作业三、基于 LDA 算法的文本分类

崔多
1479518308@qq.com

一 摘要

从上面链接给定的语料库中均匀抽取 200 个段落（每个段落大于 500 个词），每个段落的标签就是对应段落所属的小说。利用 LDA 模型对于文本建模，并把每个段落表示为主题分布后进行分类。验证与分析分类结果，（1）在不同数量的主题个数下分类性能的变化；（2）以"词"和以"字"为基本单元下分类结果有什么差异?

本文通过在给定的 16 个语料库中进行段落抽取，利用 LDA 进行文本建模并以主题分布，通过研究分类结果，对不同主题个数下 LDA 分类性能以及以字和词为区分基本单位区分的 LDA 算法进行性能评估。

二 研究内容简介

LDA 的数学模型将一篇文章看作是多个话题的组合，每个话题由某些单词组成。LDA 认为文章中的每个单词都由其中一个话题生成，并考虑了这个单词和其他单词在话题分布上的关系。通过对大量文本语料库进行 lda 主题建模，我们可以发现隐藏在数据背后的潜在话题，并根据这些话题重新组织和理解文本内容。

三 实验方法

3.1 LDA 算法

LDA（Latent Dirichlet Allocation，潜在狄利克雷分配）是一种用于处理大量文本数据并发现其中主题的统计模型。LDA 假设每个文档都由多个主题混合而成，并且每个主题又有很多单词来描述。通过对文本数据进行预处理和模型训练，可以从中提取出这些主题和单词的分布，帮助我们更好地理解文本数据。LDA 模型由三层贝叶斯结构组成，如图 1 所示。

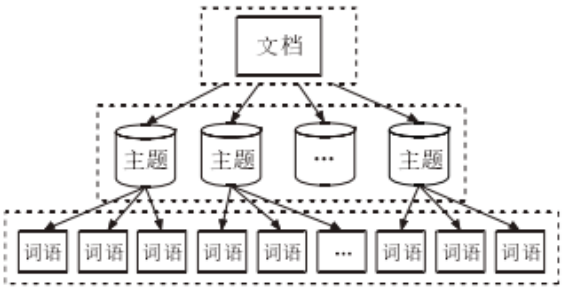

图 1 LDA 模型结构示意图

其基本做法是:

1）随机指定每篇文档的主题分布、每个主题的词语分布；

2）针对文本数据进行迭代处理，不断修正每篇文档的主题分布和每个主题的词语分布，使得每个文档都拥有能够描述其特性的主题分布，同时每个主题都由最具代表性词语组成。

3.2 LDA 文本分类模型

LDA 文本分类的基本思路是将文本数据向量化为主题特征，然后利用这些特征进行分类。模型实现的基本思路为：

一个语料库有 D 篇文档 s，K 个主题 s，

1）对于每一个主题 k∈K，计算出β_k在参数β下的狄利克雷分布；

2）对于每一篇文档 d∈D，计算出ϑ_d在参数α下的狄利克雷分布；

3）对于在文档 d 中的每一个词语 i，计算其主题分布z_{di}的多项分布，计算观察到的某一个词语w_{ji}的多项式分布。

通过计算后验概率，如式 3.1 所示，更新参数。：

$$\mathrm{p}(\mathrm{z},\vartheta,\beta|\mathrm{w},\alpha,\mathrm{n}) = \frac{p(z,\vartheta,\beta|\alpha,n)}{p(w|\alpha,n)} \tag{3.1}$$

四 实验步骤

4.1 数据描述

给定语料库攻击文章 16 篇，从每篇文章中共计均与抽取 13 个段落。

为了实现数据的均匀抽取，通过给定的停词词库进行文章分词，进行文章总词语数的计算。并通过将每篇文章均匀划分为 13 个部分，分布在 13 个部分中抽取前 500 个词语，作为段落区间。共计获取段落 208 个。抽取结果见附件 data.csv。

4.2 数据预处理

为了确保数据的有效性，删除评论数据中噪声数据和无用信息。去除英文，根据 cn_punctuation.txt 去除特殊符号等非中文的干扰信息，通过 jieba 分词进行词语划分以及汉字划分，

