

1

****

深度学习与自然语言处理作业3

|  |  |
| --- | --- |
| 院（系）名称 | 自动化科学与电气工程学院 |
| 学生姓名 | 潘翔 |
| 学生学号 | ZY2103707 |

2022年5月

## 一、问题描述

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。 截至日期：5月6日晚12点前

## 二、试验原理

1.LDA原理

LDA是常见的自然语言处理的主题模型之一，是一类无监督学习算法，在训练时不需要手工标注训练集，需要的仅仅是文档集以及指定主题是数量k即可。此外LDA的另一个优点是对于每一个主题均可找出一些词语来描述它。

LDA是一种典型的词袋模型，即它认为一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后的关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。其核心思想是寻找最佳的投影方法，将高维的样本投影到特征空间，使得不同类别间的数据距离最大，而同一类别内的数据距离最小。

LDA中几个概念：

词：数据中的基本离散单元

文档：待处理的数据对象，由词组成，不计顺序。文档对象在topic模型中是词袋概念

话题：每篇文档都有特定的一些话题，根据这些话题产生词从而组成文档

在模型中，一篇文章生成的方法如下

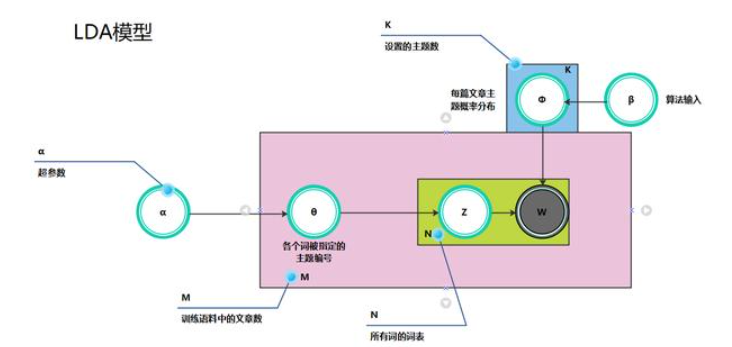
1.从Dirichlet分布中取样生成文档i的主题分布

2.从主题的多项式分布中取样生成文档i的第j个词的主题

3.从Dirichlet分布中取样生成主题对应的词语分布

4.从词语的多项式分布中采样最终生成的词语

其中，类似分布是二项式分布的共轭先验概率分布，而Dirichlet分布是多项式分布的共轭先验概率分布。LDA的图模型结构如下图所示



## 三、试验过程和结果

1.试验过程

选取金庸的十六本小说作为训练集，分别为

1).白马啸西风

2).碧血剑

3).飞狐外传

4).连城诀

5).鹿鼎记

6).三十三剑客图

7).射雕英雄传

8).神雕侠侣

9).书剑恩仇录

10).天龙八部

11).侠客行

12).笑傲江湖

13).雪山飞狐

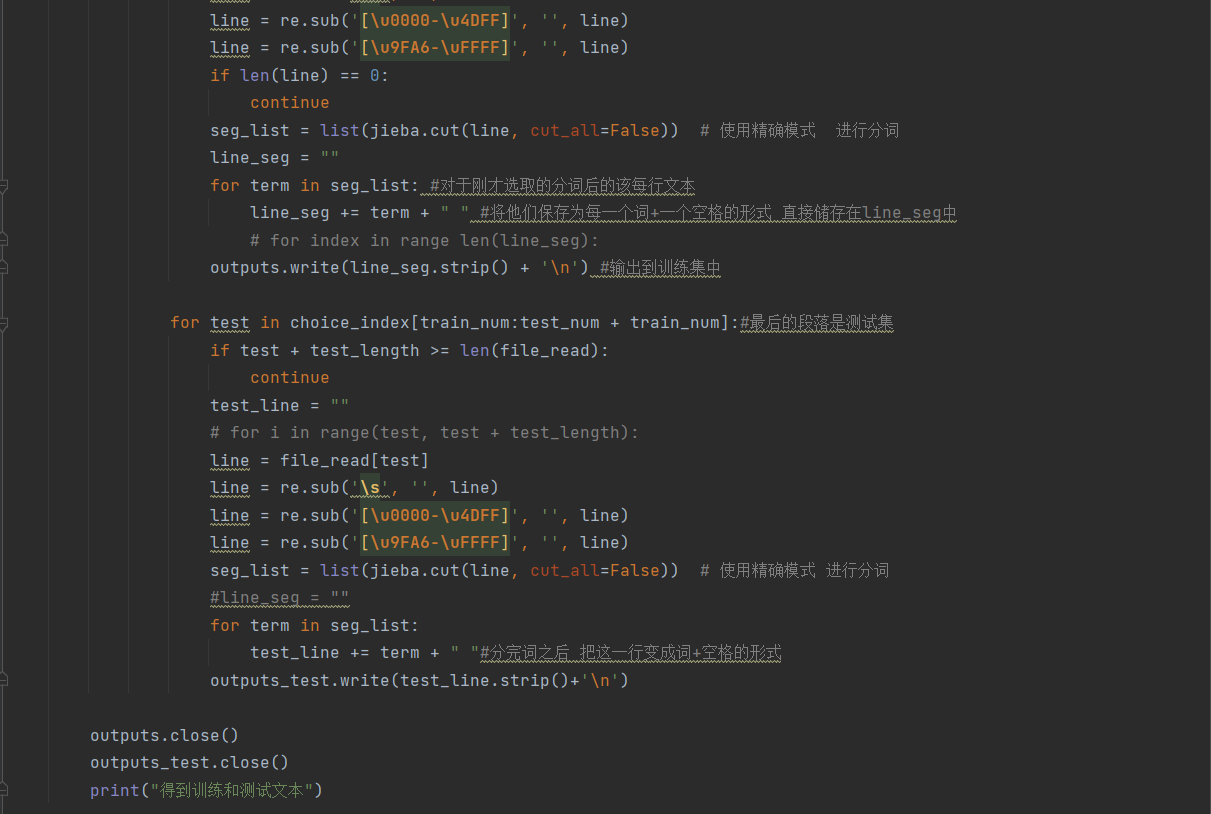
14).倚天屠龙记

15).鸳鸯刀

16).越女剑

先对语料库文本进行处理，删除特殊字符以及无价值字符等，然后在每个小说中随机选取50行文本(大概500词以上)作为测试文本，并且组合到一起，剩余所有文本作为训练文本。测试文本和训练文本的选择代码如下





利用gensim库进行LDA模型的训练，最后输出结果是十六个主题的分布，以及选取的测试文本的主题分布

2.试验结果

**16个主题的单词分布为：**

(0, '0.073\*"道" + 0.050\*"了" + 0.025\*"是" + 0.024\*"的" + 0.021\*"你" + 0.020\*"去" + 0.020\*"我" + 0.017\*"那" + 0.016\*"也" + 0.016\*"咱们"')

