Report of Deep Learning for Natural Langauge Processing

张永易
2394822700@qq.com

Abstract

本实验将从给定的语料库中均匀抽取 1000 个段落作为数据集，并根据每个段落所属的小说进行标记。接下来，你需要利用 LDA 模型对这些段落进行文本建模，并将每个段落表示为主题分布。然后，你将使用任意选择的分类器对这些主题分布进行分类，并进行 10 次交叉验证。在这个过程中，首先会探究主题数量对分类性能的影响。尝试不同数量的主题（T），比如 20、50、100 等，并观察分类性能是否有所变化。其次是探究不同 K 值的短文本和长文本对分类模型准确性的影响，在抽取段落时，尝试不同的 K 值（20、100、500、1000、3000），分别作为短文本和长文本进行建模和分类，并观察主题模型的性能是否有所差异。最后探究以"词"和以"字"为基本单元进行分类的差异。对比使用以词和以字为基本单元进行分类的结果。通过这些实验和观察，更深入地理解主题模型在不同情况下的表现，并为讨论提供有价值的信息。

Introduction

在当今信息爆炸的时代，处理和分类海量文本数据是一项极具挑战性的任务。文本分类作为自然语言处理（NLP）领域中的一个基础任务，被广泛应用于各种场景，如搜索引擎优化、情感分析、新闻推荐等。其核心目标是将文本数据按照预定义的类别进行划分和归类，以便更好地理解和利用这些信息。

在这个背景下，主题模型的出现为文本分类任务提供了全新的解决思路。其中，潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）模型作为一种强大的统计方法，引起了研究者们的广泛关注。LDA 模型通过无监督地发现文档集合中的潜在主题结构，为文本分类提供了一种全新的视角。

具体而言，LDA 模型假设文档是由多个主题的混合生成的，每个主题又被一组词的概率分布所表征。通过对文档的主题分布进行建模和推断，LDA 模型可以揭示文本数据中潜藏的语义结构，为后续的分类任务提供了重要线索。

本次实验旨在深入探究 LDA 模型在文本分类中的应用效果，并特别关注几个关键参数对分类性能的影响。通过系统地调整和比较主题数量、分词单位以及段落长度等变量，我们希望更全面地理解主题模型在文本分类任务中的表现和特性。这将为文本分类领域的研究和实践提供重要的参考和启示，有助于推动该领域的进一步发展和创新。

