# 作业 2 报告:中文语料库上的 LDA 建模与分类

朱隽凡
1362992167@qq.com

## Abstract

本报告基于作家金庸所著书籍，使用 Latent Dirichlet Allocation（LDA）模型对该语料库上的段落进行建模，本报告研究了在 10 折交叉验证下主题个数 T、以“词”或“字”为基本单元以及 token 个数 K 对分类性能的影响。

## Introduction

LDA(Latent Dirichlet Allocation) 主题模型是一种文档生成模型，是一种非监督机器学习技术。它认为一篇文档是有多个主题的，而每个主题又对应着不同的词。一篇文档的构造过程，首先是以一定的概率选择某个主题，然后再在这个主题下以一定的概率选出某一个词，这样就生成了这篇文档的第一个词。不断重复这个过程，就生成了整篇文章（当然这里假定词与词之间是没有顺序的，即所有词无序的堆放在一个大袋子中，称之为词袋，这种方式可以使算法相对简化一些）。

LDA 的使用是上述文档生成过程的逆过程，即根据一篇得到的文档，去寻找出这篇文档的主题，以及这些主题所对应的词。

## Methodology

对给定的数据集，先提取出其中的有效部分，去除其中的广告信息和停词，然后对各书籍抽取一定的段落并构造标签，确保每个段落的 token 数固定为 K，标签即为该段落所在书籍名称。指定主题数 T 后，对得到的段落数据使用 LDA 模型建模，从而得到训练数据（T 维向量的每个分量表示该主题的概率），再使用随机森林模型对该数据进行分类，使用 10 折交叉验证统计得到分类的准确率。

## Experimental Studies

# 1.探究 token 个数 K 对分类性能的影响

对文本中的数据，在不同 K 取值下以词为单位，主题数 T 固定为 10，分类结果如表 1 所示。可见 K 对准确率有较大影响，当 K 取值小于 1000 时，LDA 无法很好地提取出文本的特征，从而导致分类准确率较低，而当 K 设置为 3000 时准确率有明显提高。

表 1: K 对分类准确率的影响

| K | Accuracy |
|---|---|
| 20 | 0.13400000000000004 |
| 100 | 0.148 |
| 500 | 0.177 |
| 1000 | 0.261 |
| 3000 | 0.558 |

# 2.探究主题个数 T 对分类性能的影响

对文本中的数据，在不同 T 取值下以词为单位，token 数 K 固定为 3000，分类结果如表 2 所示。T 对分类结果的影响并不大，但 T 太小或太大都有可能导致 LDA 分类不准确。

表 2: T 对分类准确率的影响

| T | Accuracy |
|---|---|
| 5 | 0.619 |
| 10 | 0.558 |
| 20 | 0.7009999999999998 |
| 50 | 0.6020000000000001 |
| 100 | 0.622 |

# 3.探究以“词”或“字”为基本单元对分类性能的影响

对文本中的数据，将 token 数 K 固定为 3000，并以字为单位，调整主题个数 T，得到分类准确率如表 3 所示。

表 3: 以字为单位时 T 对分类准确率的影响

| T | Accuracy |
|---|---|
| 5 | 0.5970000000000001 |
| 10 | 0.8860000000000001 |
| 20 | 0.951 |

| | |
|---|---|
| 50 | 0.975 |
| 100 | 0.9810000000000001 |

# Conclusions

通过本次作业学习了 LDA 模型的原理，并将其应用在中文语料库的主题分类。