Report of Deep Learning for Natural Langauge Processing

Zhijin Cao
21231024@buaa.edu.cn

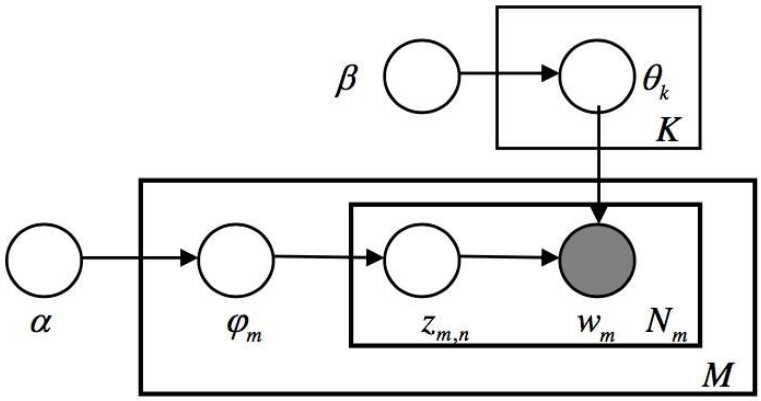
Abstract

文本分类是自然语言处理中的一个重要任务，而主题模型在提取文本特征方面具有显著优势。本文基于LDA主题模型对文本进行建模，将每个段落表示为主题分布。随后，利用SVM分类器基于主题特征对段落进行分类，通过分类结果探究不同参数下主题模型的性能。具体而言，我们关注以下三个问题：（1）不同主题数量T对分类性能的影响；（2）以“词”和以“字”为token时分类结果的差异；（3）不同文本长度K下主题模型性能的变化。

Introduction

Latent Dirichlet Allocation

Latent Dirichlet Allocation（LDA）是一种生成概率模型，用于自然语言处理和文本特征提取。这一模型的基本思想是：文档可以表示为主题的分布，而主题可以表示为词语的分布。下图为LDA模型示意图。



在上图中，K 为主题的数量，M 为文档的数量，N_m 为文档 m 中的词语数量，α 为每个文档主题分布的狄利克雷先验参数，β 为每个主题词语分布的狄利克雷先验参数。

在LDA模型下，对于文档 m 一篇文档的生成过程如下：

1. 抽取一个文档主题分布 $\varphi \sim Dir(\alpha)$ 。
2. 抽取所有主题词语分布 $\theta \sim Dir(\beta)$

3. 对于文档 m 中的每个词语：

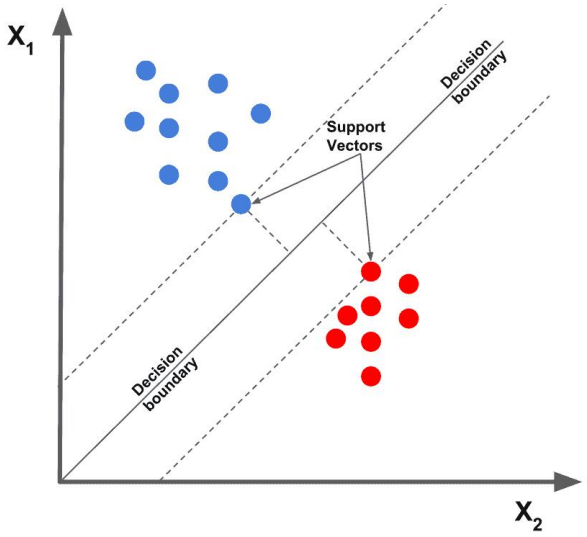
(1). 从多项式分布 φ_m 中抽取一个主题 $z_{m,n}$。

(2). 依据 $z_{m,n}$ 从 θ_k 中抽取一个词语 w_m

一般采用吉布斯采样法或变分法对LDA中的参数进行拟合。

SVM模型

SVM（Support Vector Machine，支持向量机）是一种强大的分类模型，主要用于二分类问题。它通过在特征空间中寻找一个最优的超平面来区分不同类别的数据点。其核心思想是采用最大间隔原理，这里间隔是指超平面到最近的数据点（支持向量）的距离。通过最大化间隔，SVM对于分类问题有较好的鲁棒性。下图为SVM模型示意图。



