作业 3 LDA 模型文本建模

一、实验内容

从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？

二、实验原理

LDA 模型全称隐含狄利克雷分布，是一种主题模型，可以将每篇文档的主题按照概率分布的形式给出。LDA 也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。作为一种非监督机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。它采用了词袋的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

举个例子，我们日常生活中总是产生大量的文本，如果每个文本存储为一篇文章，那么每篇文档从人的观察来说就是有序的词的序列 $d=(w_1,w_2,...,w_n)$，包含 m 篇文章的语料库，每个文档有 N_m 个单词，一共涉及到 K 个主题；每篇文档都有各自的主题，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α；每个主题都有各自的词分布，词分布为多项式分布，该多项式分布的参数服从 Dirichlet 分布，该 Dirichlet 分布的参数为β；对于某篇文档 d 中的第 n 个词，首先从该文档的主题分布中采用一个主题，然后再这个主题对应的词分布中采用一个词，不断重复该操作，直到 m 篇文档全部完成上述过程。LDA 模型的示意图如图 1 所示。

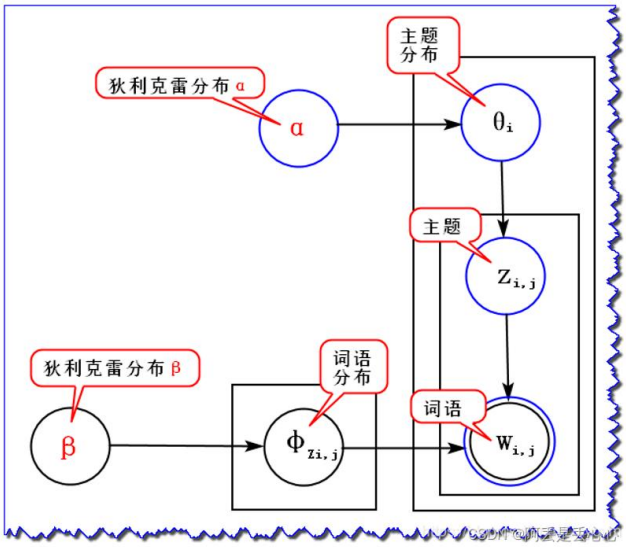

图 1

三、实验过程和结果

1 数据预处理

语料库中共 16 篇文章，要求均匀抽取 200 个段落，每个段落抽取 500 词以上。在实验中每篇文章抽取 13 个段落，共抽取 208 个段落，每个段落抽取 500 词。首先读取每篇文章的内容，删除广告内容和停用词，使用 jieba 分词系统对文章进行分段，每段选取前 500 个词组成段落。

2 探究不同主题数下分类性能变化

一般用来评价 LDA 主题模型的指标有困惑度（perplexity）和主题一致性（coherence），困惑度越低或者一致性越高说明模型越好。但一些研究表明困惑度并不是一个好的指标，因此本文将用主题一致性评价模型。

（1）字模型

图 2 给出了以“字”为基本单元时主题一致性随主题数的变化情况。从图中可以看出，不同主题数下的主题一致性分数有所差别，在本文的实验条件下主题数为 17 时一致性最高。

