Report of Deep Learning for Natural Langauge Processing

Haolei Zhang
2531563133@qq.com

Abstract

本实验旨在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）。依次验证如下的方面：（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？。

Introduction

潜在狄利克雷分配（LDA）是一种文本挖掘工具，用于从文档集合中发现主题。在自然语言处理领域，LDA 被广泛应用于主题模型构建，它可以揭示大量文档集合中隐藏的、抽象的“主题”结构。

LDA 基于以下假设：

每篇文档是由多个主题的混合而成的。每个主题则是由多个词汇的混合而成的。

LDA 模型的数学表达为：

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \beta)$$

其中：

(θ)是文档中主题的分布。

$(\mathbf{z})$是文档中词汇对应的主题。

$(\mathbf{w})$是文档中的词汇。

(α)和(β)是模型的超参数，分别控制主题分布和词汇分布的形状。

Methodology

LDA（隐含狄利克雷分配）模型的基本公式涉及到以下几个关键参数：

文档到主题的分布：每个文档(d)被表示为一个主题分布(θ_d)，它是由一个狄利克雷分布$\big(Dir(\alpha)\big)$生成的，其中(α)是分布的参数。

主题到词汇的分布：每个主题(k)被表示为一个词汇分布(ϕ_k)，它也是由一个狄利克雷分布$\big(Dir(\beta)\big)$生成的，其中(β)是分布的参数。

词汇：文档中的每个词(w)是由文档的主题分布(θ_d)和主题的词汇分布(ϕ_k)共同决定的。

LDA 模型的目标是推断出文档集合中的(θ_d)和(ϕ_k)。这通常通过迭代算法如吉布斯采样或变分贝叶斯方法来实现。

具体的 LDA 模型可以表示为以下数学公式：

对于每个主题(k)，选择$\big(\phi_k \sim Dir(\beta)\big)$

对于每个文档(d)，选择$\big(\theta_d \sim Dir(\alpha)\big)$

对于文档中的每个词(n)：选择一个主题$\Big(z_{d,n} \sim Multinomial(\theta_d)\Big)$，选择一个词$\Big(w_{d,n} \sim Multinomial\big(\phi_{z_{d,n}}\big)\Big)$

在这里，(Dir)表示狄利克雷分布，$(Multinomial)$表示多项分布。

LDA 模型的参数调整和优化对于模型的性能至关重要。常见的参数包括：

(α)和(β)：控制文档-主题分布和主题-词汇分布的狄利克雷先验。

主题数量：需要预先设定的主题的数量。

迭代次数：算法运行的迭代次数。

词汇过滤：在模型训练前过滤掉频率过低或过高的词汇。

这些参数的选择和调整需要根据具体的应用场景和数据集来进行。

Multinomial Naive Bayes（多项式朴素贝叶斯）是朴素贝叶斯分类器的一种变体，通常用于文本分类问题。它假设特征（词）的分布是多项式分布，并且在文本分类中通常表现良好。

Multinomial Naive Bayes 基于贝叶斯定理和特征条件独立性假设。在文本分类问题中，假设给定一个文档，每个特征（词）的出现概率与其他词的出现概率是独立的。根据这个假设，可以使用贝叶斯定理计算文档属于每个类别的概率，然后选择概率最高的类别作为预测结果。

Experimental Studies

基本单元	T	K	准确率
字	50	20	0.79
	50	100	0.83
	50	500	0.87
	50	1000	0.92
	50	3000	0.94
	10	3000	0.86
词	50	20	0.89
	50	100	0.91
	50	500	0.94
	50	1000	0.95
	50	3000	0.97
	10	3000	0.92

Conclusions

从 结果可以看出，当主题数 T 和 token 取值 K 相同时，以词为基本单元的主题模型性能优于以字为基本单元的主题模型。随着 K 值的增加，不同基本单元的主题模型性能变化呈现出相同的趋势，其分类结果的准确度也随之增加，主题模型表现出更好的性能。说明长文本相较于短文本的独特性更高，使用长文本进行训练能够提取出更为有效的文本主题。当 K 固定时，随着上升到，测试集准确度稳定上升，分类模型性能逐渐提高。