Methodology

LDA 模型

在 LDA 模型中，一篇文档的生成方式如下：

从迪利克雷分布α中取样生成文档 i 主题分布 θi；

从主题的多项式分布θi 中取样生成文档 i 第 j 个词的主题 zi，j；

从迪利克雷分布β中取样生成主题 zi，j 对应的词语分布 Φzi,j；

从词语的多项式分布Φzi,j 中采样最终生成词语 wi,j。

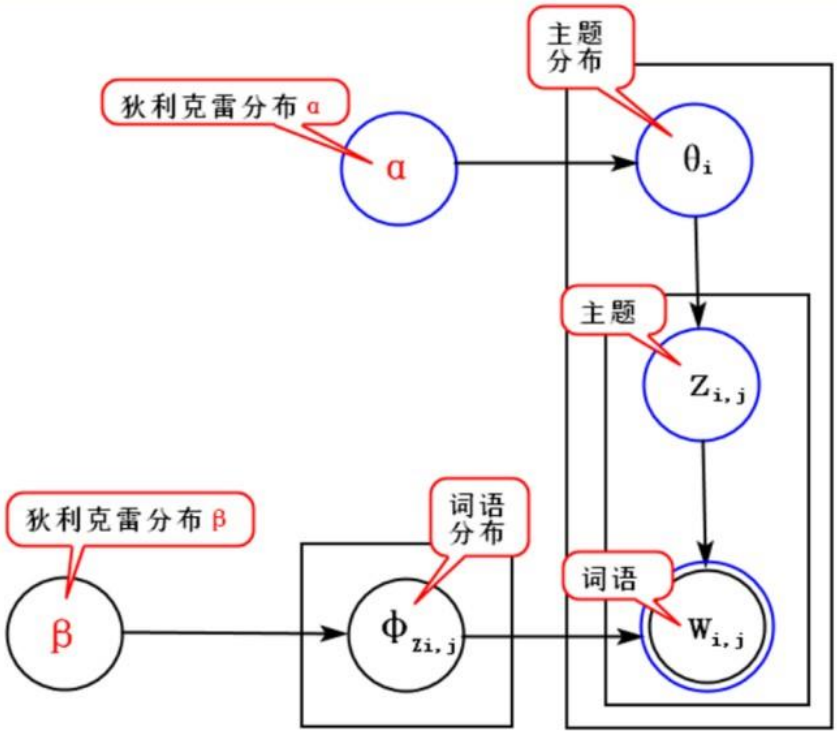


图 1 LAD 作用机理示意图

语料库的选择

本研究选取了由著名作家金庸所撰写的十六部武侠小说作为研究对象的中文语料库。值得注意的是，尽管这些小说的语言风格与现代标准中文存在显著差异，但武侠小说的表达方式本身可被视为一种独特的文学风格。这种风格不仅反映了特定的文化背景和时代特色，同时也体现了金庸先生个人的语言特点。

数据处理方法

本研究所使用的数据集包含了金庸先生创作的十六部小说。鉴于原始文本中存在大量的乱码、无效内容以及重复的中英文符号，对数据集进行彻底的预处理显得尤为重要。预处理步骤包括：首先，清理文本以去除所有隐藏字符；其次，删除所有非中文字符，以确保分析的纯净性；然后，移除停用词，以过滤掉在分析中无关紧要的常见词汇；最后，在不考虑上下文的情况下，移除所有标点符号，以避免对分词结果产生干扰。本研究采用了 jieba 分词库进行文本处理，jieba 是 Python 中广泛使用的中文分词工具。在实验中，我们采用了 jieba 的精确模式进行分词，旨在最大程度上保证文本分词的准确性和效率。这一系列预处理措施为后续的数据分析提供了干净、可靠的文本基础。为确保实验中对于不同书目的抽取相对均匀，我们在代码中加以设置，使得其中 1000 个段落的获取更具均匀性。

LDA 模型配置

在本研究中，我们采用了潜在狄利克雷分配（LDA）模型来从文本中提取主题特征，以用于文本分类任务。LDA 模型的实现基于 gensim 库，该库提供了广泛的功能，用于构建和训练主题模型。

构建词典

首先，利用 gensim 的 Dictionary 类，从预处理后的文本数据 corpus 创建一个映射（词到 ID 的映射），即词典。这个词典是后续创建词袋模型的基础。

构建词袋模型

利用上一步得到的词典，将每个文档转换为词袋模型。词袋模型是一个列表，其中每个元素是一个二元组，包括词的 ID 和该词在文档中出现的次数。

配置 LDA 模型

主题数量（num_topics）：根据实验需求设置，默认为 10。这一参数用于调整模型能够发现的潜在主题数量。迭代次数（passes）：设置为 10，以确保模型有足够的迭代次数来达到良好的收敛状态。较多的迭代次数有助于模型更准确地学习文档和主题之间的关系。

以下是步骤和参数设置：

训练 LDA 模型

使用文档的词袋表示和设置好的参数在 gensim.models.LdaModel 中训练 LDA 模型。

从训练好的 LDA 模型中提取特征涉及以下步骤：

获取文档的主题分布：对每个文档的词袋表示调用 get_document_topics 方法，这将返回一个列表，其中每个元素是一个元组，代表一个主题及其在该文档中的分布概率。

构建特征向量：初始化一个零矩阵 topic_features，其形状为文档数乘以主题数。

遍历每个文档的主题分布，将每个主题的概率填充到对应的位置上。这样，每个文档都被表示为一个固定长度的向量，向量的每个维度对应一个主题的概率。

这种基于 LDA 的特征表示方法将文本的高维信息压缩到了一个低维的主题空间中，这不仅有助于提高分类器的处理效率，还可能增强模型处理文本的能力，因为这些主题捕捉了文本中的潜在语义结构。通过这种方式，每个文档的特征向量都准备好可以被用于进一步的分类任务，如支持向量机（SVM）、logistic 回归、决策树等机器学习算法。

Experimental Studies

主题个数与分类性能的研究

为探究选取的主题个数对于，训练得到的分类器性能。分别计算数据集在[16、32、64、128、256]个主题上训练得到的支持向量机、logistic 回归模型、决策树模型的验证准确率水平，在实验中包含以词为单位（a 组）和以字为单位两种形式（b），实验结果如下所示：

