models.ldamodel.LdaModel进行文本建模，获取训练过的lda模型对应词频，分类器使用支持向量机。得到分类准确率。

实验不同主题数，不同长度文本对分类性能的影响，实验结果如下表：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Topic | 200 | 200 | 200 | 200 | 500 | 200 | 500 |
| Tokens | 20 | 80 | 100 | 120 | 250 | 500 | 500 |
| Word Accuracy | 0.275 | 0.397 | 0.197 | 0.418 | 0.727 | 0.253 | 0.723 |
| Char Accuracy | 0.506 | 0.637 | 0.427 | 0.668 | 0.812 | 0.338 | 0.849 |
| Topic | 1000 | 1000 | 1000 | 1000 |
| Tokens | 400 | 500 | 600 | 1000 |
| Word Accuracy | 0.853 | 0.574 | 0.892 | 0.856 |
| Char Accuracy | 0.899 | 0.798 | 0.923 | 0.909 |

表1 不同token、topic数对分类结果的影响。

由上表可知：

1. 短文本的效果比长文本的分类效果要差；这是可以理解的，短文本出现的更随机，更难拟合。
2. 偶然的实验发现，对于不同长度的文本，将Topic的数量控制在约0.5\*Topic数似乎相对其他情况有更高的准确率，经过多次分类后基本证明这一点。但对不同文本长度的情况，最优Topic/Token的值仍然需要进一步研究
3. 从总体的分类结果来看，字的分类效果要优于词的分类效果。原因应该是字的格式更加规则统一。

**Conclusions**

在本次实验中，成功实现了LDA文本建模与分类任务，最高准确率达到92.3%。探究了不同Token与Topic数对模型文本分类性能的影响，加深了对于NLP任务的理解，拓宽在人工智能领域的知识。