

# 深度学习与自然语言处理 作业 2 报告

魏才伦  
1260034837@qq.com

## Abstract

本文在所给语料库上均匀选取了 1000 个段落，每个段落的标签是该段落所属小说，利用 LDA 模型在所给的语料库上建模，主题数量为  $T$ ，得到每个段落的主题分布特征，然后采用贝叶斯分类器将主题分布特征分类，并进行 10 次交叉验证，分别采用“词”和“字”作为基本单元，改变每个段落的 token 数量  $K$ ，LDA 模型主题数量  $T$ ，分析基本单元、段落长短和主题数量对分类准确率的影响。

## Introduction

LDA (Latent Dirichlet Allocation) 是一种用于主题建模的概率生成模型，最早由 Blei、Ng 和 Jordan 于 2003 年提出。它是一种无监督学习方法，用于发现文本数据中隐藏的主题结构。LDA 模型的目标是通过观察到的文档集合来推断出隐藏的主题结构，以及每个文档和每个词汇在主题上的分布情况。在训练过程中，LDA 通过迭代地调整主题和词汇的分布，以最大化文档集合的似然函数。LDA 模型在文本挖掘、信息检索和推荐系统等领域有着广泛的应用。它可以用于文档主题分析、文档聚类、信息检索中的相似文档发现等任务，帮助人们理解文本数据中的隐藏结构和模式。

## Methodology

本文采用 python 中 gensim 模块生成段落词袋模型，并构建 LDA 模型，采用 scikit-learn 模块中的 MultinomialNB() 作为分类器，分别以“词”和“字”为基本单元，改变段落 token 数量，和 LDA 主题数量，进行 10 次交叉验证。

## M1: LDA 模型

LDA 由 Blei, David M.、Ng, Andrew Y.、Jordan 于 2003 年提出，用来推测文档的主题分布。它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取它们的主题分布后，便可以根据主题分布进行主题聚类或文本分类。

LDA 模型构建的方法为：假设共有  $m$  个段落，设定  $K$  个主题，每个段落都有各自的主题分布，主题分布是多项式分布，该多项式分布的参数服从 Dirichlet 分布，该 Dirichlet 分布的参数为  $\alpha$ ，每个主题都有各自的词分布，词分布也是多项式分布，该多项式分布的参数服从 Dirichlet 分布，该 Dirichlet 分布的参数为  $\beta$ ，LDA 模型的主要任务是学习这两个概率分布，为了学习这两个分布，LDA 模型使用了 EM 算法，在 E 步骤中，通过已知的参数来计算隐藏变量，即每个词的主题分布，在 M 步骤中，通过最大似然函数来更新参数，即主题分布和词分布，迭代这个过程，直到模型收敛到一个稳定的状态。在实际应用中，LDA 模型参数估计通常使用基于 Gibbs 采样的变分推断算法，Gibbs 采样通过在给定其他参数的情况下对每个隐藏变量进行采样，从而逼近后验分布。

## M2: 贝叶斯分类器

贝叶斯分类器是一类基于贝叶斯定理的统计分类器，它根据特征之间的条件独立性假设，通过计算给定类别的情况下特征的条件概率来进行分类。贝叶斯分类器在文本分类、垃圾邮件过滤、情感分析等领域被广泛应用。

朴素贝叶斯算法的步骤如下：

(1) 设某样本属性集合  $x = \{x_1, x_2, \dots, x_n\}$ ，其中  $n$  为属性数目， $x_i$  为  $x$  在第  $i$  属性上的取值；

(2) 把这个样本划分为类别集合  $c$  中的某一类， $c = \{y_1, y_2, \dots, y_m\}$

(3) 计算后验概率  $P(y_j | x) = \frac{P(y_j)P(x | y_j)}{P(x)} = \frac{p(y_j) \prod_{i=1}^n p(x_i | y_j)}{p(x)}$ ,  $j = 1, 2, \dots, m$

(4) 如果  $P(y_j | x) = \max\{P(y_1 | x), P(y_2 | x), \dots, P(y_m | x)\}$ ，则样本在属性集  $x$  下属于类别  $y_j$

其中，每个类别的先验概率  $p(y_j)$  和每个特征在每个类别下的条件概率  $p(x | y_j)$  是从训练数据中估计得到的。

### M3: 程序流程

本文程序中构建 LDA 模型的流程如下：

Step 1: 读取语料库，当以“词”为基本单元时，采用 jieba 模块进行中文分词，去除停用词，每个词视为 1 个 token，分别以 20、100、500、1000、2500 个 token 组成每个段落，从所有段落中均匀抽取 1000 组作为实际使用的数据集；以“字”为基本单元时，去除停用字，每个字视为 1 个 token，分别以 20、100、500、1000、2500 个 token 组成每个段落，从所有段落中均匀抽取 1000 组作为实际使用的数据集；

Step 2: 采用 gensim.corpora 模块，生成数据集的词典，并将每个段落转化为词袋向量，分别设定主题数量为 8、16、32、64，采用 gensim.models.LdaMulticore 函数，输入段落数据和主题数量，构建 LDA 模型；

