Report of Deep Learning for Natural Langauge Processing

Homework 2

Ao Xie
xieao2019@buaa.edu.cn

Abstract

本研究旨在探讨基于潜在狄利克雷分配（LDA）模型的文本分类方法在不同主题数、段落长度、和基本单元（词 vs 字）设置下的表现。通过对中文小说数据集进行系统的实验，本文评估了分类性能随参数变化的趋势。结果表明，调整这些参数可以显著影响分类准确率，为中文文本处理提供了实证参考。

Introduction

LDA(Latent Dirichlet Allocation)模型是一种文本挖掘技术中的主题模型，主要用于从大量文档集合中发现隐藏的主题信息。LDA 是一种无监督学习技术，广泛应用于自然语言处理和文本分析领域。

LDA 模型的基本思想是：文档是由一系列的主题混合而成的，而每个主题则是由一系列特定的词汇组成。LDA 的目标是识别出文档集合中的主题，并解析每个文档的主题结构。

通过迭代的方式调整模型参数（使用变分贝叶斯方法或吉布斯抽样等算法），以最大化文档集合的似然函数，从而学习到每个文档的主题分布和每个主题的词分布。

文本分类是自然语言处理中的一个核心任务，旨在将文本内容按照特定标准分配到预定义类别。LDA 主题模型作为一种强大的文本建模工具，被广泛应用于文档的主题发现和文本分类。本研究通过实验探讨了 LDA 模型在中文文本分类中的应用效果，尤其是考虑到中文的语言特性，比如处理单个汉字与处理词语的差异。主要研究问题包括：1、主题数量（T）对分类性能的影响。2、以“词”和“字”为单位的分类性能差异。3、不同段落长度（K）对主题模型性能的影响。

Methodology

LDA（Latent Dirichlet Allocation）是一种用于主题建模的概率生成模型，其基本原理是假设每篇文档都由一组主题构成，而每个主题又由一组单词构成。LDA 模型的生成过程如下：

确定模型参数：首先，需要确定模型的参数，包括主题数目 K 和两个狄利克雷分布的参数α和β。其中，α是文档-主题分布的参数，β是主题-词语分布的参数。

生成文档的主题分布：对于每篇文档 i，首先从狄利克雷分布α中采样生成该文档的主题分布θ_i。这里的θ_i是一个 K 维的向量，表示了文档 i 中每个主题的概率分布。

生成文档中每个词的主题：对于文档 i 中的每个词 j，从文档的主题分布θ_i中采样生成该词的主题 z_{ij}。这里 z_{ij} 是一个整数，表示第 j 个词的主题编号。

生成词语的分布：对于每个主题 z_{ij}，从狄利克雷分布β中采样生成该主题对应的词语分布$\varphi(z_{ij})$。这里的$\varphi(z_{ij})$是一个词汇表大小的向量，表示了主题 z_{ij} 中每个词的概率分布。

生成最终的词语：对于每个词 j，从词语分布$\varphi(z_{ij})$中采样生成最终的词语 w_{ij}。这样就完成了一篇文档的生成过程。

通过这个生成过程，LDA 模型假设了文档的生成过程是由主题和词语的多项式分布以及狄利克雷分布共同作用完成的。在训练过程中，LDA 通过观察文档中的词语来推断主题的分布，从而得到文档的主题表示。

Experimental Studies

1 数据集与预处理

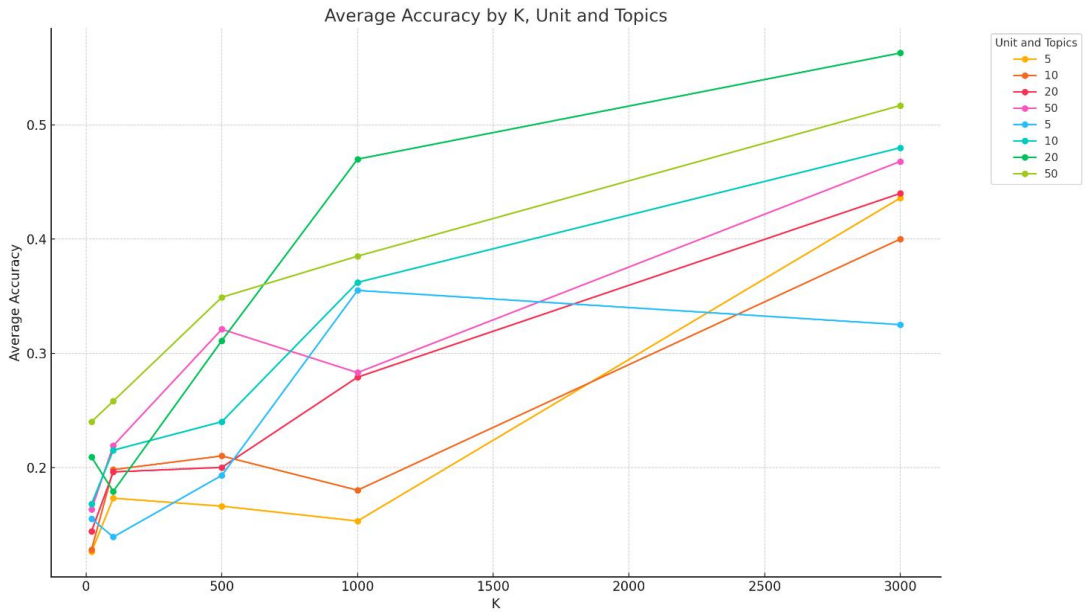
实验使用的数据集包括多部中文小说，每本书的文本被随机分割为长度不等的段落。根据实验需求，分割长度 K 设定为 20, 100, 500, 1000, 3000。文本预处理包括使用 Jieba 进行分词（对于词级处理）和直接处理汉字（对于字级处理）。

2 实验方法

实验中，首先利用 CountVectorizer 转换文本数据，然后应用 LDA 模型从转换后的文本中提取主题。LDA 模型的主题数分别设定为 5, 10, 20, 50。设置 token 值为 20、100、500、1000、3000。每个段落通过其主题分布被转换为特征向量，随后使用多项式朴素贝叶斯分类器进行分类。分类的准确性通过 10 折交叉验证来评估。一共进行 40 次实验。实验结果如下

表：

K	Unit	Topics=5	Topics=10	Topics=20	Topics=50
20	char	0.126	0.128	0.144	0.163
	word	0.155	0.168	0.209	0.24
100	char	0.173	0.198	0.196	0.219
	word	0.139	0.215	0.179	0.258
500	char	0.166	0.21	0.2	0.321
	word	0.193	0.24	0.311	0.349
1000	char	0.153	0.18	0.279	0.283
	word	0.355	0.362	0.47	0.385
3000	char	0.436	0.4	0.44	0.468
	word	0.325	0.48	0.563	0.517



Conclusions

实验结果显示，主题数量的增加通常提升了分类的准确度，但在超过一定阈值后准确度提升幅度减小。此外，以词为基本单元的分类性能普遍优于以字为单元，反映了中文处理中分词的重要性。段落长度 K 的增加对分类性能有正面影响，尤其是在较长文本中主题分布

更为明显。总的来说，这些发现强调了在设计基于 LDA 的文本分类系统时合理选择参数的重要性。