第三次作业

范想明-ZY2207304 xmfan@buaa.edu.cn

摘要

本研究通过对给定的 16 篇小说的语料进清洗和均匀段落采样，得到语料的采样结果对其进行 LDA 文本建模，得到了其不同主题数下的特征词，并分析了困惑度和分类准确度。

一、研究背景

LDA（Latent Dirichlet Allocation）是一种无监督的文本建模技术，用于从大量文本中发现主题的分布，并将每个文档表示为不同主题的概率分布。LDA 假定每个文档都由多个主题组成，每个主题都由单词的概率分布定义，主题分布是从 Dirichlet 分布中采样得到的。LDA 可以帮助我们理解文档集合中的主题和单词之间的关系，进而帮助我们进行文本分类、信息检索等任务。

二、实验方法

2.1 数据采样

本研究首先对语料进行了清洗并基于停用词对清洗的语料进行了划分，对于得到的结果进行均匀采样，每个文章选取 13 个段落，每个段落选取 500 词，并将采样结果分别以词和字的形式保存在一个 json 文件中，如图 1 所示。

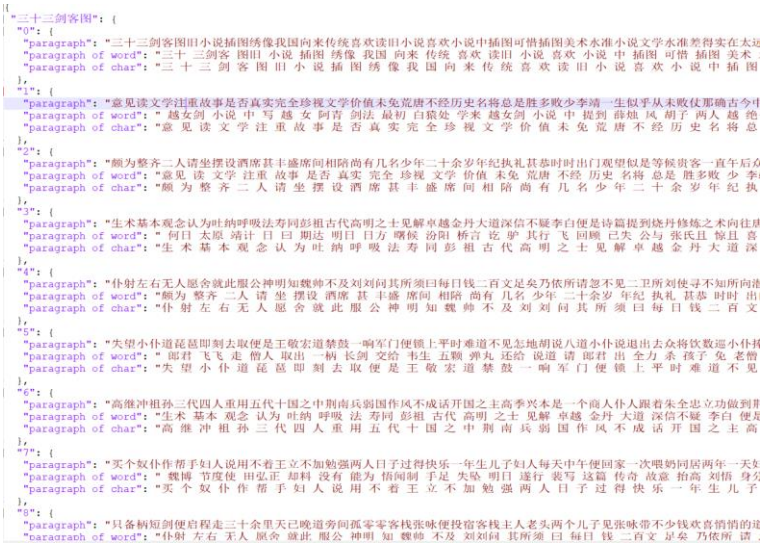

图 1 采样结果

2.2 LDA 建模

在 LDA 建模中，我们首先需要将文本数据转换为数字向量表示，常用的方法是将文本分词、去停用词后使用词频或 TF-IDF 表示。接着，我们可以使用 sklearn 等工具库中的 LDA 算法进行建模，通过调节主题数和超参数等参数，可以得到不同的主题模型。模型训练完成后，我们可以通过困惑度和一致性等指标来评估模型的好坏，并通过可视化工具如 pyLDAvis 来展示模型结果。

本研究基于 sklearn 工具库中的 LDA 算法进行建模，通过调节主题数得到不同的主题模型。如图 2 所示为主题数为 5 的 LDA 模型的特征词，如图 3 所示为主题数为 5 的 LDA 模型的特征字，其结果均符合金庸武侠小说的主题词、字。

