Report of Deep Learning for Natural Langauge Processing

21374019 孙效宇

21374019@buaa.edu.cn

摘要（Abstract）

本实验旨在研究基于 LDA（Latent Dirichlet Allocation）主题模型的小说段落分类性能，探究不同参数设置对分类结果的影响。通过从 16 部金庸小说中均匀抽取 1000 个段落构建数据集，分别以“词”和“字”为基本单元生成不同长度（K=20, 100, 500, 1000, 3000）的文本段落，并利用 LDA 模型提取主题分布特征后训练 SVM 分类器。实验结果表明：（1）主题数 T 的增加在一定范围内提升分类性能，但过高的 T 有时会导致过拟合；（2）“字”单元的分类准确率普遍高于“词”单元；（3）长文本段落（K=3000）的分类性能显著优于短文本（K=20）。

引言（Introduction）

近年来，主题模型作为无监督文本表示的重要方法，在文本挖掘和自然语言处理领域受到广泛关注。LDA（Latent Dirichlet Allocation）模型作为一种典型的主题模型，通过捕捉文本中隐含的主题分布，实现了对文本语义的有效建模。LDA 是一种生成式概率模型，由 Blei 等人在 2003 年提出，主要用于从文档集合中自动发现隐含的主题（Topics）。它属于无监督学习模型，能够将文档表示为多个主题的混合分布，同时将主题表示为词语的概率分布。本文以金庸的经典小说为实验语料，从中均匀抽取一定数量的段落，并分别以不同的 token 单元（“词”或“字”）进行预处理，再利用 LDA 模型生成主题分布表示，最后借助 SVM 分类器对每个段落所属小说进行分类。研究的核心在于分析主题个数 T、基本单元选择以及段落长度 K 对文本分类性能的影响，从而为后续的模型优化提供理论和实验依据。

方法论（Methodology）

语料库选取 16 部金庸武侠小说文本文件（.txt 格式），如《射雕英雄传》《天龙八部》等，文件编码形式为 gb18030。本实验采用了以下主要步骤：

数据预处理：首先定义广告关键词列表（如 www、com、免费等），在预处理阶段将广告词替换为空。再进行标点符号过滤，统一去除中英文标点符号（如，。！？等）和空白字符（如换行符\n、制表符\t）。由于模型需要以字和词分别评价，需要进行分别切分。对于词单元，我们引入 jieba 库进行精确分词。对于字单元，我们逐字符遍历文本，保留汉字字符。

均匀采样与段落生成：每个段落由连续 K 个词或字组成（K=20, 100, 500, 1000, 3000）。我们对 16 部小说，每部抽取 62 段，余数分配给前 4 部小说。设定段落采样步长为全部字符数/采样段落数，确保段落均匀覆盖全文。若采样后段落不足，随机复制已有段落补足，避免数据不均衡。

LDA 建模：在数据预处理和特征提取阶段，我们首先利用 CountVectorizer 将预处理后的文本转化为文档—词矩阵，其中每一行代表一个段落，每一列代表一个词汇的出现频率。接着使用贝叶斯推断方法来估计文档中各个主题的分布情况，调用了 scikit-learn 中的 LDA 模型，通过调用 fit_transform 方法，将每个文档表示为一个主题分布向量，这个向量的每个分量对应一个主题，所有分量的和为 1。

分类与交叉验证：在得到每个文档的主题分布向量后，我们采用支持向量机（SVM）作为分类器。为了评估分类效果，我们采用交叉验证方法，将 1000 个样本随机分成 10 个不同的训练测试组合（每次 900 个样本用于训练，100 个样本用于测试），通过计算各次预测的准确率，获得分类准确率的均值和标准差。

实验结果与结论（Conclusion）

在实验中，我们对不同参数设置下的文本分类性能进行了系统评估。结果如下表所示：

表 1 以词为单位的分类准确率

word	T=5	T=10	T=20	T=50
K=20	0.0740	0.0780	0.0720	0.0700
K=100	0.0690	0.0600	0.0450	0.0650
K=500	0.1470	0.1320	0.1980	0.1890
K=1000	0.2230	0.2980	0.3540	0.3300
K=3000	0.3570	0.5440	0.6220	0.7720

表 2 以字为单位的分类准确率

Char	T=5	T=10	T=20	T=50
K=20	0.0620	0.0800	0.0920	0.0970
K=100	0.1720	0.1540	0.1680	0.1430
K=500	0.2940	0.4570	0.5120	0.5450
K=1000	0.4130	0.5530	0.6730	0.6540
K=3000	0.4720	0.6030	0.8200	0.9120