在给定的主题数量下，SVM 模型具有最高的性能，其次是 logistic 回归模型，决策树模型的性能最低。这表明在这个特定的问题领域中，增加 topic 数量对于 SVM 模型的性能提升效果更为显著，而决策树模型对于 topic 数量的变化更为敏感，性能变化较大。随着 topic 数量的增加，SVM 模型和 logistic 回归的性能呈现出明显的改善趋势。这表明 SVM 模型和 logistic 回归对于更多的主题能够更好地进行分类和预测。相比之下，决策树模型的性能随 topic 数量的增加而呈现出逐渐下降的趋势。这可能意味着这两种模型对于更多的主题可能存在一定的过拟合或不适应性。因此，总体来说，增加 topic 数量对于 SVM 模型和 logistic 回归模型性能的提升效果更为显著，而对于决策树模型的性能可能有一定的负面影响。总体来看，当 topic 的增量在一定范围时模型的性能均会随着 topic 的增加而增加，因此在选择 topic 时可以先进行最优值的探索，从而保证模型的性能水平。

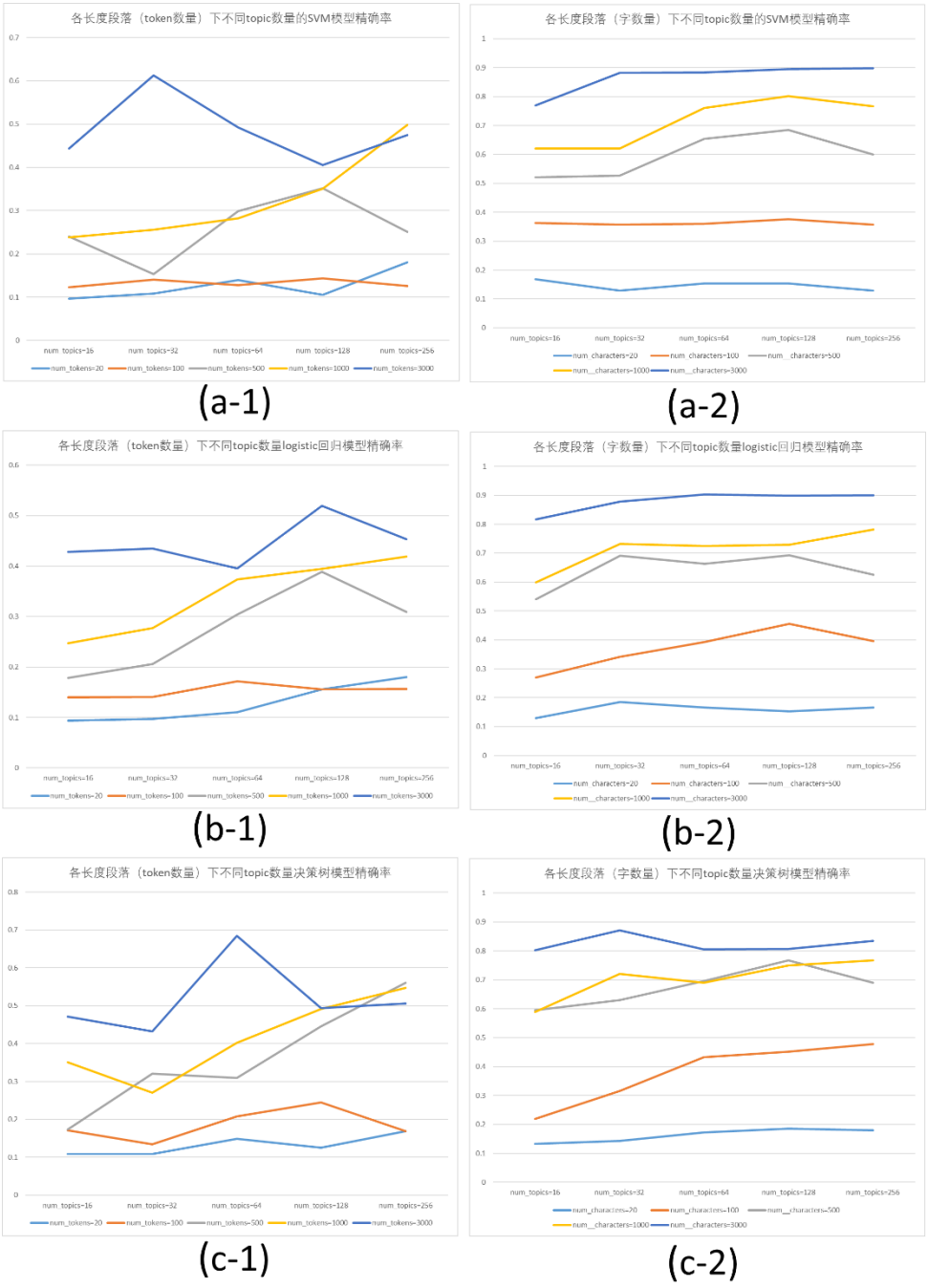
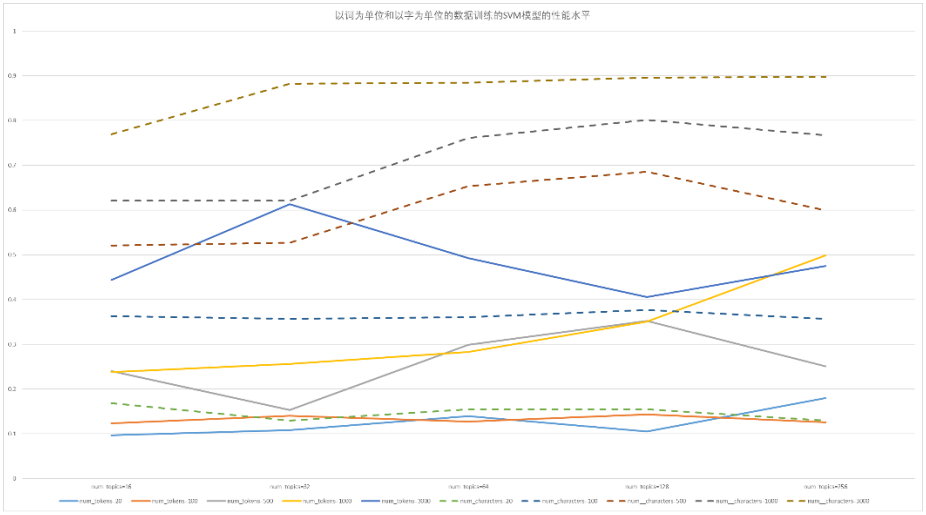


