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

何明璞

1113490780@qq.com

**Abstract**

本研究基于潜在狄利克雷分配（LDA）模型，在给定的语料库上进行建模，探究了不同主题数量、段落长度和文本粒度单元（词/字）对小说文本分类性能的影响。本次实验从语料库中均匀抽取了1000个段落，进行了10次交叉验证，发现了以下现象：主题数量存在最优取值区间，以字为单元在长文本中表现更优，段落长度与分类准确率呈显著正相关。当段落长度达到3000词时，基于字符单元和50个主题的准确率最高，达到了76.1%。

**Introduction**

LDA模型（Latent Dirichlet Allocation，潜在狄利克雷分配）‌是一种用于文本挖掘和自然语言处理的概率生成模型，旨在从大量文档中自动发现隐藏的“主题”（即语义相关的词集合）。其核心思想是假设每篇文档由多个主题混合构成，而每个主题则表现为一组相关的词汇。因此对LDA模型来说，主题数量、文本长度、文本粒度单元的选取都能在不同程度上影响其分类性能，尤其对于中文文本来说，选取不同的文本粒度单元（字和词）对于分类的影响是极其显著的。本研究以武侠小说为研究对象，系统分析了主题数量、文本粒度单元（字和词）和段落长度对分类性能的影响。本研究旨在回答以下问题：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**

本研究主要包括三个部分，第一个部分为语料处理和采样，第二个部分为LDA模型搭建，第三个部分为性能评估。

在语料处理和采样的过程中，首先从指定的小说目录中读取文本，使用jieba分词分别根据文本粒度单元进行分词，在这个过程中去除停用词。采样时从语料库中均匀抽取了1000个段落，段落长度分别设置为20、100、500、1000和3000个token，每个段落的标签为小说名。

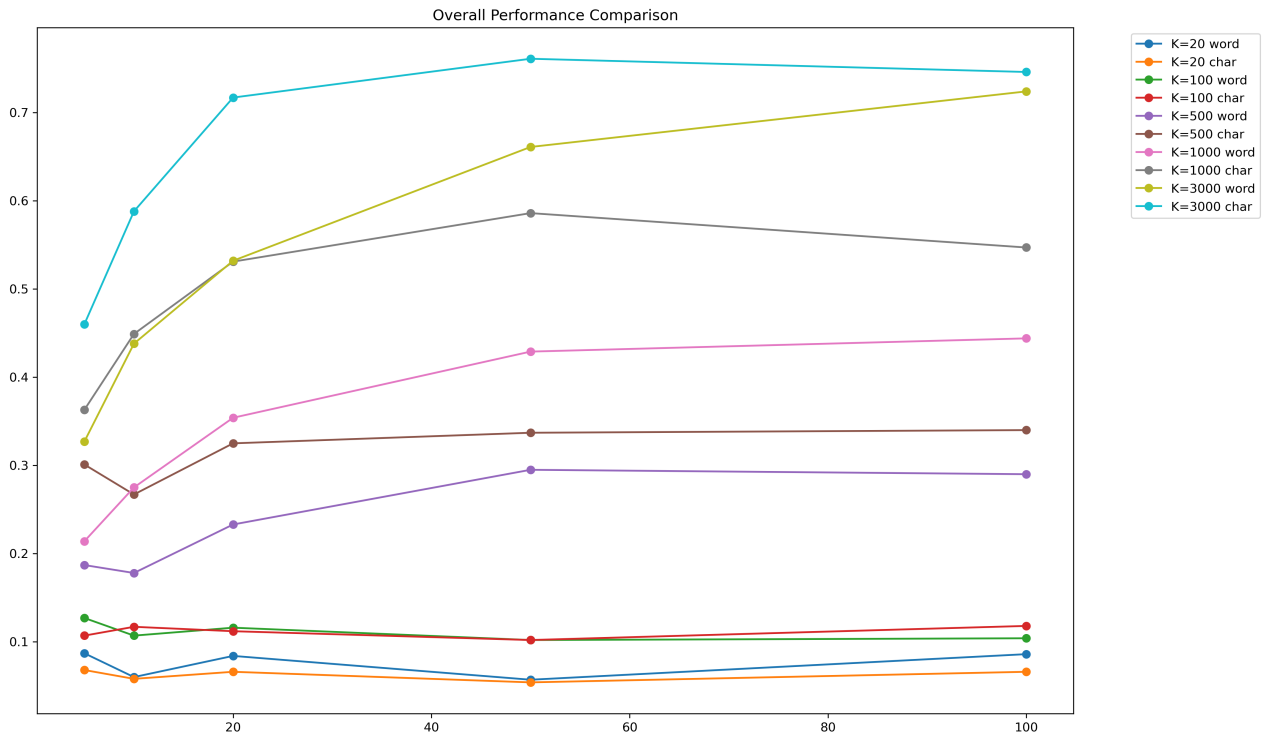
在LDA建模时，主题数量分别设置为5、10、20、50和100，并进行10次交叉验证。代码使用的分类器为LogisticRegression逻辑回归分类器。