Methodology

实验思路

本文从语料库中均匀抽取1000个段落，每个段落的长度K有20、100、500、1000和3000共5种类型。token有char字和word词2种类型，在不同情况下对段落进行抽取。我们使用LDA模型对文本进行建模，主题数量T有5、10、15、20、25共5种类型。随后，将每个段落表示为主题分布，并输入到分类器中进行分类。分类器采用SVM，使用10次交叉验证（每次900个样本用于训练，100个样本用于测试）来评估分类性能。

实验数据

该实验中文数据集为16本金庸小说，具体内容见data，其中data/inf.txt中存储了16本小说的小说名，其余txt文件文件名为小说名且文件内容为小说内容。对于划分段落，在inf.txt

文件中的前8本小说每本抽取62段，后8本小说每本抽取63段。为了保证抽取的段落在小说中均匀分布，本文设置抽取起始点的位置均匀分布在整本小说中。

代码解析

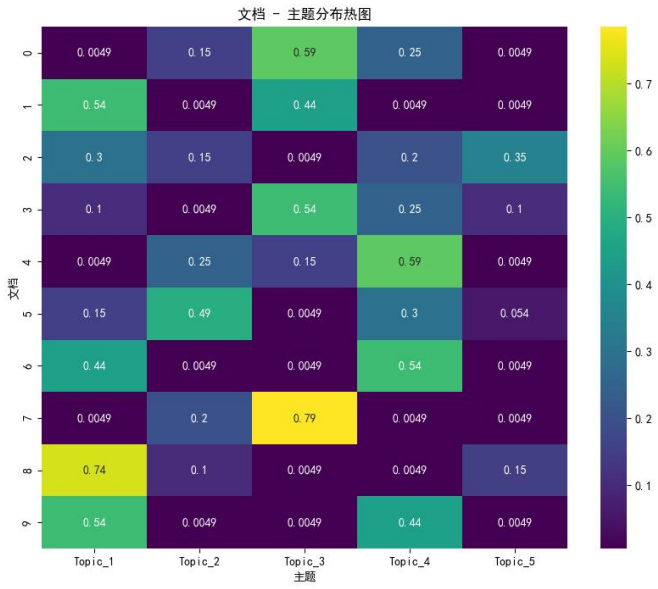
该实验的主函数主要分为加载文档数据all_docs，获取文档-单词频率矩阵df，LDA建模并分类进行交叉验证三个部分（这部分在主函数中）。

