Report of Deep Learning for Natural Langauge Processing

Jiale Wang
hirw@qq.com

Abstract

本文通过LDA主题模型对金庸小说语料库中的文本进行了建模与分类实验，探讨了不同段落长度、主题数和文本基本单元对分类性能的影响。实验从语料库中均匀抽取了1000个段落，段落长度分别为20、100、500、1000和3000个token，并采用“词”和“字”两种基本单元进行实验。使用LDA模型对文本进行建模，并利用随机森林分类器对每个段落的主题分布进行分类。通过10次交叉验证评估分类性能。实验结果显示，随着段落长度的增加，分类准确率逐渐提高。例如，在段落长度为3000个token时，准确率达到了最高值0.7458（以字为基本单元，主题数100），而段落长度为20时，准确率最低，仅为0.0524（以字为基本单元，主题数5）。此外，在长文本的情况下，通常使用“词”作为基本单元会相对更优，而在短文本中，“字”表现得更好。总体来看，较长段落和较高主题数（如50、100）对分类性能有显著提升作用。

Introduction

文本分类是自然语言处理中的一项核心任务，LDA（Latent Dirichlet Allocation）作为一种经典的主题建模方法，已广泛应用于各类文本分析任务。本研究基于金庸小说语料库，探讨了LDA模型在文本分类中的表现，并重点分析了段落长度、主题数及文本基本单元对分类效果的影响。通过实验，旨在为基于主题模型的文本分类任务提供经验指导。

Methodology

Step1: 数据集与预处理

实验使用了金庸小说语料库。通过对文本进行清理，移除非中文字符和标点符号，然后根据基本单元（“词”或“字”）对文本进行分词或字切分。实验从每本小说中均匀抽取1000个段落，段落长度分别设定为20、100、500、1000和3000个token，以探索不同段落长度对分类性能的影响。

Step2: LDA主题建模

LDA模型通过无监督学习对文本进行主题建模，将每个文档表示为多个潜在主题的概率分布。在实验中，我们选择了不同的主题数（5、10、20、50、100）进行建模，以探讨主题数对分类性能的影响。

Step3: 分类方法与实验设计

分类实验使用了随机森林分类器（Random Forest Classifier）。为了确保实验结果的可靠性，采用了10次交叉验证，每次使用900个段落进行训练，100个段落进行测试，最终计算分类准确率和标准差。

实验围绕以下几个问题进行：

（1）不同主题数（T）是否影响分类性能？

（2）以“词”和“字”作为基本单元下，分类结果是否存在差异？

（3）不同段落长度（K）对短文本和长文本的主题建模效果是否有差异？

Experimental Studies

Step1: 试验结果

经过对设计程序进行运行，得到结果如下：

表1: 字为单元的准确率统计

段落长度(K)	主题数(T)	基本单元	准确率	标准差
20	5	word	0.052444444444444446	0.016222914057476664
20	10	word	0.05448484848484849	0.016501922233617107
20	20	word	0.04638383838383839	0.01373832438778402
20	50	word	0.05440404040404041	0.010080706867773922
20	100	word	0.060444444444444446	0.021035881618748774
100	5	word	0.07660606060606061	0.024385570641355523
100	10	word	0.1634141414141414	0.03723799536969126
100	20	word	0.15327272727272725	0.026101103918272767
100	50	word	0.15123232323232325	0.03328606382090435
100	100	word	0.12105050505050503	0.027345626477000874
500	5	word	0.27026262626262626	0.04163545693708381
500	10	word	0.37492929292929295	0.04903277931758903
500	20	word	0.4534949494949495	0.061762753258186524
500	50	word	0.4898989898989899	0.051505247612048326
500	100	word	0.4656363636363636	0.06228517083437076
1000	5	word	0.3020833333333333	0.03423265984407288
1000	10	word	0.5375000000000001	0.0473242362150023
1000	20	word	0.6104166666666667	0.06236095644623236
1000	50	word	0.6520833333333333	0.04039733214513607
1000	100	word	0.6125	0.046304397439360157
3000	5	word	0.48730506155950754	0.04696873634568113
3000	10	word	0.7319288645690835	0.014762015478451351
3000	20	word	0.7761422708618332	0.0295554787470024
3000	50	word	0.8297674418604652	0.0337778188937029
3000	100	word	0.7273324213406293	0.05096763734668537