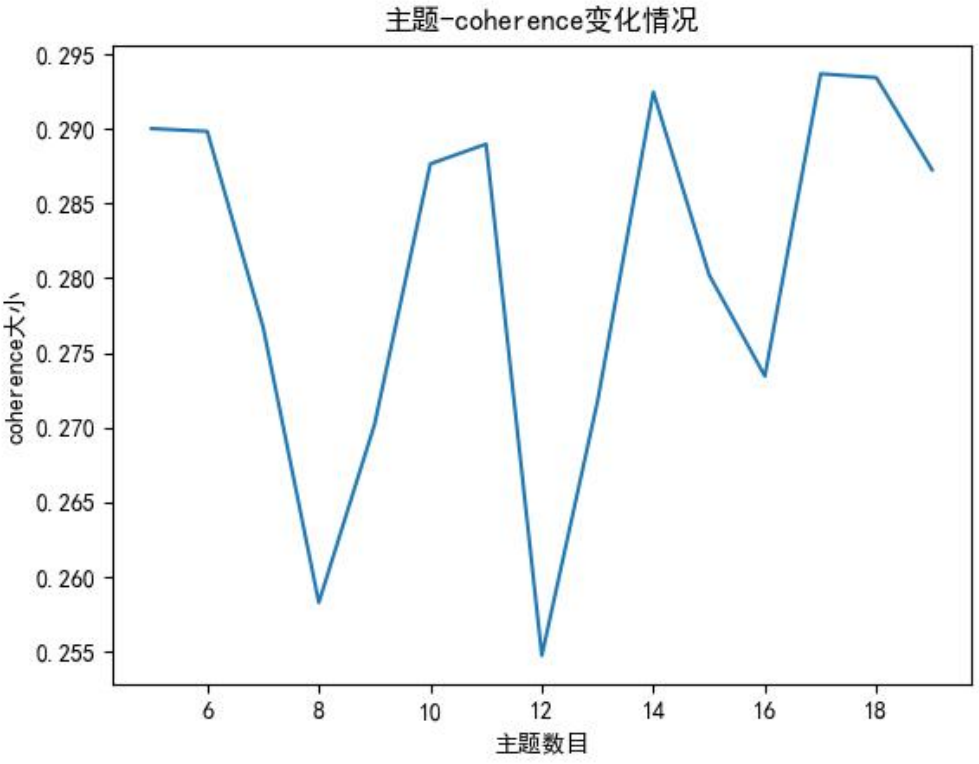

图 2

表 1 给出了主题数为 17 时每一个主题频率最高的 9 个关键词。从表中可以看出在所有的主题中，“道”、“手”等字出现频率都很高。

表 1

主题	出现频率最高的 10 个字								
1	道	李	克	文	苏	见	说	秀	汉
2	道	令	狐	说	手	剑	心	笑	见
3	道	说	手	剑	见	心	身	老	青
4	道	手	说	剑	心	见	身	老	知
5	道	手	说	见	身	心	陈	洛	老
6	道	手	心	说	剑	见	身	生	知
7	道	手	说	见	身	心	笑	刀	高
8	道	苏	李	秀	文	克	说	著	见
9	剑	道	手	说	身	见	青	心	吴
10	道	说	手	心	见	剑	知	时	想
11	说	道	张	行	年	官	杀	相	李

12	道	手	说	刀	见	心	身	石	胡
13	道	韦	宝	说	手	想	心	身	笑
14	道	说	手	身	见	心	知	年	然
15	道	说	身	见	手	剑	心	老	想
16	道	手	见	青	袁	张	说	年	承
17	道	心	手	见	说	黄	笑	郭	杨

