NLP_REPORT_2

基于中文语料库的的 LDA 模型训练分类

ZY2343226　朱子航

1010951286@qq.com

Abstract

本文使用所给的中文语料库文件（jyxstxtqj_downcc.com）中文本均匀抽取段落，将从同一文本中抽取的段落组合到一起，形成文档作为训练集。测试集的建立方式同训练集，通过 LDA 模型对文本进行训练建模，利用训练集训练出不同主题的词频分布，然后将所抽取的段落表示为主题分布后进行分类，验证分类结果。以下验证考虑了不同主题个数 T、不同段落长度 K 以及以“词”和“字”为基本单元下分类对验证结果的影响。

Introduction

LDA（Latent Dirichlet Allocation）是一种用于文本数据主题建模的概率生成模型。它是由 Blei、Ng 和 Jordan 于 2003 年提出的。LDA 可以从文档集合中发现隐藏的主题，并且将每个文档表示为这些主题的概率分布，其背后的核心假设是文档中的每个单词都是由一些主题生成的，而主题又是由单词的分布组成的。

LDA 的基本思想是将文档集合中的每个文档表示成一个主题的概率分布，而每个主题又表示成一个单词的概率分布。具体来说，LDA 模型的生成过程包括以下三个步骤：

1）对一篇文档的每个位置，从主题分布中抽取一个主题；

2）从上述被抽到的主题所对应的单词分布中抽取一个单词；

3）重复上述过程直至遍历文档中的每一个单词。

本文所要研究的问题包括：从给定的语料库中均匀抽取 1000 个段落作为数据集（每个段落可以有 K 个 token, K 可以取 20，100，500, 1000, 3000），每个段落的标签就是对应段落所属的小说。利用 LDA 模型在给定的语料库上进行文本建模，主题数量为 T，并把每个段落表示为主题分布后进行分类（分类器自由

选择），分类结果使用 10 次交叉验证（i.e. 900 做训练，剩余 100 做测试循环十次）。实现和讨论如下的方面：

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？

2）以"词"和以"字"为基本单元下分类结果有什么差异？

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？

Methodology

LDA 的基本思想是将文档集合中的每个文档表示成一个主题的概率分布，而每个主题又表示成一个单词的概率分布。具体来说，LDA 模型的生成过程包括以下三个步骤：

1.对一篇文档的每个位置，从主题分布中抽取一个主题；

2.从上述被抽到的主题所对应的单词分布中抽取一个单词；

3.重复上述过程直至遍历文档中的每一个单词。

具体实现过程如下：

1.中文语料库预处理。对语料库中文本进行训练前需要删除开头无用信息，删除中文停词及标点符号，删除空白符号（空格、换行符等），同时利用 jieba 分词对文本进行分词处理。

2.段落抽取。根据每篇文本的字\词数占总语料库字\词数的比例确定该篇文本应抽取的段落数，即比例权重抽取段落，将抽取的段落以字和词形式以及训练集和测试集的形式分类保存。

3.模型训练。将从步骤 2 中抽取的所有段落组合到一起作为训练集，进行训练。首先为每篇文章中的词随机分配一个 topic，然后统计每篇文章的 topic 频率以及每个 topic 的词频。再计算每个 topic 下的词出现在这个位置的概率，然后更改该词的 topic 为概率最大的 topic，随后进行文章 topic 频率以及 topic 词频的更新，当频率无变化时，说明算法已经收敛，迭代终止。

4.分类测试。将 K 个段落利用训练好的 topic 词的分布计算段落的主题分布，并且利用欧氏距离选择该段落与哪一个样例更加接近，便认为该段落属于这一类。

Experimental Studies

设置段落的 token 分别为 K＝20、100、500、1000、3000，分类的 topic 分别为 T＝10、50、100，再分别以字和词进行验证。由于当 K 取 20 时，计算频率时出现 INF，运算时间过载，故无此类数据。以“字”为基本单元验证准确率结果如表 1 所示，以“词”为基本单元验证准确率结果如表 2 所示。

表 1 分类准确率（以字为单元）

T \ K	20	100	500	1000	3000
10		0.1373	0.2059	0.3137	0.4314
50		0.1961	0.5392	0.6471	0.7745
100		0.2353	0.6373	0.7549	0.8333

表 2 分类准确率（以词为单元）

T \ K	20	100	500	1000	3000
10		0.1176	0.2353	0.2843	0.4804
50		0.1863	0.5196	0.5882	0.8627
100		0.2059	0.6078	0.6765	0.9314

由上述表格数据分析可得：当设定不同的 topic 时，即 T 取不同值时，分类性能不同，随着 T 值的不断增大，分类的准确率在不断提高；当设定不同的 token 时，即 K 取不同值时，分类性能差距明显，对分类准确率的影响比改变 T 的取值影响大，随着 K 值的不断增大，分类的准确率也在不断提高；分别以字和词两种单元分类，整体来看，字的分类准确率比词的分类准确率高，但这体现在 K 的取值不大（K＜1000）时，当 K 的取值为 3000 时，则出现词的分类准确率比字

的分类准确率高的情况。

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[2]https://blog.csdn.net/weixin_44966965/article/details/124556948