

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 篇小说的语料进行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。

```
{
  "三十三剑客图": {
    "0": {
      "paragraph": "三十三剑客图旧小说插图绣像我国向来传统喜欢读旧小说喜欢小说中插图可惜插图美术水准小说文学水准差得实在太远了",
      "paragraph of word": "三十三剑客图旧小说插图绣像我国向来传统喜欢读旧小说喜欢小说中插图可惜插图美术",
      "paragraph of char": "三十三剑客图旧小说插图绣像我国向来传统喜欢读旧小说喜欢小说中插图",
    },
    "1": {
      "paragraph": "意见读文学注重故事是否真实完全珍视文学价值未免荒唐不经历史名将总是胜多败少李靖一生似乎从未败仗那确是古今",
      "paragraph of word": "意见读文学注重故事是否真实完全珍视文学价值未免荒唐不经历史名将总是胜多败少李靖",
      "paragraph of char": "意见读文学注重故事是否真实完全珍视文学价值未免荒唐不经历史名将总是",
    },
    "2": {
      "paragraph": "颇为整齐二人请坐摆设酒席甚丰盛席间相陪尚有几名少年二十余岁年纪扶礼甚恭时出门观望似是等候客人一直午后",
      "paragraph of word": "意见读文学注重故事是否真实完全珍视文学价值未免荒唐不经历史名将总是胜多败少李靖",
      "paragraph of char": "颇为整齐二人请坐摆设酒席甚丰盛席间相陪尚有几名少年二十余岁年纪扶礼",
    },
    "3": {
      "paragraph": "生术基本观念认为吐纳呼吸法寿同彭祖古代高明之士见解卓越金丹大道深信不疑李白便是诗篇提到炼丹修炼之术向往",
      "paragraph of word": "何日太原端计日日期达明日日方曙候汾阳桥言迄驴其行飞回颺已失公与张氏且惊且喜",
      "paragraph of char": "生术基本观念认为吐纳呼吸法寿同彭祖古代高明之士见解卓越金丹大道深信",
    },
    "4": {
      "paragraph": "仆射左右无人愿舍就此服公神明知魏卿不及刘刘问其所须曰每日钱二百文足矣乃依所请忽不见二卫所使寻不知所向",
      "paragraph of word": "颇为整齐二人请坐摆设酒席甚丰盛席间相陪尚有几名少年二十余岁年纪扶礼甚恭时出",
      "paragraph of char": "仆射左右无人愿舍就此服公神明知魏卿不及刘刘问其所须曰每日钱二百文",
    },
    "5": {
      "paragraph": "失望小仆道甚即刻去取便是王敬宏道禁鼓一响军门便锁上平时难道不见怎地胡说八道小仆说退出去众将饮数巡小仆",
      "paragraph of word": "失望小仆道甚即刻去取便是王敬宏道禁鼓一响军门便锁上平时难道不见",
      "paragraph of char": "失望小仆道甚即刻去取便是王敬宏道禁鼓一响军门便锁上平时难道不见",
    },
    "6": {
      "paragraph": "高维冲祖孙三代四人重用五代十国之中荆南兵弱国作风不成话开国之主高季兴本是一个商人仆人跟着朱全忠立功做到",
      "paragraph of word": "生术基本观念认为吐纳呼吸法寿同彭祖古代高明之士见解卓越金丹大道深信不疑李白便是",
      "paragraph of char": "高维冲祖孙三代四人重用五代十国之中荆南兵弱国作风不成话开国之主高",
    },
    "7": {
      "paragraph": "买个奴仆作帮手妇人说不着王立不加勉强两人日子过得快乐一年生儿子妇人每天中午便回家一次喂鸭同居两年一天",
      "paragraph of word": "魏博节度使田弘正却料没有能为情闻制手足失聪明日遂行装写这篇传奇故豪抬高刘悟身",
      "paragraph of char": "买个奴仆作帮手妇人说不着王立不加勉强两人日子过得快乐一年生儿子",
    },
    "8": {
      "paragraph": "只备柄短剑便启程走三十里天已晚道旁间孤零零客棧张咏便投宿客棧主人老头两个儿子见张咏带不少钱欢喜悄悄的",
      "paragraph of word": "仆射左右无人愿舍就此服公神明知魏卿不及刘刘问其所须曰每日钱二百文足矣乃依所请",
      "paragraph of char": "仆射左右无人愿舍就此服公神明知魏卿不及刘刘问其所须曰每日钱二百文足矣乃依所请",
    }
  }
}
```