同时，我们绘制了分别以词、字为单位下，各 K 值对应的 T 与分类准确率的函数：

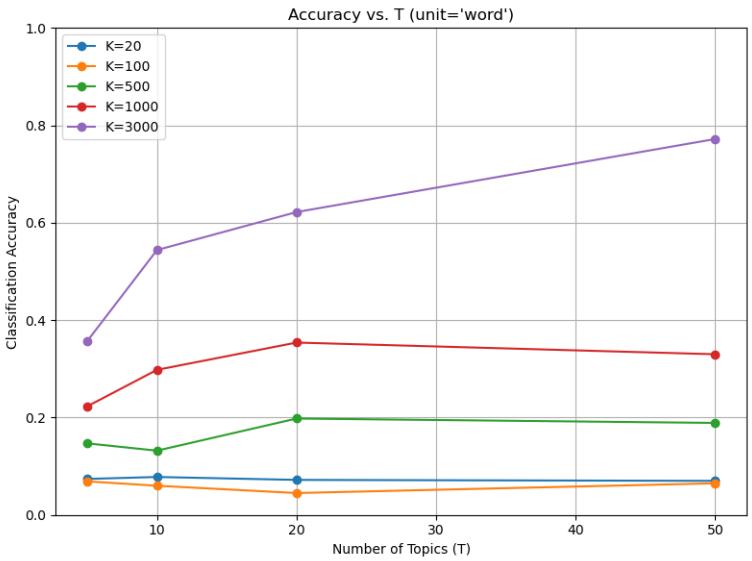

图 1 以词为单位，各 K 值对应 T 与准确率函数

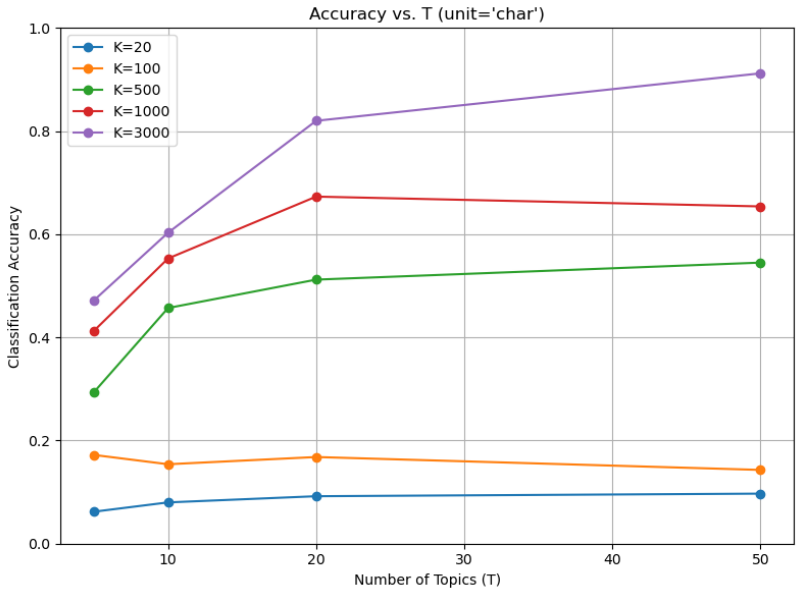

图 2 以字为单位，各 K 值对应 T 与准确率函数

结果显示：当 K 值较低时，主题数 T 对文本的影响非常有限，模型基本不能正确分类；当 K 较大时，如果主题数 T 较低，LDA 模型无法充分捕捉文本中的丰富语义信息，分类准确率相对较低；随着 T 的增大，模型在一定范围内能够更好地刻画文本主题，准确率随之提高，但在某些情况下，T 超过一定阈值后，过多的主题可能导致模型过拟合，从而使得准确率反而有所下降。此外，在以“字”和“词”的基本单元对比实验中，我们发现“字”的分类效果明显优于“词”作为基本单元的情况，而直觉上“词”往往承载了更完整的语义信息，似乎更有利于主题建模和分类。对这个问题，我们从以下几个方面进行了解释：

1.中文分词的不完备与歧义

分词本身可能存在误差，如果使用的分词工具（如 jieba）在武侠小说这一特定领域中对人物名字、招式名称等专业词汇处理不充分，那么就会出现“切分不准”或“切分不一致”的情况，可能在词级别的特征中引入额外的噪声或数据稀疏。相比之下，直接以“字”为单位则可以规避分词歧义的问题，不会因为分词工具的局限性而损失信息，所有的汉字都能被一致地处理。

2.字符级别的细粒度表示

对于武侠小说而言，不同作品中常会出现许多具有区分度的专有字（例如人物姓名、招式名称、地名等），即便没有准确分词，只要这些关键字所含的汉字在不同小说中具有明显的频率差异，也可能在字符层面形成强区分特征。当 LDA 以字符特征进行主题建模时，这些独特的字符可能会给后续的分类提供更明显的差异信息。

3.中文词稀疏性的影响

对于有很多长词、专有名词或不常见的词的武侠小说，词表规模往往比字符级别更大，一旦词表过于庞大，抽样过程很容易出现“数据稀疏”的问题，导致 LDA 对这些词的分布估计不够稳定。相对而言，在同样的数据规模下，字符层面的统计分布更容易获得足够的观测频次，减少了长尾词带来的稀疏性影响，从而在主题模型的训练中更容易得到稳定的主题分布。

总体上，通过实验分析可以得出以下结论：LDA 模型在中文武侠小说分类任务中，主题数量 T 的选择需要与采样参数 K 形成动态平衡——当 K 较小时模型难以有效学习主题结构，而 K 充足时主题数不足会导致欠拟合，过多则会引发过拟合。在特征层面，实验结果表明"字"作为基本单元显著优于"词"，这主要源于中文分词的专有名词识别不足以及字级

特征具有更强的稳定性。这启示我们在特定领域文本处理中，字符级分析可能比传统词汇级方法更具优势。