Step 3: 采用构架的 LDA 模型生成所有段落数据的主题分布，将该主题分布作为段落的特征向量，即分类器的输入；

Step 4: 采用 scikit-learn 模块的 MultinomialNB()函数构造朴素贝叶斯分类器，并采用 cross\_val\_score()进行 10 次交叉验证，得到 10 次交叉验证的平均准确率。

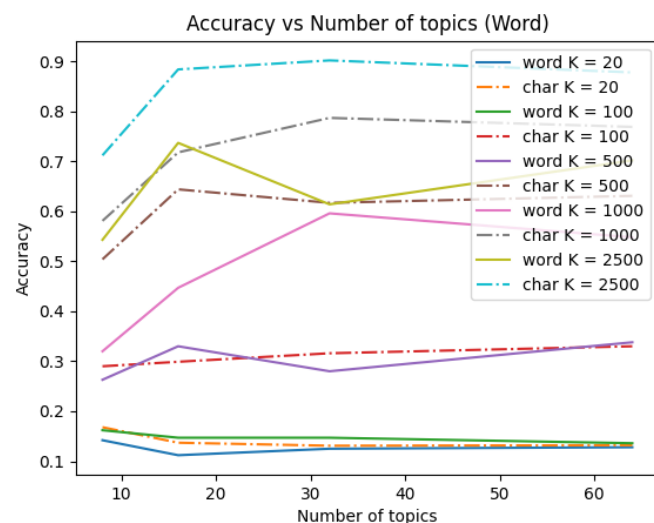
## Experimental Studies

本文分别以“词”和字作为基本单元，段落长度 K 为 20、100、500、1000、2500 个 token，在语料库均匀抽取 1000 个段落，以及主题数量 T 为 8、16、32、64 的条件下，构造 LDA 模型，将每个段落的主题分布输入贝叶斯分类器，进行 10 次交叉验证，得到的平均准确率如下表所示。

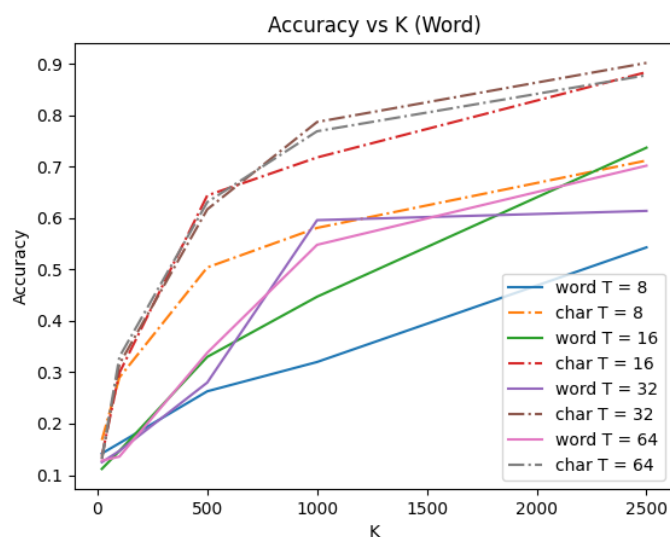
表 1 分类平均准确率

主题数量 T 段落长度 K	以“词”为基本单元				以“字”为基本单元			
	8	16	32	64	8	16	32	64
20	0.142	0.112	0.125	0.128	0.168	0.137	0.131	0.132
100	0.162	0.142	0.147	0.136	0.290	0.299	0.316	0.330
500	0.263	0.330	0.280	0.338	0.504	0.644	0.617	0.631
1000	0.320	0.447	0.596	0.548	0.581	0.718	0.787	0.769
2500	0.543	0.737	0.614	0.702	0.712	0.884	0.902	0.878

段落长度 K 相同时，比较不同主题数量 T 对准确率的影响，绘制曲线图如图 1 所示，发现段落长度相同时，主题数量增加可以提高准确率，但如果主题数量过高，准确率反而会下降，其原因可能是主题数量过高会导致特征过拟合，泛化能力较弱。



主题数量  $T$  相同时，比较不同段落长度  $K$  对准确率的影响，绘制曲线图如图 2 所示，发现主题数量相同时，增加段落长度可以提高准确率，且段落越长准确率越高，其原因是更长的段落包含的信息越多，越有助于分类特征的提取。



观察图 1 和图 2 以“词”为基础单元的准确率要低于以“字”为基础单元的准确率，但此结果与常识并不相符，理论上中文词组应该包含更多的语义信息，更有助于分类特征的提取。实验结果“词”的准确率较低的原因可能与语料库有关，该语料库均是金庸的小说，所用词组可能比较相似，导致词组的区分度较弱。

## Conclusions

本文通过实验发现段落长度相同时，主题数量增加可以提高准确率，但如果主题数量过高，导致主题分布过拟合，泛化能力较弱，准确率反而会下降。主题数量相同时，增加段落

长度，包含的信息越多，越有助于分类特征的提取，可以提高准确率。本实验中以“词”为基本单元的准确率低于以“字”为基本单元，与理论不符，原因可能与数据集词组区分度小有关。

## References

- [1] <https://zhuanlan.zhihu.com/p/97627567>
- [2] <https://blog.csdn.net/Hjh1906008151/article/details/127867912>
- [3] [https://blog.csdn.net/weixin\\_45934622/article/details/130383085](https://blog.csdn.net/weixin_45934622/article/details/130383085)