Deep NLP 第二次作业

ZB2303019　汪婧伶

Abstract

本实验旨在探究在中文小说语料库上使用 Latent Dirichlet Allocation (LDA) 模型进行文本分类的性能。实验分为三个部分：（1）研究 token 数目 K 对文本分类性能的影响；（2）研究主题数 T 对文本分类性能的影响；（3）比较以“字”和“词”为基本单位对分类性能的影响。通过这些实验，旨在深入理解 LDA 模型在文本分类任务中的应用，并探究其在中文小说语料库上的效果。

Introduction

文本分类是将文本文档自动划分到预定义的类别中的任务。在本实验中，我们使用了 Latent Dirichlet Allocation (LDA) 模型来完成文本分类任务。LDA 是一种生成式概率模型，被广泛应用于文本挖掘和主题建模领域。其基本原理是假设每个文档是由一组主题构成的概率分布生成的，而每个主题则是一组词的概率分布。通过对文档-词频矩阵进行分解，LDA 可以推断出文档的主题分布和主题的词分布。

为了完成文本分类任务，我们将每本小说的文本划分为段落，并为每个段落分配正确的标签。然后，我们使用 LDA 模型对这些段落进行建模，并提取段落的主题分布。最后，我们使用随机森林分类器将段落文本分类到正确的小说标签上。

LDA 模型之所以适用于文本分类任务，是因为它能够自动地从文档中学习主题，并将文档表示为主题分布的形式。这种表示形式具有良好的语义解释性，能够在文本分类任务中提供有用的信息。因此，选择使用 LDA 模型来完成本实验中的文本分类任务。

Methodology

M1: 研究 token 数目 K 对文本分类性能的影响

本实验旨在探究 token 数目 K 对文本分类性能的影响。通过改变 token 数目 K，使用不同的特征表示方法来构建文本特征，然后使用随机森林分类器对文本进行分类，并通过交叉验证评估分类性能。

在实验中，首先从给定的语料库中读取小说文本数据，并对每个小说的文本进行分段处理，去除其中的停用词和无用文本。随后，根据不同的 token 数目 K（20、100、500、1000、3000）构建文档-词频矩阵，其中每个文档表示一个段落，每个词的频率表示该词在该段落中出现的次数。接着，使用 LDA（Latent Dirichlet Allocation）模型对文档进行主题建模，将每个段落表示为主题分布。最后，我们使用随机森林分类器对文本进行分类，并通过交叉验证评估分类器的性能。

实验中，固定主题数为 100，以字为基本单位进行训练分类。通过记录不同 token 数目 K 下的分类性能，并进行分析。

M2：研究主题数 T 对文本分类性能的影响

实验中，固定 token 数目 K=3000，以字为基本单位进行训练分类。通过改变主题数 T，使用不同的主题数来进行文本建模，并进行分类性能分析。

M3：比较以“字”和“词”为基本单位对分类性能的影响

本实验旨在比较以“字”和“词”为基本单位对文本分类性能的影响。使用不同的基本单位来构建文本特征，并使用随机森林分类器对文本进行分类，并通过交叉验证评估分类性能。

以“词”为基本单位：使用词作为基本单位，构建文档-词频矩阵，每个词的频率表示该词在该段落中出现的次数。

以“字”为基本单位：使用字作为基本单位，构建文档-字频矩阵，每个字的频率表示该字在该段落中出现的次数。

Experimental Studies

表 1 实验一结果

K（T=100）	ACCURACY
20	0.163
100	0.15
500	0.447
1000	0.745
3000	0.983

根据实验结果发现，分类性能基本随着 K 的增大而变好，当 K>1000 时，分类正确率有

大幅度提升。当 K>=3000，分类性能趋于稳定。

表 2 实验二结果

T（K=3000）	ACCURACY
5	0.661
10	0.928
15	0.932
20	0.943

根据实验结果发现，当 K 固定时，T 越大，分类性能越好，当 T<10 时，分类性能提升幅度大；当 T>10 时，分类性能提升幅度小

表 3 实验三结果

基本单位（K=3000,T=10）	ACCURACY
词	0.558
字	0.928

根据实验结果发现，当 K、T 固定时，统计频率的基本单位对分类性能影响较大。以字为基本单位时，分类性能优于以词为基本单位。

Conclusions

通过三个实验，加强了对 LDA 模型的理解，对文本建模有了一定的了解。