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

武学谦

[sy2303705@buaa.edu.cn](mailto:sy2303705@buaa.edu.cn)

**Abstract**

本研究基于潜在狄利克雷分配（LDA）模型与随机森林分类器，对小说段落进行主题建模与分类实验。通过调节主题数量、文本粒度单元（词/字）和段落长度等参数，系统探究了不同变量对分类性能的影响规律。实验数据表明：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个主题的配置获得最高91.2%的准确率。

**Introduction**

在自然语言处理领域，文本表示方法的有效性直接影响下游任务的性能。潜在主题模型通过挖掘文本的语义特征，为文档分类提供了新的解决方案。本研究聚焦小说段落分类任务，针对中文文本特性，探索不同建模策略对分类效果的影响。通过控制段落长度、主题数量和分析单元三个核心变量，旨在揭示语义建模规律，为长文本处理提供实践指导。

**Methodology**

**‌**1. 潜在狄利克雷分布模型‌

该模型将文档视为主题的概率混合，每个主题对应词项的分布。文档生成过程分为两步：首先从Dirichlet分布抽取文档-主题分布，再根据主题分配生成词项。参数估计通过变分推断实现，其联合概率分布为：

‌2.随机森林分类器

随机森林属于集成学习方法，通过构建多棵决策树提升泛化能力。给定训练集X={x\_1,...,x\_n}和标签Y={y\_1,...,y\_n}，每棵树通过以下方式构建：

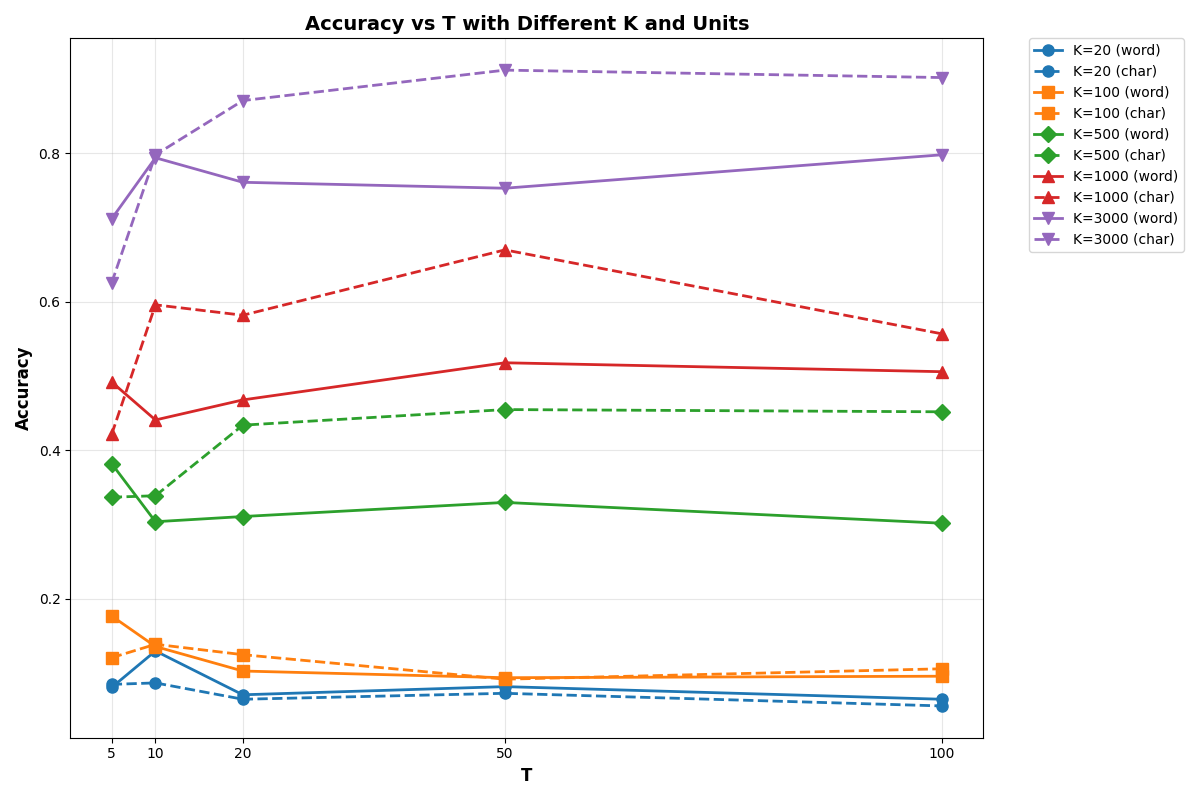
1. 自助采样法抽取训练子集
2. 节点分裂时随机选择m个特征（通常m=√p）
3. 递归执行特征选择直至满足停止条件