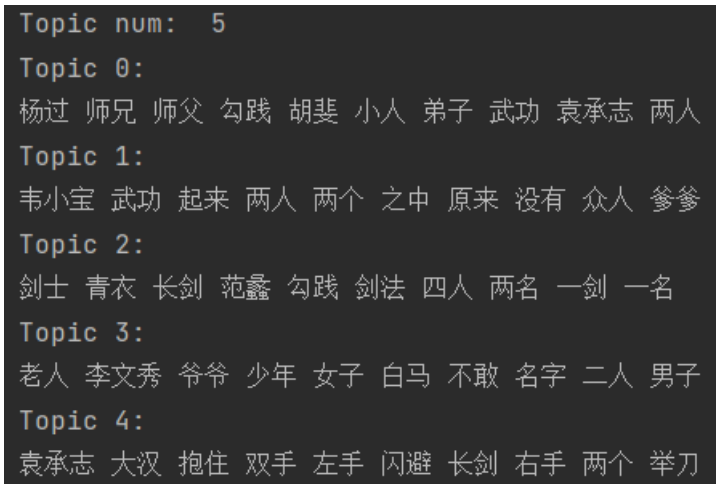

图 2 主题数为 5 的 LDA 模型的特征词

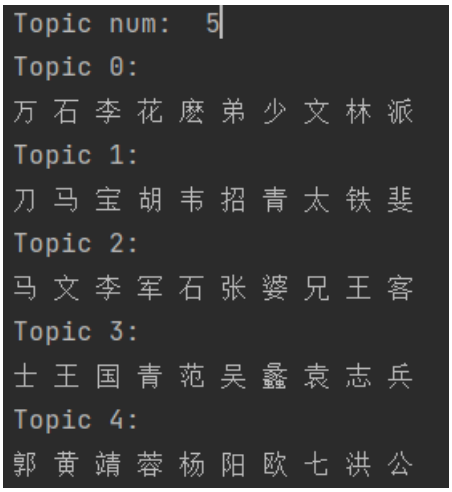

图 3 主题数为 5 的 LDA 模型的特征字

2.3 模型分析

如图 4 所示为不同主题数下的模型困惑度，由于该语料库为同一作者所作的类型相似的文章，因此主题大致相同，如图中在主题数相对较低时模型困惑

度较低。

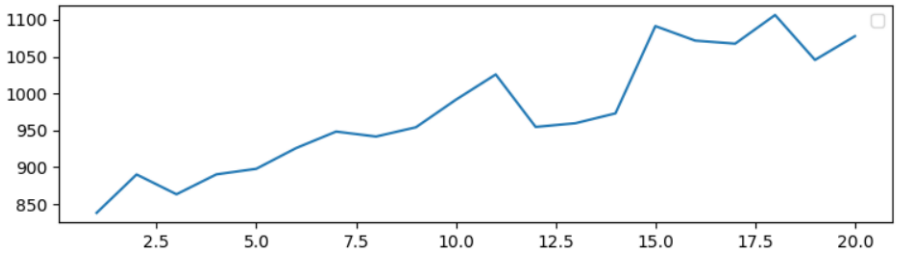

图 4 不同主题数下的模型困惑度

如图 5 所示为不同主题数下以词划分的模型预测准确度曲线，可以看到，当主题数为 15 时，模型的准确度最高，而语料的文章数为 16，建模结果与真实内容相当。而以字划分的模型预测准确度没有一个明显的峰值，如图 6 所示，且最值小于以词划分的模型的准确度的峰值，说明在预料主题高度相似情况下以词划分的模型优于以字划分的模型。

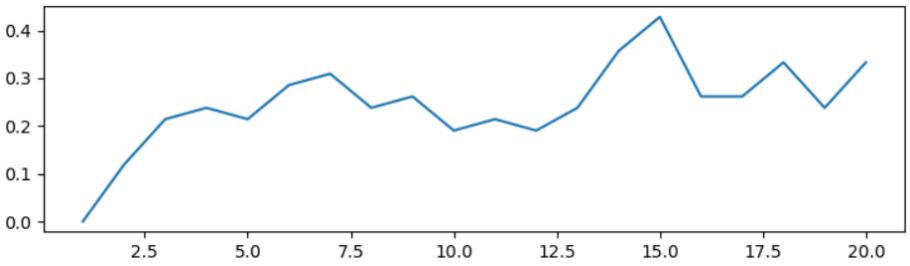

图 5 不同主题数下以词划分的模型预测准确度

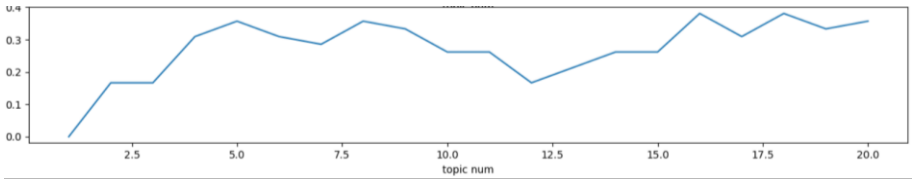

图 6 不同主题数下以字划分的模型预测准确度

三、结论

本研究针对金庸的 16 篇武侠小说作为语料库进行 LDA 建模，在语料库主题高度相似的情况下，LDA 建模的困惑度较高，且无论是以词划分的模型还是以字划分的模型，其建模的质量均不高，但是以词划分的模型在主题数与文章数相近时模型的准确度到达一个峰值。