Report of Deep Learning for Natural Langauge Processing

Zengchang Qin
zengchang.qin@gmail.com

摘要

本实验从本地中文小说语料库中随机均匀抽取1000个段落，并对每个段落按token数截断为不同长度（K取20、100、500、1000、3000），构建数据集。实验分别以“词”（采用jieba分词）和“字”（逐字符拆分）为基本单位进行预处理；利用LDA模型生成主题分布向量，再使用SVM分类器进行小说标签预测，采用10折交叉验证评估分类性能。评价指标包括分类准确率、精确率、召回率、F1分数及混淆矩阵。本文对不同文本长度、分词单位及主题数T（候选值10、20、30、50）对分类性能的影响进行了详细的定量分析。

简介

中文小说文本由于无明显词界、格式多样等特点，在文本预处理、主题建模和分类任务中存在一定难度。为此，本实验首先对语料库进行分段、分词、停用词过滤和文本截断处理，然后利用LDA模型提取每个段落的主题分布向量，最后以SVM分类器对小说标签进行预测。实验重点探究三个变量对分类效果的影响：即文本长度K、分词单位、和主题数量T。

理论方法

There are models of my research.

M1: LDA Model

本文使用的Latent Dirichlet Allocation (LDA) [2]是一种生成式概率主题模型，其目的是从大量文档中自动发现隐含的主题结构。该模型假设每个文档由若干主题构成，而每个主题则对应于一个词分布。具体而言，LDA 的数学过程可以描述如下：

首先，对于每个文档，假定其主题分布 θ 是从一个 Dirichlet 先验分布中抽取的；接着，对于文档中的每个词，先从文档的主题分布中抽取一个主题 z，然后再从该主题对应的多项式分布 φ 中生成一个词。这里，主题-词分布 φ 同样假定服从 Dirichlet 先验分布。整个模型通过联合分布 $p(\theta, z, w \mid \alpha, \beta)$ 来描述文档生成过程，其中 α 和 β 分别是控制文档主题分布和主题词分布稀疏性的超参数。对于后验分布，LDA 模型利用 Dirichlet 分布与多项式分布之间的共轭性质，简化了参数更新的数学推导，从而使得模型能够有效地从数据中提取出每个文档的主题分布向量，作为后续文本分类或其他任务的低维语义特征表示。

实验过程

T本报告的整体方法可分为以下几步：

文本预处理

1. 数据来源与段落抽取： 语料库存放于指定文件夹中，其中每个txt文件代表一部小说，文件名（不含扩展名）作为小说标签。文件中各段落以“全角空格空格+换行”作为分隔符；同时，文件夹中存在记录文档标题的inf.txt文件，将其排除。为应对编码问题，程序首先尝试使用utf-8编码读取，若失败则采用gbk编码并忽略错误。
2. 文本预处理： 对抽取到的段落进行两种预处理流程：

以“词”为基本单元：利用jieba进行中文分词，再使用预设停用词列表过滤无效词项。

以“字”为基本单元：将文本中每个汉字作为一个token，不进行额外分词。

随后，根据不同K值（20、100、500、1000、3000）截取前K个token，构建不同长度的文本样本。

主题建模

采用LDA模型对预处理后的文本数据进行建模。具体步骤为：构建gensim字典和语料库；训练LDA模型，模型的主题数量T（10、20、30、50）、迭代次数、α和β参数均以参数形式提供，便于后续调优；利用训练好的模型，将每个段落转换为主题分布向量，作为后续SVM分类的特征表示。

分类实验

采用SVM分类器对每个段落的主题分布向量进行分类。分类实验采用10折交叉验证（每折900个样本训练、100个样本测试），并记录准确率、精确率、召回率和F1分数等指标。