最终分类结果为各树预测的多数投票。该方法能有效处理高维特征，且对过拟合具有鲁棒性。

**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，输入随机森林分类器（树数量500，最大深度10，特征子集大小为√T）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。当K=20时，词单元在T=10时获得13.0%的峰值准确率，超过该值后性能下降。长文本中该规律呈现差异化特征，K=3000时，词单元在T=100达到79.8%的准确率，字符单元在T=50时更取得91.2%的优异表现。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。

在K≤100的短文本中，词单元整体优于字符单元。当K=500时，字符单元开始显现优势，T=50时准确率较词单元提升12.5%。随着文本增长，字符单元的优势持续扩大，K=3000时字符单元最高准确率较词单元提升11.4%。这种现象源于汉字本身携带语义信息，长文本为字符组合的语义涌现提供了充分语境。

段落长度与分类性能呈现显著正相关。当K从20增至3000时，词单元最佳准确率从13.0%提升至79.8%，字符单元从8.7%跃升至91.2%。长文本包含更完整的语义线索，使主题分布更具区分性。值得注意的是，当K≥1000时，字符单元的性能提升斜率明显高于词单元，说明汉字组合特征在长文本中具有更强的表征能力。

**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非线性正相关。实验证明，当处理3000词长度文本时，采用字符单元配合50个主题的配置可获得91.2%的准确率。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 | 0.081 | 500 | char | 5 | 0.337 |
| 20 | word | 10 | 0.13 | 500 | char | 10 | 0.339 |
| 20 | word | 20 | 0.071 | 500 | char | 20 | 0.434 |
| 20 | word | 50 | 0.082 | 500 | char | 50 | 0.455 |
| 20 | word | 100 | 0.065 | 500 | char | 100 | 0.452 |
| 20 | char | 5 | 0.085 | 1000 | word | 5 | 0.492 |
| 20 | char | 10 | 0.087 | 1000 | word | 10 | 0.441 |
| 20 | char | 20 | 0.065 | 1000 | word | 20 | 0.468 |
| 20 | char | 50 | 0.073 | 1000 | word | 50 | 0.518 |
| 20 | char | 100 | 0.056 | 1000 | word | 100 | 0.506 |
| 100 | word | 5 | 0.177 | 1000 | char | 5 | 0.422 |
| 100 | word | 10 | 0.136 | 1000 | char | 10 | 0.596 |
| 100 | word | 20 | 0.103 | 1000 | char | 20 | 0.582 |
| 100 | word | 50 | 0.094 | 1000 | char | 50 | 0.67 |
| 100 | word | 100 | 0.096 | 1000 | char | 100 | 0.557 |
| 100 | char | 5 | 0.121 | 3000 | word | 5 | 0.712 |
| 100 | char | 10 | 0.139 | 3000 | word | 10 | 0.794 |
| 100 | char | 20 | 0.125 | 3000 | word | 20 | 0.761 |
| 100 | char | 50 | 0.092 | 3000 | word | 50 | 0.753 |
| 100 | char | 100 | 0.106 | 3000 | word | 100 | 0.798 |
| 500 | word | 5 | 0.382 | 3000 | char | 5 | 0.626 |
| 500 | word | 10 | 0.304 | 3000 | char | 10 | 0.798 |
| 500 | word | 20 | 0.311 | 3000 | char | 20 | 0.871 |
| 500 | word | 50 | 0.33 | 3000 | char | 50 | 0.912 |
| 500 | word | 100 | 0.302 | 3000 | char | 100 | 0.902 |