图 2 主题个数与分类性能的实验结果

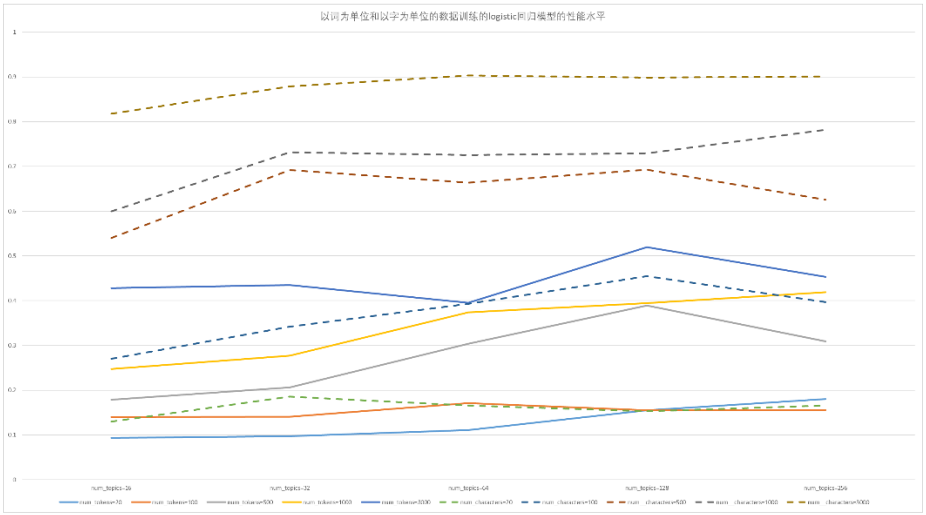
基本单元选取方法与分类性能的研究

为探究选取的主题个数对于，训练不同的的分类器观察其性能。分别计算数据集在[16、32、64、128、256]个主题上训练得到的支持向量机（a 组）、logistic 回归模型（b）、决策树模型（c）的验证准确率水平，其中以词为单位的实验组由实线表示，以字为单位的实验组由虚线表示，实验结果如下所示：

