Report of Deep Learning for Natural Langauge Processing 2

Di Wu
wd_0509@163.com

Abstract

使用金庸语料库，利用 LDA 模型进行主题建模，并构建文本分类器，以探究在不同主题个数、不同分词（字、词）以及段落长度对主题模型分类性能的影响。

Introduction

在信息时代，文本数据的数量急剧增加，如何从大量的文本数据中有效地提取出有用的信息是一个至关重要的问题。这不仅关系到信息检索的效率，还关系到知识管理、决策支持和许多自动化过程的有效性。

在自然语言处理和文本挖掘领域，有许多技术和模型被开发出来应对这一挑战，例如主题模型、机器学习分类器、情感分析等。其中，LDA 模型就是一个非常流行的工具，它能够从大规模文本集中自动发现主题，帮助人们理解文档的主要内容和结构。这种类型的技术可以应用于多种场景，例如新闻聚类、社交媒体趋势分析、客户反馈主题提取等等。通过自动提取主题，我们可以更快地获得关键信息，提高处理文本数据的效率。这不仅使得我们能够管理更大量的数据，而且还能帮助我们从中发现趋势、模式和见解，从而做出更明智的决策。

本文旨在研究 LDA 模型在文学作品领域影响其分类性能的因素，针对以下三个方面进行探讨：

1、主题数量 T 对分类性能的影响：探究不同主题数量对模型的解释能力和分类效果的影响。

2、基于“词”与“字”的分类差异：分析在中文文本处理中，选择“词”或“字”作为基本单元对信息捕捉和模型表现的影响。

3、不同长度 K 的文本对模型性能的影响：段落长度 K 是决定文本覆盖的细节程度的重要因素，分析不同长度的段落对 LDA 模型性能的影响。

Methodology

Part 1:文本读取与预处理

遍历目录下的文件，读取小说的文本内容，去除停用词以减少数据的噪声和冗余，并进行数据清洗，分别针对以“字”和“词”为单位进行文本预处理，将文本数据转换为词袋模型。

Part 2：LDA 建模与训练

。在创建词典和语料库之后，使用 Gensim 库中的 LdaModel 类来训练 LDA 模型，并指定主题的数量。最后使用训练好的模型对文档进行主题分布的推断，得到每个文档的主题分布向量。

Part 3：分类与验证

选择随机森林分类器作为我们的分类器，并分别对以词为单位和以字为单位的特征数据进行十折交叉验证，输出平均准确率。

Experimental Studies

Part 1:不同主题个数 T 以及基本单元对分类性能的影响

设定主题个数 T 分别为 5，10，20，100，500，1000 进行试验，分别以字和词为单位，在 K=1000 的情况下使用随机森林分类器进行建模分析，得到分类准确率如表 1 所示。

Table 1:不同主题个数 T 以及基本单元对分类性能的影响

主题个数	以字为单位的准确率	以词为单位的准确率
5	0.5761363636363636	0.5829545454545455
10	0.7420454545454546	0.7159090909090909
20	0.7795454545454545	0.8295454545454545
100	0.8352272727272727	0.8625
500	0.8613636363636366	0.8772727272727273
1000	0.8545454545454545	0.8579545454545455

Part 2:不同长度 K 的文本对分类性能的影响

设定 K 分别为 20，100，500，1000，3000 进行试验，分别以字和词为单位，在 T=100 的情况下使用随机森林分类器进行建模分析，得到分类准确率如表 1 所示。

Table 2:不同长度 K 的文本对分类性能的影响

段落长度	以字为单位的准确率	以词为单位的准确率
20	0.17373737373737375	0.1797979797979798
100	0.34424616792495993	0.2636363636363636
500	0.7011363636363637	0.7125000000000001
1000	0.859090909090909	0.8534090909090908
3000	0.9727272727272729	0.9647727272727273

Conclusions

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。实验表明，随着 T 的增加，分类准确率先上升后下降，在 T=500 左右到达峰值。

基本单元对分类性能的影响：在中文文本处理中，选择“词”或“字”作为基本单元会影响信息的捕捉和模型的表现。以“词”为单元能捕获更多的语义信息，而以“字”为单元可能更关注形式和结构。这两种方法在处理同样的文本数据时，会显示出不同的分类效果。

文本长短对分类性能的影响：较长的文本包含更多的主题信息，有助于模型学习更准确的主题分布。