加载文档数据all_docs部分，其内容包括将文本数据按路径加载，去除无用内容，按模式对数据进行拆分，将拆分后的数据进行筛选并去除停词。

获取获取文档-单词频率矩阵df部分，其内容包括根据加载好的docs找出1000段文档中的全部单词表并对每个文档中单词表频次进行统计。

Experimental Studies

运行代码，在运行LDA模块后，可以查看文档-主题分布，将该分布映射到热力图上，具体如下图所示（只展示前十个文档的主题分布情况）



代码运行完成后，从result.txt中可以看到实验结果，如下表所示

Table 1: Mode为char时，不同长度K和主题数T下分类准确率

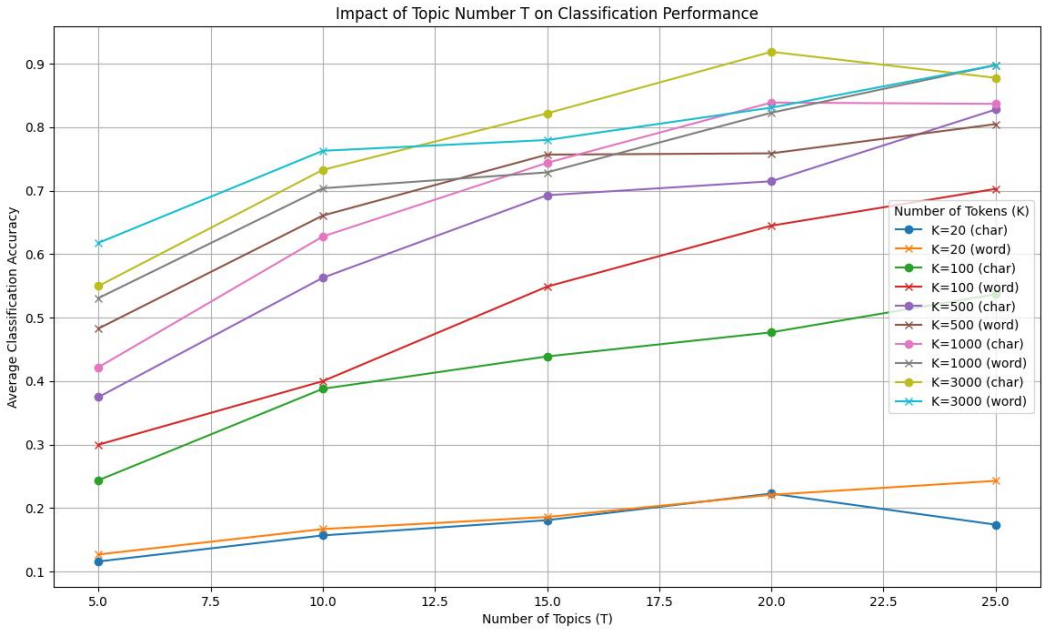
平均分类准确率	T=5	T=10	T=15	T=20	T=25
K=20	0.116	0.157	0.181	0.223	0.174
K=100	0.244	0.388	0.439	0.477	0.537
K=500	0.375	0.563	0.693	0.715	0.828
K=1000	0.422	0.628	0.744	0.839	0.837
K=3000	0.55	0.733	0.822	0.919	0.878

Table 2 : Mode为word时，不同长度K和主题数T下分类准确率

平均分类准确率	T=5	T=10	T=15	T=20	T=25
K=20	0.127	0.167	0.186	0.221	0.243
K=100	0.300	0.400	0.549	0.645	0.703
K=500	0.483	0.661	0.757	0.759	0.805
K=1000	0.531	0.704	0.729	0.823	0.898
K=3000	0.618	0.763	0.78	0.831	0.898

Conclusions

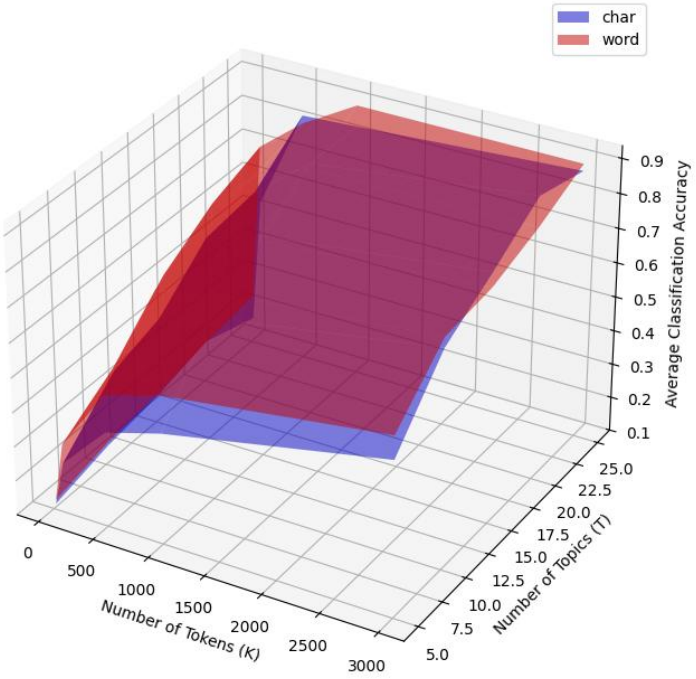
1. 主题数量T的影响



对比一行中的数据，或通过上图，结果显示，随着主题数量T的增加，分类准确率上升至饱和后在一个值附近上下波动。这是由于以下原因，当T较小时，模型提取特征的维度有限，导致在低维上分类性能较低；而当T进一步增大，模型提取特征的维度进一步增大，分类的效果越来越好。