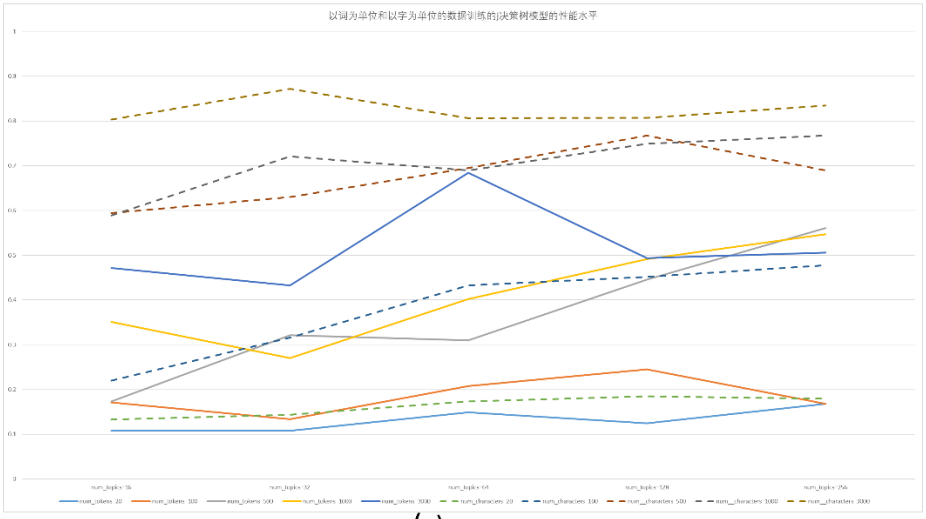
随着数据基本单位从词到字的转变，模型性能也有所改变。具体而言，当数据基本单位为词时，模型性能可能较低，而当数据基本单位为字时，模型性能可能有所提高。这说明在这个特定的问题领域中，使用更细粒度的数据基本单位（词）可能有助于提高模型的性能。



(a)



(b)



(c)

图 3 基本单元选取方法与分类性能的实验结果

文本长度与分类性能的研究

随着文本长度的增加，模型的性能呈现出不同的变化规律。具体而言，SVM 模型在不同文本长度下的性能表现均较好，相对于 logistic 回归和决策树模型，SVM 模型的性能更稳定

且更高。这表明 SVM 模型在处理不同文本长度时具有更好的分类和预测能力。

logistic 回归和决策树模型的性能在不同文本长度下存在一定的波动。在某些文本长度下，它们的性能可能接近或甚至优于 SVM 模型，但在其他文本长度下可能表现较差。这可能是由于这两种模型对于文本长度的变化更为敏感，可能需要更多的数据或调整来适应不同长度的文本。

随着文本长度的变化，SVM 模型表现出更稳定和优越的性能，而 logistic 回归和决策树模型的性能可能有所波动。然而，为了更准确地分析模型性能与文本长度的变化规律，我们可能需要更多的信息和数据来进行进一步的研究和验证。

