**Report of Deep Learning for Natural Langauge Processing 2**

Yongquan Chen

cyqss@buaa.edu.cn

**Abstract**

使用金庸语料库，通过LDA模型文本建模，从而构建文本分类器。以此探究在不同主题、不同基本单元（字、词）和不同token的分类性能。

**Introduction**

随着互联网中文文本数据的爆炸式增长，如何有效对中文文本进行分类成为一项重要的任务。在文本分类中，LDA模型作为一种无监督的主题模型，以其在主题建模方面的卓越性能引起了广泛关注。然而，中文文本的特点使得分类过程中面临着许多挑战：一是中文没有明显的分词界限，二是中文语料库通常呈现出多样性与稀疏性。

本研究旨在通过对以下三个方面进行探讨，为中文文本分类提供一些新的见解：

1.主题数量T对分类性能的影响：不同的主题数量可能会对分类性能产生不同影响。

2.分词单位的影响：以“字”或“词”为分词单位，对LDA模型的性能是否有显著影响。

3.文本长度K对性能的影响：不同的段落长度可能会对主题建模和分类产生不同效果。

4.不同分类器的性能差异：不同的分类器可能会对主题建模和分类产生不同效果。

**Methodology**

**Part 1: 数据准备**

从小说文本中抽取段落，构建用于分类的数据集。遍历小说目录，逐个读取所有小说的文本内容，保存在字典中，每个键对应一个小说文件名，值是该小说的全部文本内容。随后进行分段处理，生成具有不同特征长度的段落集合，针对每个小说文本，将其按设定长度（K）的词数分割成段落。每个段落被标记为对应小说的标签。

**Part 2: 文本预处理**

使用Jieba进行中文分词，将整个文本转化为一系列词汇。然后通过停用词列表，将常见的无意义词汇过滤掉，以保留对分类和主题建模有用的词汇。

**Part 3: LDA建模**

先构建词袋模型，将文本转换为词频矩阵。随后使用LDA模型，得到每个段落的主题分布（特征向量）。

**Part 4: 分类与验证**

利用不同分类器，根据段落的主题分布特征向量预测其所属的小说标签，并进行交叉验证。最后输出分类准确率。

**Experimental Studies**

**M1: 设定不同的主题个数T的情况下的分类性能变化**

设主题个数T分别为5,10, 20, 100, 500, 1000, 3000 进行实验，在设定文本长度K为1000的情况下使用随机森林分类器进行建模分析。得到分词情况下模型分类性能的情况。

|  |  |  |
| --- | --- | --- |
| 主题个数T | 分类准确率 | F1分数 |
| 5 | 0.5458 | 0.95 |
| 10 | 0.6917 | 0.95 |
| 20 | 0.7899 | 0.97 |
| 100 | 0.8927 | 0.99 |
| 500 | 0.8917 | 0.99 |
| 1000 | 0.8965 | 0.99 |
| 3000 | 0.8983 | 0.99 |

主题个数在100以上，分类性能提升缓慢。主题个数在100左右能达到性能和效率的平衡。

**M2: 以"词"和以"字"为基本单元,分类结果差异**

设主题个数T为20，设定文本长度K为1000的情况下进行实验，使用随机森林分类器进行建模分析。得到分词和分字情况下模型分类性能的情况。

|  |  |  |
| --- | --- | --- |
|  | 分类准确率 | F1分数 |
| 字 | 0.25 | 0.44 |
| 词 | 0.79 | 0.97 |

以"词"为基本单元，分类效果更好。

**M3: 不同的分割文本长度K下，模型性能的差异**

设设定文本长度K分别为20, 100, 500, 1000, 3000进行实验，在设定主题个数T为20的情况下使用随机森林分类器进行建模分析。得到分词情况下模型分类性能的情况。

|  |  |  |
| --- | --- | --- |
| 文本长度K | 分类准确率 | F1分数 |
| 20 | 0.3404 | 0.89 |
| 100 | 0.6234 | 0.95 |
| 500 | 0.7785 | 0.97 |
| 1000 | 0.7896 | 0.97 |
| 3000 | 0.8073 | 0.98 |

设定文本长度越大，分类性能最好。

**M4: 不同分类器的性能差异**

设主题个数T为20，设定文本长度K为1000的情况下进行实验，使用随机森林分类器进行建模分析。得到分词 情况下不同模型分类性能的情况。

|  |  |  |
| --- | --- | --- |
| 分类器 | 分类准确率 | F1分数 |
| 逻辑回归 | 0.75 | 0.76 |
| 朴素贝叶斯 | 0.70 | 0.71 |
| SVM | 0.76 | 0.78 |
| 随机森林 | 0.79 | 0.97 |

实验发现随机森林表现最好。