表2: 词为单元的准确率统计

段落长度(K)	主题数(T)	基本单元	准确率	标准差
20	5	char	0.08880808080808081	0.040702253017166026
20	10	char	0.10080808080808082	0.006401203845557516
20	20	char	0.08664646464646467	0.03633042254201595
20	50	char	0.052444444444444446	0.023427933640982012
20	100	char	0.08472727272727273	0.0261422571796682
100	5	char	0.11696969696969697	0.022823812917169 46
100	10	char	0.08676767676767676	0.023715657377704022
100	20	char	0.10282828282828282	0.01961157349416016
100	50	char	0.09274747474747476	0.022515276134715916
100	100	char	0.14723232323232324	0.025488001102158927
500	5	char	0.238	0.028202222800348967
500	10	char	0.4011919191919192	0.06451224703376228
500	20	char	0.485979797979798	0.04751790000983588
500	50	char	0.4393737373737373	0.050678506180378134
500	100	char	0.46379797979797976	0.05811393346109152
1000	5	char	0.4104166666666666	0.024295632895188754
1000	10	char	0.5354166666666667	0.07412686197773831
1000	20	char	0.5520833333333333	0.050603990839546316
1000	50	char	0.6125	0.06123724356957945
1000	100	char	0.5791666666666667	0.058034951154933845
3000	5	char	0.4846785225718194	0.048754178038090804
3000	10	char	0.7038850889192886	0.03329715712177072
3000	20	char	0.7669767441860464	0.03633865113418535
3000	50	char	0.7435294117647059	0.04743079850270367
3000	100	char	0.745827633378933	0.0489896726165631

从表1和表2中可以看出，段落长度、主题数以及基本单元的选择对准确率有显著影响。

首先，段落长度（K）对分类结果影响较大。随着段落长度的增加，准确率显著提升。例如，在以“字”作为基本单元的情况下，当段落长度从20增加到3000时，准确率从0.0888提升至0.7458。这表明，较长的段落提供了更多的上下文信息，有助于LDA模型更好地提取潜在主题，进而提升分类性能。

其次，主题数（T）的增加也对准确率有明显的提升作用，尤其是在较长段落（如1000和3000个token）中。例如，当段落长度为3000个token时，主题数从5增加到100，准确率从0.4873提高到0.7458，表明增加主题数能够帮助更好地细化文本的主题结构，增强分类性能。

再次，基本单元的选择对分类结果也有显著影响。与“词”作为基本单元的情况相比，在长文本情况下，使用“词”作为基本单元的分类结果普遍更好。在段落长度较短（如K=20）的情况下，字单元的优势尤其明显。例如，在段落长度为20个token时，“字”作为基本单元