最后计算准确率，并通过数据可视化生成图标。

**Experimental Studies**

实验以1000个段落为数据集，设定五组段落长度（20、100、500、1000、3000词或字符），文本单元分为词（经分词去停用词处理）与字符两类，主题数量设置为5、10、20、50、100。LDA模型提取主题分布特征后，进行10折交叉验证，训练集与测试集比例为9:1。

得到的实验结果如下图所示（完整实验数据在附录中给出）：



实验数据显示主题数量T存在最优区间。对于短文本来说，T的取值为5时准确率最高比如K=20、T=5时，以词为单位的准确率为8.7%，此后有所下降，这在长文本中则出现差异化，当K=3000时，T的取值需要达到50准确率才能达到峰值。这表明较长的文本需要更多主题刻画语义细节，但需避免过度细分导致的特征稀疏。

根据以上实验结果我们还可以发现，以字为单元在长文本中表现更优，在K≤100时，以词为单位的准确率大于以字为单位的准确率。当K的取值变得更大之后，现象则完全相反，变成以字为单位的准确率高于以词为单位，比如K=500，T=5时，以字为单位的准确率比以词为单位时甚至高了11.4%。

除此之外，我们发现段落长度与分类准确率呈显著正相关。当K从20增至3000时，词单元最佳准确率从8.7%提升至72.4%，字符单元从6.8%跃升至76.1%。长文本包含更完整的语义线索，使主题分布更具区分性。

**Conclusions**

本研究通过系统实验揭示了中文小说分类任务中的关键规律：主题数量需根据文本长度动态调整，长文本宜选择较大T值；字符单元在K≥500时展现显著优势；段落长度与分类性能呈正相关。这些发现为中文长文本处理提供了重要参考，未来可进一步研究跨语言场景下的普适性规律。‌

**Appendix**

**完整实验数据如下：**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| K | Unit | T | Accuracy | K | Unit | T | Accuracy |
| 20 | word | 5 | 0.087 | 500 | char | 5 | 0.301 |
| 20 | word | 10 | 0.06 | 500 | char | 10 | 0.267 |
| 20 | word | 20 | 0.084 | 500 | char | 20 | 0.325 |
| 20 | word | 50 | 0.057 | 500 | char | 50 | 0.337 |
| 20 | word | 100 | 0.086 | 500 | char | 100 | 0.34 |
| 20 | char | 5 | 0.068 | 1000 | word | 5 | 0.214 |
| 20 | char | 10 | 0.058 | 1000 | word | 10 | 0.275 |
| 20 | char | 20 | 0.066 | 1000 | word | 20 | 0.354 |
| 20 | char | 50 | 0.054 | 1000 | word | 50 | 0.429 |
| 20 | char | 100 | 0.066 | 1000 | word | 100 | 0.444 |
| 100 | word | 5 | 0.127 | 1000 | char | 5 | 0.363 |
| 100 | word | 10 | 0.107 | 1000 | char | 10 | 0.449 |
| 100 | word | 20 | 0.116 | 1000 | char | 20 | 0.531 |
| 100 | word | 50 | 0.102 | 1000 | char | 50 | 0.586 |
| 100 | word | 100 | 0.104 | 1000 | char | 100 | 0.547 |
| 100 | char | 5 | 0.107 | 3000 | word | 5 | 0.327 |
| 100 | char | 10 | 0.117 | 3000 | word | 10 | 0.438 |
| 100 | char | 20 | 0.112 | 3000 | word | 20 | 0.532 |
| 100 | char | 50 | 0.102 | 3000 | word | 50 | 0.661 |
| 100 | char | 100 | 0.118 | 3000 | word | 100 | 0.724 |
| 500 | word | 5 | 0.187 | 3000 | char | 5 | 0.46 |
| 500 | word | 10 | 0.178 | 3000 | char | 10 | 0.588 |
| 500 | word | 20 | 0.233 | 3000 | char | 20 | 0.717 |
| 500 | word | 50 | 0.295 | 3000 | char | 50 | 0.761 |
| 500 | word | 100 | 0.29 | 3000 | char | 100 | 0.746 |