结果分析

实验结果汇总：

	token_mode	K	num_topics	accuracy	precision	recall	f1_score
0	word	20	10	0.197	0.030143	0.067588	0.036451
1	word	20	20	0.189	0.035610	0.068583	0.045203
2	word	20	30	0.188	0.031204	0.065474	0.040504
3	word	20	50	0.182	0.053167	0.068248	0.050804
4	word	100	10	0.196	0.038564	0.070102	0.046784
5	word	100	20	0.179	0.033707	0.063787	0.041363
6	word	100	30	0.177	0.040769	0.065688	0.047287
7	word	100	50	0.210	0.054038	0.081265	0.060518
8	word	500	10	0.190	0.041769	0.069036	0.045592
9	word	500	20	0.205	0.039333	0.074723	0.050508
10	word	500	30	0.200	0.045392	0.073008	0.050773
11	word	500	50	0.167	0.036104	0.061602	0.043893
12	word	1000	10	0.216	0.042875	0.078790	0.052780
13	word	1000	20	0.203	0.036029	0.071582	0.043767
14	word	1000	30	0.205	0.054777	0.077099	0.057491
15	word	1000	50	0.203	0.060456	0.079713	0.061871
16	word	3000	10	0.195	0.030006	0.067473	0.038073
17	word	3000	20	0.201	0.039124	0.073481	0.048868
18	word	3000	30	0.203	0.036936	0.071648	0.045856
19	word	3000	50	0.165	0.049164	0.062585	0.046895
20	char	20	10	0.185	0.038374	0.067418	0.043745
21	char	20	20	0.224	0.041625	0.079652	0.051226
22	char	20	30	0.203	0.047607	0.078013	0.056370
23	char	20	50	0.183	0.060017	0.072179	0.057047
24	char	100	10	0.218	0.036571	0.076587	0.045575
25	char	100	20	0.198	0.034606	0.069729	0.043499
26	char	100	30	0.203	0.049653	0.075790	0.054746
27	char	100	50	0.198	0.045110	0.072428	0.050425
28	char	500	10	0.202	0.033862	0.071961	0.044619
29	char	500	20	0.192	0.039286	0.069619	0.047331
30	char	500	30	0.215	0.055425	0.080507	0.059548
31	char	500	50	0.196	0.069447	0.083180	0.067759
32	char	1000	10	0.213	0.054073	0.082300	0.061638
33	char	1000	20	0.211	0.055818	0.080997	0.062045
34	char	1000	30	0.195	0.047274	0.073958	0.054561
35	char	1000	50	0.208	0.054353	0.078724	0.058416
36	char	3000	10	0.231	0.051689	0.104369	0.066892
37	char	3000	20	0.196	0.065644	0.091028	0.073762
38	char	3000	30	0.198	0.061931	0.082538	0.066406
39	char	3000	50	0.207	0.075499	0.091066	0.073478