的准确率为0.0888，而“词”作为基本单元的准确率仅为0.0524。

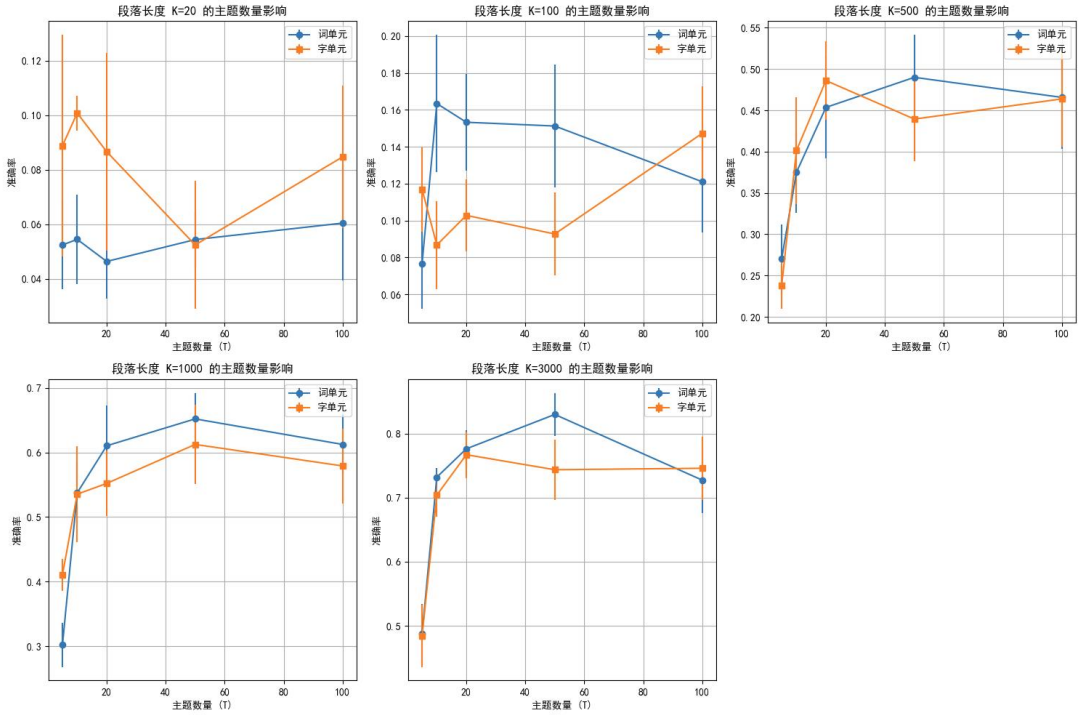

图1: 不同段落长度准确率统计图

图1是对比不同段落长度（如K=20, K=100, K=500, K=1000, K=3000）时，主题数（T）变化对准确率的影响。每个图中，两条曲线代表“字”作为基本单元和“词”作为基本单元的分类结果。在长文本的情况下，通常使用“词”作为基本单元会相对更优，而在短文本中，“字”表现得更好。

具体分析如下：

1. 长文本中“词”优于“字”的现象

在图表中，随着段落长度的增加，特别是当段落长度为1000或3000个token时，可以看到使用“词”作为基本单元的准确率通常较高。这是因为长文本提供了更丰富的上下文信息，使用“词”能够更好地捕捉到更长范围内的语义结构。相对于“字”，“词”能够更准确地反映出文本的主题和含义，特别是在长文本中，分词能够将常见词语合并为更具代表性的特征，从而增强模型对文本的理解和分类能力。例如，在段落长度为1000时，使用“词”作为基本单元的准确率显著高于使用“字”的情况，这表明随着文本长度的增加，词汇的层次感更能提升分类效果。

2. 短文本中“字”优于“词”的现象

相比之下，在较短的段落（例如段落长度为20个token）中，使用“字”作为基本单元的分类性能通常较好。这是因为短文本提供的信息较少，无法充分利用长文本中常见的词汇结构，而“字”可以更加细致地捕捉到文本中的局部信息。因此，在短文本中，“字”作为单元能够更好地描述文本的局部特征，提供更多细节上的信息，进而提升模型的分类能力。

综上所述，本实验结果表明，段落长度、主题数和基本单元的选择对LDA模型在文本分类任务中的表现具有重要影响。在实际应用中，选择合适的参数设置对于提升分类准确率至关重要。

Conclusions

通过对金庸小说语料库进行LDA主题建模与分类实验，我们得出以下结论：

主题数（T）的增加通常能够提高分类性能，尤其是在主题数达到50或100时，分类准确率显著提升。例如，在段落长度为1000个token时，主题数为100时的准确率为0.6125（以词为基本单元），而在3000个token时，主题数为100时的准确率为0.7458（以字为基本单元）。在长文本的情况下，通常使用“词”作为基本单元会相对更优，而在短文本中，“字”表现得更好。段落长度（K）对分类结果有重要影响，较长的段落（如1000、3000个token）提供了更多上下文信息，导致分类性能更为稳定，准确率更高。例如，段落长度为3000时，以字为基本单元，分类准确率可达到0.7458。

本文的实验结果表明，LDA模型在文本分类任务中具有良好的应用前景，特别是在处理长文本时，模型能够有效提取潜在的主题信息，从而提升分类性能。未来的研究可以进一步探讨更复杂的模型和其他文本特征对分类性能的影响。