用 pyLDAvis 可视化分析如图 3 所示，可见字模型不能很好地区分各个主题。

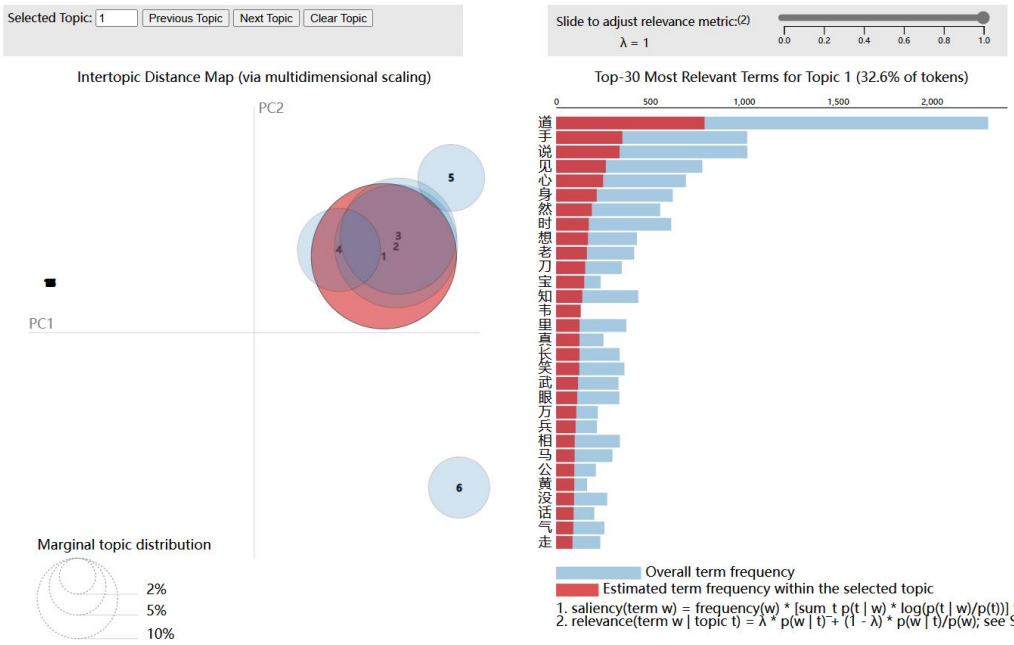

图 3

（2）词模型

图 4 给出了以“词”为基本单元时主题一致性随主题数的变化情况。从图中可以看出，主题数为 17 时一致性最高，与字模型相同。并且与字模型相比，词模型的主题一致度分数要更高。