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。

```
Topic num: 5
Topic 0:
杨过 师兄 师父 勾践 胡斐 小人 弟子 武功 袁承志 两人
Topic 1:
韦小宝 武功 起来 两人 两个 之中 原来 没有 众人 爹爹
Topic 2:
剑士 青衣 长剑 范蠡 勾践 剑法 四人 两名 一剑 一名
Topic 3:
老人 李文秀 爷爷 少年 女子 白马 不敢 名字 二人 男子
Topic 4:
袁承志 大汉 抱住 双手 左手 闪避 长剑 右手 两个 举刀
```

图 2 主题数为 5 的 LDA 模型的特征词

```
Topic num: 5
Topic 0:
万 石 李 花 廖 弟 少 文 林 派
Topic 1:
刀 马 宝 胡 韦 招 青 太 铁 斐
Topic 2:
马 文 李 军 石 张 婆 兄 王 客
Topic 3:
士 王 国 青 范 吴 蠡 袁 志 兵
Topic 4:
郭 黄 靖 蓉 杨 阳 欧 七 洪 公
```

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。

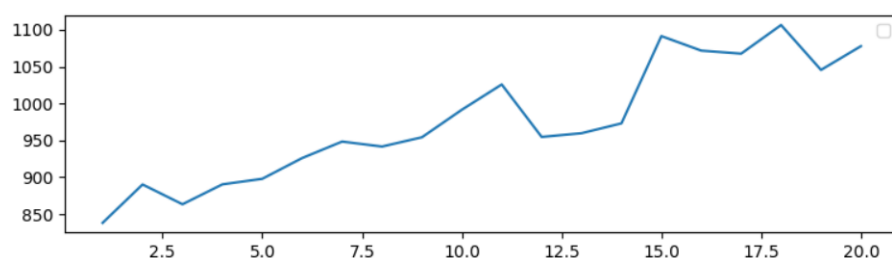


图 4 不同主题数下的模型困惑度

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，且最大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。

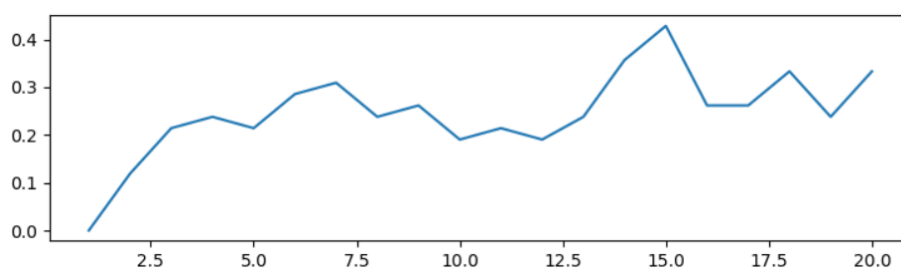


图 5 不同主题数下以词划分的模型预测准确度

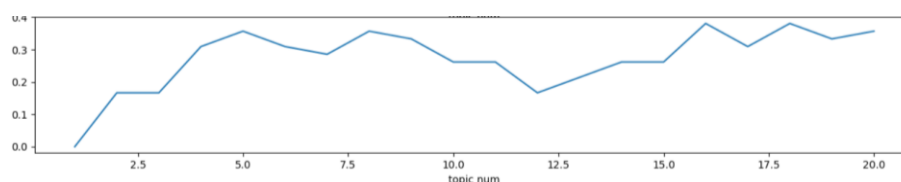


图 6 不同主题数下以字划分的模型预测准确度

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。