DNLP 作业 3

利用 LDA 模型在给定的语料库上的文本建模

朱道涵 20231202
20231202@buaa.edu.cn

Abstract

利用 LDA 模型，通过中文语料库进行文本建模。把每个段落表示为主题分布后进行分类，探究不同的主题个数 T 下分类性能的变化、以"词"和以"字"为基本单元下分类结果的差异。

Introduction

近年来，随着互联网技术和数字化文本存储的飞速发展，大规模中文文本的获取和分析逐渐成为自然语言处理领域的研究热点。作为文本挖掘中的重要任务之一，文本分类旨在将不同文本根据其内容归类到预定义的类别标签中。传统的分类方法通常依赖于丰富的先验知识与特征工程，然而面对如今海量的文本数据，自动化的特征学习与分类方法愈发受到研究者的关注。

主题模型（Topic Model）作为一种无监督的特征学习方法，可以有效挖掘文本中的潜在语义结构，成为文本表示与分类的有力工具。LDA 分布（Latent Dirichlet Allocation）作为经典的主题模型，可以将文本表示为多个主题的概率分布，从而将文本映射到主题空间进行分类。然而，主题模型在中文文本上的应用仍面临若干挑战，如中文的复杂字符结构与多义性。此外，由于中文文本的粒度多样（字、词、短句、长段落等），其在不同粒度文本上的表现也值得深入探讨。

本文将基于给定的中文小说语料库，探究 LDA 模型在不同设置下的文本分类性能表现。具体而言，本文将解决以下问题：

1. 不同主题个数 T 对分类性能的影响：通过设置不同的主题数量，评估 LDA 模型在不同主题空间下的分类性能变化。
2. “词”与“字”作为基本单元的分类差异：比较以“词”与“字”为基本单元的 LDA 模型对文本分类的准确性差异。
3. 不同段落长度 K 对分类性能的影响：探讨段落长度（以词或字计）不同对 LDA 模型性能的影响，特别是短文本与长文本之间的差异。

Methodology

本研究旨在探索 LDA 模型在中文小说语料库上的文本建模与分类性能，并研究不同段落长度与主题数量对分类性能的影响。为此，我们制定了如下的实验方法。

数据集构建

从语料库中均匀抽取 1000 个段落，每个段落的标签即为其所属的小说名称。我们设定段落长度 K 为 20、100、500、1000 和 3000。段落的表示形式有两种：以“词”作为基本单元和以“字”作为基本单元。每个段落经过预处理后，将移除停用词并转化为相应的向量表示。

① 语料库处理：

编码格式：将语料库文件转码为 gb18030 进行处理。

停用词移除：使用给定的 cn_stopwords.txt 停用词表对文本进行停用词处理。

② 段落构建与预处理：

按照设定的 K 值，随机从每部小说中均匀抽取段落，确保每部小说都有足够的代表性样本。

③ 特征表示：

字级别：将文本直接按单字进行分割，构建字级别词袋模型。

词级别：使用 Jieba 分词工具对文本进行分词，构建词级别词袋模型。

LDA 模型训练

在不同段落长度 K 和不同主题个数 T (5, 10, 20, 50, 100) 的组合下，使用 LDA 模型将每个段落表示为主题分布向量。

① 使用 gensim 库中的 LdaModel 模型进行训练。

② 在构建 LDA 模型时，将文本表示转换为主题分布表示。

分类器设计

将段落的主题分布向量作为特征，使用支持向量机（SVM）进行文本分类。

① 使用 sklearn 库中的 SVM 实现分类器。

② 采用 10 次交叉验证的方式，将 1000 个段落划分为训练集 (900) 和测试集 (100)。

③ 对每种 K 和 T 的组合进行独立的分类实验，计算每次交叉验证的准确率并取平均值作为最终结果。

：

Experimental Studies

表 1：以“词”为分类单位的准确率

T\K	20	100	500	1000	3000
5	0.1460	0.1710	0.3240	0.4160	0.5230
10	0.1540	0.1920	0.3750	0.4710	0.6460
15	0.1680	0.1870	0.4090	0.5010	0.6980
20	0.1630	0.1860	0.3720	0.5060	0.7230
25	0.1650	0.1920	0.3960	0.5310	0.7350

表 2：以“字”为分类单位的准确率

T\K	20	100	500	1000	3000
5	0.1360	0.2430	0.4010	0.4530	0.5920
10	0.1660	0.2480	0.5430	0.5660	0.7940
15	0.1560	0.2800	0.6100	0.6610	0.8220
20	0.1700	0.2780	0.6120	0.6600	0.8420
25	0.1710	0.2960	0.6210	0.6850	0.8600

1. 在设定不同的主题个数 T 的情况下，分类性能是否有变化？

可以观察到，随着主题数量的增加，分类性能逐渐提高，尤其在段落长度较长时（如 1000 和 3000）。当 T 达到 25 时，准确率达到最佳。

2. 以"词"和以"字"为基本单元下分类结果有什么差异？

可以观察到，段落长度较长时，以“字”为分类单位的准确率总体优于“词”。

3. 不同的取值的 K 的短文本和长文本，主题模型性能上是否有差异？

总体来看，段落长度越长提供的上下文信息越丰富，模型性能也越好。

Conclusions

① 主题数量 T 的影响：增加主题数量通常能提升分类性能，但 T 的增长需要结合段落长度和文本内容的复杂度。

② 基本单元差异：以“字”为分类单位通常能提供更高的准确率，尤其在长段落情况下。

③ 段落长度 K 的影响：段落长度增加能够显著提升分类性能，但也会增加计算复杂度。