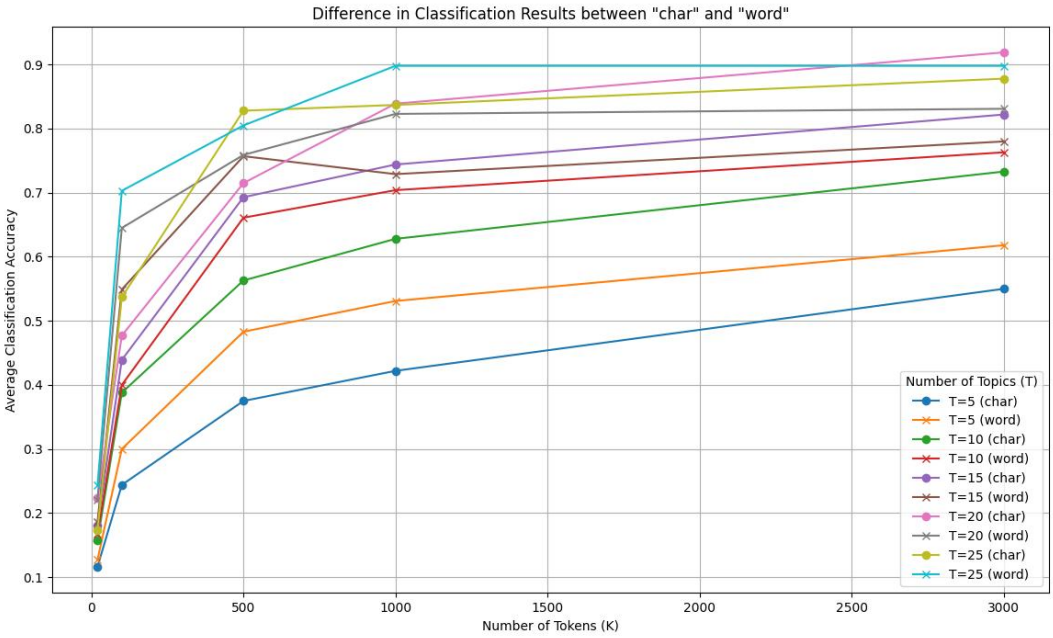
同时也可以发现，**当K值较小时，T的增加对准确率的提升较小；而K较大时，随着T的增加准确率增加较为显著。**这可能是因为，K过小时信息有限，导致LDA的结果较差，此时主要是K这一因素制约性能，因此T提升效果不明显。而K较大时制约量主要就成了T，因此T增大对分类的影响显著。

2. “词”与“字”为基本单元的差异



对比两张表中相同位置的数据，或通过上图，结果显示，以“词”为基本单元时，分类性能大部分是优于以“字”为基本单元的。这是可能因为“词”相比“字”有更高的信息熵，能够更好地捕捉文本的语义信息，从而为分类器提供更有效的特征。

3. 文本长度 K 的影响：



对比一列中的数据，或通过上图，结果显示，**随着文本长度K的增加，分类准确率上升至饱和值**。这可能是因为，对于较短的文本（如K=20），由于信息量有限，主题模型难以

准确提取有效的特征，分类性能较差。随着文本长度的增加（如 K=100、500），主题模型能够更好地捕捉文本的主题分布，分类性能显著提升。然而，当文本长度进一步增加到K=3000时，分类性能的提升趋于平缓，这表明过长的文本虽然包含更多的信息，但也会引入更多的噪声，对分类性能的提升作用有限。

References

[1] https://rxfz7565awz.feishu.cn/docx/EECSdIqVDo60BGxfLNtc7SiqnYe

[2] https://web.stanford.edu/~jurafsky/slp3/A.pdf