Report of Deep Learning for Natural Langauge Processing\_2

ZY2303109 金子棋

**Abstract**

基于LDA主题模型对中文语料库进行主题提取，并将主题作为训练集，对应的小说名称作为标签进行分类训练，最后采用交叉验证方式验证训练准确率。

**Introduction**

Latent Dirichlet Allocation（LDA）是一种概率生成模型，常用于文本数据分析中的主题建模。LDA假设文档是由多个主题混合组成的，每个主题又由一组单词的分布表示。通过推断文档中的主题分布和单词所属主题，LDA能够自动发现文本数据中的潜在主题，帮助理解文本结构和内容，适用于文档分类、信息检索等任务，尽管在处理短文本或噪声较多的情况下可能存在一定局限性。本文将采用LDA主题模型对中文语料库进行主题提取，并将主题作为训练集，对应的小说名称作为标签进行分类训练，最后采用交叉验证方式验证训练准确率，以回答以下问题：（1）在设定不同的主题个数T的情况下，分类性能是否有变化？；（2）以"词"和以"字"为基本单元下分类结果有什么差异？（3）不同的取值的K的短文本和长文本，主题模型性能上是否有差异?

**Methodology**

1. 文本预处理

本次使用的中文语料库为金庸小说集，有效正文文本文件共16个，此外还有包含所有文件的zip包、文本文件来源网站url链接以及包含所有章节标题的文本文件inf.txt，直接在文件夹中删去。

对于正文文本，一需要删除开头结尾网站附上的广告信息，二需要去除文本中非中文字符、空格换行等无效信息，三需要去除中文文本中的标点符号，最后仅剩中文文字作为后续处理的信息源。

1. 段落提取与主题分类训练估计

得到预处理后的分字/分词文本后，均匀地从中抽取1000个段落作为数据集。首先统计共有16篇文本文档，故平均从每个文本文档中抽取63个段落(这样最后一篇文本文档是抽55个段落)。由此，将每个文本文档预处理后的分字/分词文本平均分为63份，并从每份的开头开始向后取token\_num个字/词作为一个段落，加入训练集。将对应的小说名称加入标签。得到了训练集以后利用sklearn库中的fit\_transform()方法得到lda主题模型，并将900个段落用作训练，100个段落用作测试，采用SVM分类器训练后进行10次交叉验证得到准确率。以下是实验结果：

|  |  |  |  |
| --- | --- | --- | --- |
| 主题数量  段落大小 | 50 | 100 | 200 |
| 20 | 7.11% | 6.78% | 8.56% |
| 100 | 11.44% | 17.68% | 14.57% |
| 500 | 25.44% | 35.00% | 17.89% |
| 1000 | 34.56% | 39.89% | 46.78% |
| 3000 | 72.33% | 87.67% | 84.56% |

表1 按词分解的LDA分类训练准确率

|  |  |  |  |
| --- | --- | --- | --- |
| 主题数量  段落大小 | 50 | 100 | 200 |
| 20 | 4.00% | 4.00% | 10.00% |
| 100 | 14.00% | 16.00% | 19.00% |
| 500 | 20.00% | 24.00% | 15.00% |
| 1000 | 25.00% | 35.00% | 46.00% |
| 3000 | 63.00% | 93.00% | 87.00% |

表2 按词分解的LDA分类测试准确率

|  |  |  |  |
| --- | --- | --- | --- |
| 主题数量  段落大小 | 50 | 100 | 200 |
| 20 | 8.67% | 8.89% | 7.78% |
| 100 | 10.89% | 10.44% | 11.46% |
| 500 | 15.79% | 15.45% | 18.00% |
| 1000 | 14.22% | 13.11% | 29.86% |
| 3000 | 36.44% | 66.22% | 72.22% |

表3 按字分解的LDA分类训练准确率

|  |  |  |  |
| --- | --- | --- | --- |
| 主题数量  段落大小 | 50 | 100 | 200 |
| 20 | 7.00% | 12.00% | 9.00% |
| 100 | 11.00% | 11.00% | 14.00% |
| 500 | 12.00% | 12.00% | 13.00% |
| 1000 | 11.00% | 10.00% | 30.00% |
| 3000 | 32.00% | 69.00% | 72.00% |

表4 按字分解的LDA分类测试准确率

**Conclusion**

1. 随着主题数量的提升，在按字分解的情况下训练/测试准确率呈现出与主题数量的正比例关系，而在按词分解的情况下主题数量100个与200个的情况基本相近，主题数量为200个时还出现了准确率相对下降的情况，对比于按字分解的情况，可以看到主题数量并不是越多越好，而是会在一个值后出现饱和甚至下降的情况。
2. 对比按字分解与按词分解的情况，可以看到按词分解的情况下训练/测试准确率都比按字分解的情况高，可见按词分解的性能会更加优秀，由此可见，LDA对短文本的主题分类效果比较差。