Report of Deep Learning for Natural Langauge Processing

冷元飞
ffxdds@163.com

Abstract

以文学作品段落为对象，运用 LDA（Latent Dirichlet Allocation）模型进行主题建模，并以此为基础构建文本分类器，旨在探究不同主题数量、分词单位（词与字）以及段落长度对分类性能的影响。

Introduction

随着数字化技术的发展，大量的文学作品以电子形式存储和传播，为文学研究带来了前所未有的便利。然而，面对海量文本数据，如何有效地提取、理解和分析其中蕴含的丰富信息，成为现代文学研究面临的重要挑战。文本分类作为一种基础且重要的文本挖掘技术，旨在依据文本内容将其自动归入预定义的类别中，对于文学作品的研究、检索、推荐等方面具有广泛应用价值。

本文聚焦于文学作品段落的文本分类问题，以期通过自动化手段揭示段落与其所属小说之间的内在联系，进一步深化对作品主题、风格及作者创作特色的理解。特别地，我们采用LDA（Latent Dirichlet Allocation）模型进行主题建模，这是一种统计学习方法，能够在大规模文本数据中发现隐藏的主题结构，为文本分类提供高层次、抽象的特征表示。LDA 模型的优势在于，它能够将复杂的文本内容简化为一组主题及其在文本中的概率分布，有助于捕捉文本的核心思想，减轻传统基于词汇或短语特征的分类方法面临的维度灾难问题。尽管 LDA 模型在文本分类领域的应用已取得一定成果，但在文学作品这一特定领域，其性能受多种因素影响，如主题数量的选择、分词单位的选择（词或字）以及段落长度等。这些因素如何影响 LDA 模型在文学作品段落分类任务中的表现，尚未得到充分探讨。因此，本研究旨在通过实证分析，深入研究以下问题：

1) 主题数量对分类性能的影响：探究在文学作品段落分类中，设定不同数量的主题（T）对分类准确性的影响，寻找既能捕捉文本主题结构又能避免过拟合的最优主题数。
2) 分词单位对分类结果的影响：对比以“词”和“字”为基本单元进行主题建模及分类的结果，分析两种分词方式在捕捉语义层面主题与关注词汇形态特征及上下文依赖方面的差异，确定更适合文学作品段落分类的分词单位。
3) 段落长度对主题模型性能的影响：分析不同长度（K 值）的段落（短文本与长文本）对 LDA 模型性能的影响，理解模型在处理不同长度文本时的适应性和局限性。

为实现上述目标，本研究设计并实施了一套完整的实验流程，包括文本预处理、LDA 主题建模、基于主题分布的文本分类以及交叉验证性能评估。通过对实验结果的深入分析，我们揭示了主题数量、分词单位和段落长度等因素对文学作品段落分类性能的具体影响，为今后在类似任务中有效应用 LDA 模型提供了实证依据与实践指导。

Methodology

一、LDA 模型建模

LDA（Latent Dirichlet Allocation，潜在狄利克雷分配）是一种统计模型，用于发现大规模文档集合中的隐藏主题结构。它假设文档是由一系列不可见的主题生成的，而每个主题又由一组词语的概率分布所定义。LDA 通过概率统计方法对文档集进行建模，旨在揭示文档间的潜在主题关系以及主题内部的词汇关联。

- 文档（Document）：LDA 处理的对象是文档集合，每个文档由一系列词语组成。
- 主题（Topic）：主题是文档中词汇出现模式的抽象概括，由一组词语及其对应的概率分布构成。例如，一个关于“科技”的主题可能包含“AI”、“机器学习”、“算法”等词语，每个词语都有其在该主题下的概率。
- 词袋（Bag of Words）：在 LDA 模型中，文档被视为无序的词袋，即忽略词语的顺序和语法关系，仅关注文档中出现的词语及其频次。

- 生成过程（Generative Process）：对于每个文档从全局主题分布中抽取主题比例向量 θ（Dirichlet 分布）对于文档中的每个词语：从文档的主题比例向量 θ 中抽取一个主题 z（多项式分布）从该主题对应的词语分布 φ 中抽取一个词语 w（多项式分布）参数估计（Parameter Estimation）：LDA 模型的实际应用中，我们已知的是文档集合 D，但 θ、φ 和 z 都是未知的。通常使用 Gibbs Sampling、Variational Inference 或者其他近似推断方法来估计这些隐变量。
- 主题表示（Topic Representation）：对于每个文档，LDA 模型输出其主题分布 θ，即该文档在各个主题上的概率分布。这样，每个文档就可以被表示为其主题分布，而不是原始的词序列。

本研究采用定量分析方法，通过构建 LDA 主题模型并结合文本分类技术，对文学作品段落进行分类，以探究主题数量、分词单位以及段落长度对分类性能的影响。

Experimental Studies

LDA 模型用于对抽取的段落进行主题建模。具体步骤如下：

1) 文本预处理：对每个段落进行分词（按词或字），形成词袋表示。

2) LDA 模型训练：使用 LDA 模型对预处理后的文本数据进行训练，指定主题数量 T。训练完成后，每个段落将得到一个 T 维的主题分布向量，表示该段落在各个主题上的概率权重。

3) 分类器训练：将每个段落的主题分布向量作为特征，其所属小说的标签作为目标变量，使用你选择的分类器进行训练。这里，分类器的任务是学习如何根据段落的主题分布将其正确分类到对应的小说类别中。

4) 交叉验证与性能评估：使用 10 次交叉验证对分类器进行评估，每次保留 10%的数据作为测试集，其余 90%作为训练集。计算每次交叉验证的分类性能指标（如准确率、F1 分数等），并分析不同主题数量 T、分词单位（词或字）以及段落长度（K 值）对分类性能的影响。

SVM	K	测试准确率
Char	20	14.3%
	100	13%
	500	19%
	1000	24.6%
	3000	17%
Words	20	14%
	100	12.5%
	500	18%
	1000	24%
	3000	16.5%

不同主题数量 T 对分类性能的影响：随着 T 的增加，模型可能会捕捉到更细致的主题结构，但也可能导致过拟合。理想情况下，应选择能使分类性能最佳的主题数量。

“词”与“字”作为基本单元的分类差异：使用词作为基本单元时，模型可能更能捕捉到语

义层面的主题；而使用字时，模型可能更关注词汇的形态特征和上下文依赖。比较两种情况下的分类性能，可以了解哪种分词方式更适合当前任务。
不同取值的 K（短文本与长文本）对主题模型性能的影响：较长的文本可能包含更多主题信息，有助于模型学习更准确的主题分布；而较短的文本可能主题信息较少，导致模型学习难度增大。对比不同 K 值下的分类性能，可以分析模型在处理不同长度文本时的表现。

Conclusions

通过本次实验，我们能够深入理解 LDA 用于发现大规模文档集合中的隐藏主题结构