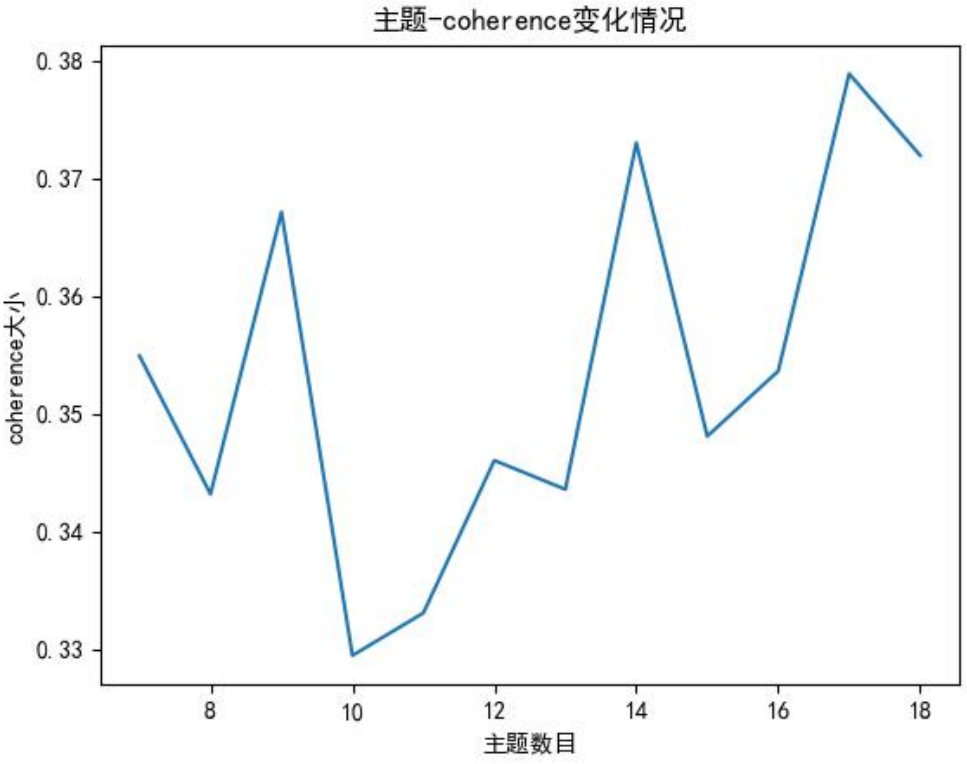

图 4

表 2 给出了主题数为 17 时每一个主题频率最高的 9 个关键词。从中可以看出，不同主题出现频率最高的关键词相差较大。

表 2

主题	出现频率最高的 9 个词
1	杨过 小龙女 麽 牋 周伯通 潇湘子 武三通 蜈蚣 於
2	一声 麽 剑 范蠡 石破天 心想 袁承志 姑娘 李文秀
3	韦小宝 康熙 公主 皇上 双儿 一声 皇帝 侍卫 请
4	麽 郭靖 著 曹云奇 一声 心想 爹爹 铁木真 姑娘
5	李文秀 陈家洛 麽 苏鲁克 车尔库 苏普 一声 汉人 著
6	麽 一声 韦小宝 袁承志 爹爹 著 心想 剑 李文秀
7	袁承志 胡斐 青青 一声 程灵素 心想 大汉 袁崇焕 银子 赵半山
8	郭靖 黄蓉 欧阳克 洪七公 铁木真 哲别 穆念慈 欧阳锋 一声
9	张无忌 张翠山 宋青书 谢逊 殷素素 赵敏 那小鬟 俞岱岩 一声 少林
10	麽 著 萧中慧 曹云奇 卓天雄 周威信 瞎子 於 一声 刀
11	韦小宝 一声 麽 心想 剑 姑娘 著 陈家洛 胡斐

12	麽 韦小宝 一声 著 萧中慧 心想 江湖 派 剑士
13	范蠡 剑士 石破天 青衣 剑 勾践 阿青 一声 长剑 薛烛
14	数页 数百人 点头称是 留恋 触手 疗伤 丫鬟 祖深 注意 曲子
15	麽 范蠡 一声 韦小宝 著 心想 令狐冲 剑
16	杨过 小龙女 麽 牋 周伯通 潇湘子 武三通 蜈蚣 於
17	一声 麽 剑 范蠡 石破天 心想 袁承志 姑娘 李文秀

用 pyLDAvis 可视化分析如图 5 所示，可见词模型的分类结果要好于字模型，但对于一些主题如主题 1 和 10、7 和 9、4 和 8 等，词模型也没有很好地区分开。

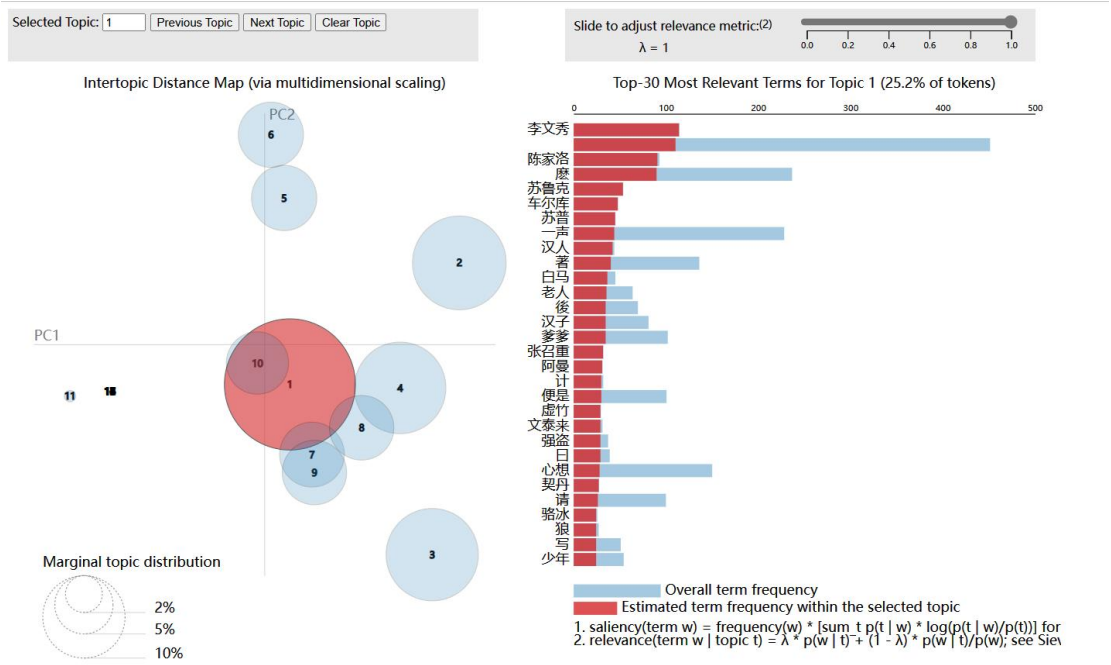

图 5

四、实验总结

本次实验使用 LDA 模型对文本建模，通过验证与分析分类结果可知，不同数量的主题个数下分类性能有区别；以"词"和以"字"为基本单元下分类结果有较大差异，以“词”为基本单元的分类结果要明显好于以“字”为基本单元。