4.3 LDA 模型主题聚类模型

本研究通过 sklearn 进行 LDA 建模，首先将文本中的词语转化为词频矩阵，然后通过困惑度与一致性分析进行主题数目选择。

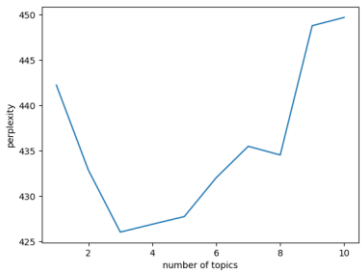

图 2 LDA 困惑度分析结果

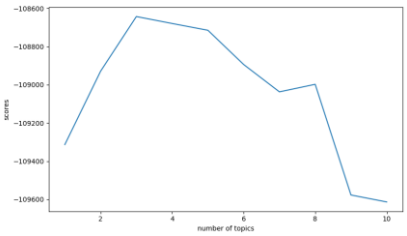

图 3 LDA 一致性分析结果

可以看到，当主题数目为 3 时，LDA 的困惑度最高，这是由于数据虽然来自语料库的 16 篇文章，但由于都是武侠性质，文章的主题重合度较高，区分难度较大。以主题数目为 3 作为示例，在 3 种主题下，文章的三个主题的前十五个特征词为：

表 1 在 3 种主题下前十五个特征词

主题	特征词
Toic1	他们 咱们 二人 心中 一声 什么 两人 不是 武功 只见 不知 如此 兄弟 心想 如何
Toic2	剑士 姑娘 范蠡 少女 李文秀 长剑 汉子 一声 突然 老人 少年 女子 黄蓉 汉人 女孩
Toic3	小宝 韦小宝 韦小 令狐 令狐冲 皇帝 三十 中国 知道 甚么 什么 还是 不是 第三 天下

同样的，当以单独汉字作为划分依据，同样在 3 种主题下，获得的主题特征汉字为：

表 2 在 3 种主题下前十五个特征汉字

主题	特征词
Toic1	剑 她 士 青 刀 黄 文 女 范 阿 克 招 左 苏 马
Toic2	师 山 主 弟 女 她 龙 林 兄 第 胡 半 少 公 点
Toic3	马 兵 王 国 宝 官 袁 军 百 李 皇 文 金 志 其

4.3 决策树进行文本分类

以 LDA 模型获取到的段落在某一主题下的概率向量作为输入，以段落所属文章作为分类标签，通过决策树模型进行文本分类，将训练集额测试集以 8：2 划分。在不同主题数下，模型的准确率如图 4 所示。

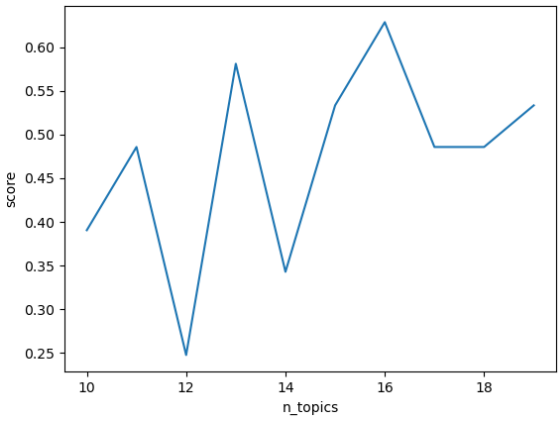

图 4 以词为单位不同主题数下的模型准确率

可以看到，当主题数选择为 16 时，模型的准确率最高，为 0.631。这与我们 16 篇文章是一致的。

而以汉字为单位，模型的准确率变化如图 5 所示。在主题数 44 时产生峰值，准确率最高为 0.619。

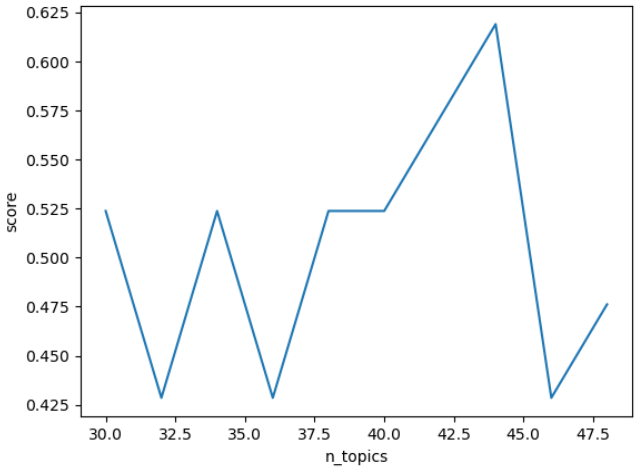

图 5 以汉字为单位不同主题数下的模型准确率

五 实验结果及评估

可以看到，模型的准确率并不高，这是由于模型的主题重复度较大，导致区分准确度较低，模型的收敛能力较差。

且模型受到主题划分个数的影响较大，可以看到，随着主题划分数目的增加，准确率总体呈现增长趋势，当主题数过多时，会产生过拟合的现象。

且以词语为基本单位的主题分类效果相较与于以汉字进行划分的准确率更高，以词进行主题划分的可靠性比汉字更强。