文学文本智能分析的关键参数优化研究

核心发现

影响因素	最优配置	精度提升
语义维度	长文本(t=50)	**+29.9%**
处理粒度	字符特征	**+19.7%**
文本长度	≥3000字符	**+13.2%**

技术实现

- **语料构建**：16部武侠小说平衡采样，生成5种长度(20-3000字符)的8000个文本片段
- **特征工程**：双粒度处理（字符拆解 vs 结巴分词），配合542词停用表过滤
- **建模框架**：LSI语义空间构建 → Logistic分类器（L2正则化）

关键结论

1. **维度适配规律**：语义空间维度与文本长度呈强正相关（r=0.92），长文本最佳维度为50，短文本维度＞20时精度下降18%

2. **粒度优势区间**：字符特征在＞1000字符文本中优势显著，3000字符时相较词语特征提升19.7pp

3. **长度阈值效应**：1000字符为关键转折点，上下文信息量增长曲线出现明显拐点（k=0.43→0.87）

实践价值

> 本研究提出的动态参数配置策略已在古籍数字化项目中验证，使《射雕三部曲》的自动分类准确率从72.1%提升至89.3%。研究结论为中文NLP任务提供三重启示：①长文本优先采用字符级特征 ②建立维度-长度的动态映射表 ③确保关键段落≥1000字符