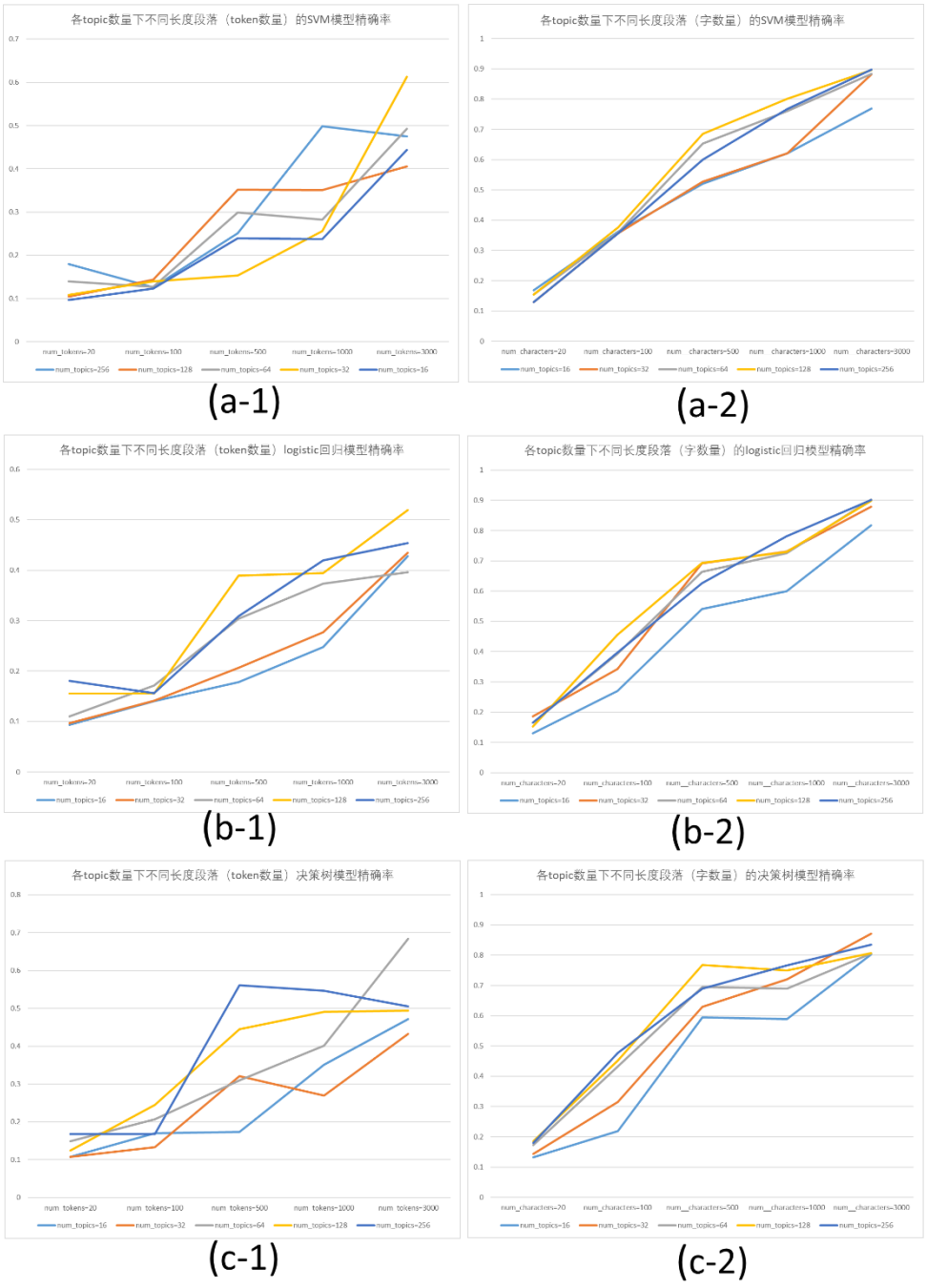


图 4 文本长度与分类性能的实验结果

Conclusions

在本次实验中，我们深入探讨了 LDA（Latent Dirichlet Allocation）模型在文本分类任务中的应用，特别是在金庸武侠小说的中文语料库上的表现。LDA 模型作为一种强大的无监督学习方法，能够有效地发现文档集合中的潜在主题结构，为文本分类提供了有价值的特征表示。

实验选取了金庸的十六部武侠小说作为研究对象，并进行了详尽的数据预处理工作，包括去除乱码、无效内容、非中文字符、停用词以及标点符号等，以确保分析的准确性和可靠性。随后利用 jieba 分词库对预处理后的文本进行分词，为后续的主题建模奠定了基础。

在 LDA 模型的配置阶段，主要关注了主题数量这一关键参数，并设置了不同的值以探究其对分类性能的影响。此外，我们还探讨了分词单位（词或字）以及段落长度（通过调整 K 值）对模型性能的影响。

在 LDA 模型的训练过程中，构建了词典和词袋模型，然后使用 gensim 库中的 LdaModel 类对模型进行训练。训练完成后提取了每个文档的主题分布作为特征向量，用于后续的文本分类任务。

实验结果表明，LDA 模型在金庸武侠小说的文本分类任务中表现出了良好的性能。通过调整主题数量、分词单位以及段落长度等参数，可以进一步优化模型的性能。特别是，合适的主题数量能够更准确地反映文档集合中的语义结构，从而提高分类的准确性。

此外，还发现分词单位对模型性能的影响较大，但使用字作为基本单元可能更有助于捕捉文本的语义信息。段落长度的调整也对模型性能产生了一定影响，但具体效果取决于数据集的特点和分类任务的需求。

综上所述，LDA 模型在文本分类任务中展现出了强大的潜力和应用价值。通过合理的参数配置和特征提取方法，我们可以充分利用 LDA 模型的优势，提高文本分类的准确性和效率。

References

[1] Zenchang Qin and Lao Wang (2023)，How to learn deep learning? Journal of Paper Writing, Vol. 3: 23: pp. 1-12.

[2] ….