图 1：实验结果

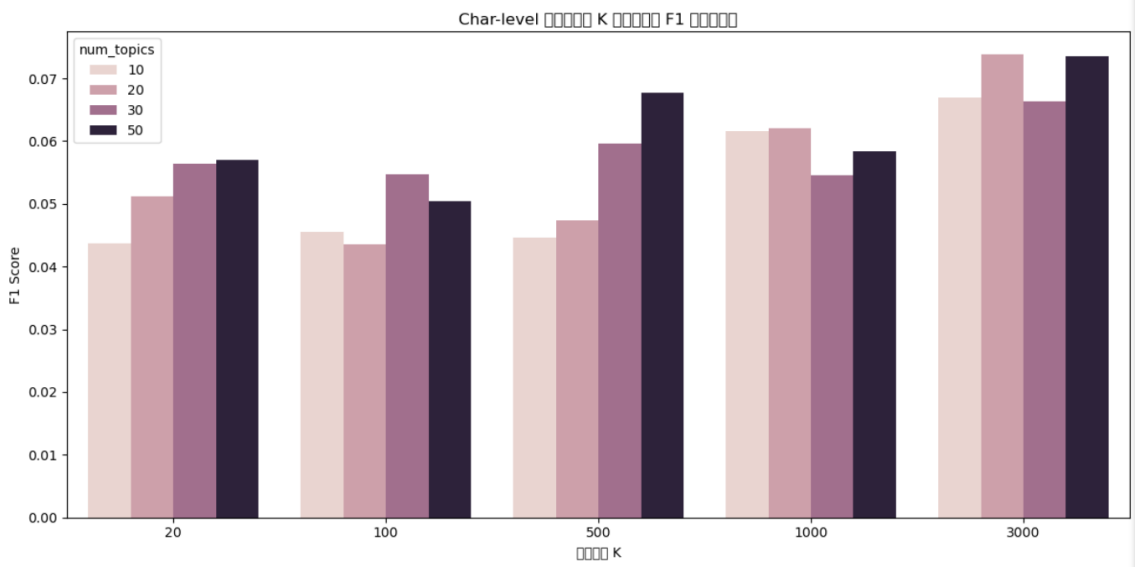

图 2：Char-level 分词下不同 K 与主题数对 F1 分数的影响

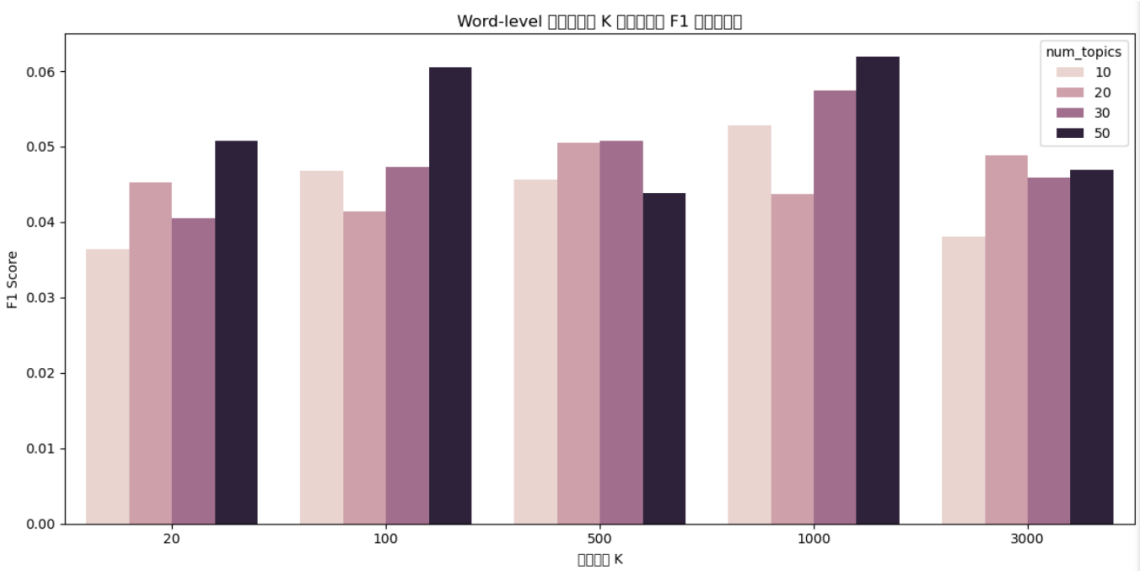

图 3：Word-level 分词下不同 K 与主题数对 F1 分数的影响

由实验结果可以看出，在短文本情况下（K=20、100），整体准确率及F1分数较低，这表明较短文本难以提供充分语义信息，导致LDA模型提取的主题分布不够稳定，进而影响SVM分类性能。而在中长文本情况下（K=500、1000）：实验数据显示，在“word”模式下，K=1000时准确率达到约0.216，F1分数在0.052～0.062之间；而在“char”模式下，表现略有提升，部分配置准确率在0.211~0.213左右。当长文本情况下时（K=3000），部分配置在“char”模式下显示较高的准确率（最高0.231），但在“word”模式下，K=3000的效果反而略有下降（例如，当T=50时准确率降至0.165）。这可能表明，文本过长可能引入噪声或冗余信息，影响主题模型效果。

至于分词单位对准确率的影响，整体上，字符模式在部分配置下取得了更高的准确率（例如K=3000, T=10时达到0.231），而词模式下的表现相对稳定，但准确率略低。从F1分数看，虽然两种模式的数值相近，但在某些主题数配置下，“char”模式能略微提高召回率，可能是因为逐字符统计能捕捉到更多细粒度的语言信息，但也可能带来高维稀疏性的问题。

在相同的K和token_mode下，不同主题数对分类性能影响明显。例如，在“word”模式下，K=100时，T=50的F1分数为0.0605，而T=10、20、30的F1均低于此值；但在其他文本长度下，最佳T值并不完全一致，说明主题数的选择需结合具体数据情况进行调优。综合来看，适中的主题数量（如20～30）似乎能在一定程度上平衡主题表达能力与噪声引入，获得较优的分类效果。

从实验结果来看，无论采用哪种分词模式及文本长度，整体分类准确率均在0.160.07之间。虽然数值较低，但这可能与数据集的多类别、样本噪声以及文本本身的复杂性有关。此外，实验中只使用了经典的SVM分类器进行分类，导致实验代码在高维主题分布向量上的表现还存在较大提升空间，未来可考虑更复杂的分类模型或特征降维方法。

根据生成的柱状图可以看出，随着K值的增加，字符模式下的F1分数呈现一定上升趋势，但在K过大时趋于平稳甚至略降；词模式下不同T值间的差异较为明显，最佳T值约在50左右，但整体波动不大；两种分词模式下，F1分数的波动范围相近，但字符模式在部分长文本配置下略有优势。

结论

本实验通过对中文小说段落进行预处理、LDA主题建模和SVM分类，探讨了文本长度、分词单位和主题数量三个变量对分类性能的影响。结果表明：

- 较短的文本（K=20、100）由于信息量不足，分类性能低下；中长文本（K=500、1000）能较好地提取主题信息，而过长文本（K=3000）则可能引入多余噪声；
- 相较于以“词”为基本单元的预处理，以“字”为基本单元在部分实验条件下取得了稍高的准确率和F1分数，但两者之间的差距不大；
- 主题数量T的选择对分类效果有显著影响，适中的主题数（例如20～30）较能平衡语义表达与噪声控制。

References

[1] Zenchang Qin and Lao Wang (2023), How to learn deep learning? Journal of Paper Writing, Vol. 3: 23: pp. 1-12.

[2] LDA主题模型及Python实现_python lda-CSDN博客