基于中文语料库的 LDA 模型分类验证

ZY2303118　Yisong Wang
2740416657@qq.com

Abstract

Latent Dirichlet Allocation（LDA）是一种用于主题建模的概率图模型。它的基本思想是，每个文档是由一组主题混合而成的，每个主题又由一组词汇构成。LDA 试图找到最佳的主题和词汇组合，以解释给定的文本数据。也可以这么理解一篇文章的生成：先以一定的概率选取某个主题，然后再以一定的概率选取该主题下的某个词，不断重复这两步，直到完成整个文档。而 LDA 解决的问题就是，分析给定的一篇文章都有什么主题，每个主题出现的占比大小是多少。

Introduction

基于给定的中文语料库（金庸 16 本武侠小说），本文首先从中均匀抽取 1000 个段落作为数据集，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。段落表示为主题分布后使用分类器进行分类。最后采用 10 次交叉验证。实现和讨论如下的方面：（1）在设定不同的主题个数 T 的情况下对分类性能影响；（2）以"词"和以"字"为基本单元下对分类结果影响？（3）不同的取值的 K 的短文本和长文本对主题模型性能的影响？

Methodology

实验参数选择如下：

- 每个段落的 token 数 K 取 20，100，500，1000，3000
- T 取 50，100，200
- 分类器选择为 SVM
- 前 900 个段落数据集做训练，剩余 100 段落数据集做测试循环十次

Experimental Studies

根据上述理论和方法，若采用按“词”为基本单元，得到以下结果：

表 1: 按“词”分类结果

K	T	Classifier	type	Training accuracy	accuracy
20	50	SVM	word	0.183333333	0.14
20	100	SVM	word	0.175555556	0.13
20	200	SVM	word	0.178888889	0.14
100	50	SVM	word	0.454444444	0.43
100	100	SVM	word	0.510000000	0.48
100	200	SVM	word	0.517777778	0.53
500	50	SVM	word	0.826666667	0.80
500	100	SVM	word	0.831111111	0.83
500	200	SVM	word	0.743333333	0.81
1000	50	SVM	word	0.918888889	0.93
1000	100	SVM	word	0.908888889	0.91
1000	200	SVM	word	0.847777778	0.84
3000	50	SVM	word	0.996888889	0.99
3000	100	SVM	word	0.975555556	1
3000	200	SVM	word	0.962222222	1

若采用按“字”为基本单元，得到以下结果：

表 2: 按“字”分类结果

K	T	Classifier	type	Training accuracy	accuracy
20	50	SVM	char	0.077777778	0.08
20	100	SVM	char	0.075555556	0.08
20	200	SVM	char	0.124444444	0.11
100	50	SVM	char	0.123333333	0.07
100	100	SVM	char	0.126666667	0.11
100	200	SVM	char	0.141111111	0.08
500	50	SVM	char	0.166666667	0.16
500	100	SVM	char	0.173333333	0.18
500	200	SVM	char	0.152222222	0.20
1000	50	SVM	char	0.162222222	0.12
1000	100	SVM	char	0.163333333	0.14
1000	200	SVM	char	0.390000000	0.33
3000	50	SVM	char	0.420000000	0.46
3000	100	SVM	char	0.813333333	0.89
3000	200	SVM	char	0.711111111	0.66

观察分析可知，对比不同的 K 值，也就是不同的 token 数，随着 K 的增长，分类性能变得越来越好，且影响比其他参数更大，对于本案例来说，当 K 为 1000 以上时，准确率可以达到 90%以上；对于主题数 T 而言，可以发现不宜过低和过高，在取中间适当值时可以取得不错的分类性能；而对于以“词”“字”两种基本单元下进行分类，可以发现，以“字”为基本单位时，准确率明显降低，由此可见，LDA 对短文本的主题分类效果比较差。

References

[1] 梦家.齐夫定律(Zipf'law)理论及其应用场景[EB/OL].2022:[2024-4-7].https://dreamhomes.top/posts/202204221003/.

[2] historyasamirror.Zipf's law[EB/OL].2008:[2024-04-07].https://blog.csdn.net/historyasamirror/article/details/3125223.

[3] Kanglei Zhou.中文平均信息熵[EB/OL].2021:[2024-04-07].https://kangleizhou.github.io/nlp/2021/04/08/Chinese_entropy/.

[4] Brown P F, Della Pietra S A, Della Pietra V J, et al. An estimate of an upper bound for the entropy of English[J]. Computational Linguistics, 1992, 18(1): 31-40.