(1, '0.026\*"的" + 0.025\*"他" + 0.018\*"剑法" + 0.017\*"已" + 0.017\*"剑" + 0.014\*"又" + 0.013\*"在" + 0.012\*"了" + 0.011\*"是" + 0.011\*"如此"')

(2, '0.019\*"武林" + 0.019\*"功夫" + 0.013\*"了" + 0.013\*"的" + 0.013\*"嗤" + 0.012\*"抓" + 0.011\*"阿" + 0.011\*"心头" + 0.010\*"前来" + 0.009\*"吩咐"')

(3, '0.033\*"听" + 0.025\*"得" + 0.024\*"的" + 0.021\*"了" + 0.019\*"只" + 0.017\*"道" + 0.016\*"天下" + 0.015\*"一" + 0.012\*"叫" + 0.012\*"有人"')

(4, '0.036\*"这次" + 0.022\*"举起" + 0.020\*"啷" + 0.017\*"谢" + 0.016\*"呛" + 0.011\*"危难" + 0.011\*"嗓子" + 0.008\*"辛苦" + 0.008\*"怀疑" + 0.008\*"经历"')

(5, '0.048\*"他" + 0.039\*"的" + 0.037\*"了" + 0.029\*"是" + 0.026\*"着" + 0.016\*"在" + 0.014\*"道" + 0.014\*"便" + 0.012\*"得" + 0.010\*"那"')

(6, '0.033\*"了" + 0.031\*"在" + 0.029\*"的" + 0.020\*"她" + 0.017\*"得" + 0.013\*"上" + 0.012\*"他" + 0.012\*"听" + 0.010\*"中" + 0.007\*"到"')

(7, '0.017\*"向" + 0.017\*"长剑" + 0.016\*"突然" + 0.014\*"左手" + 0.014\*"身子" + 0.014\*"穴道" + 0.013\*"弟子" + 0.013\*"只见" + 0.012\*"右手" + 0.012\*"在"')

(8, '0.073\*"派" + 0.015\*"在" + 0.014\*"与" + 0.013\*"二人" + 0.013\*"未必" + 0.011\*"除了" + 0.011\*"是" + 0.010\*"此时" + 0.009\*"当下" + 0.009\*"怎能"')

(9, '0.043\*"一位" + 0.037\*"至" + 0.028\*"银子" + 0.025\*"号令" + 0.020\*"敌手" + 0.017\*"位" + 0.012\*"西" + 0.011\*"停步" + 0.009\*"移动" + 0.009\*"老大"')

(10, '0.038\*"了" + 0.021\*"师哥" + 0.015\*"的" + 0.013\*"须" + 0.012\*"在" + 0.012\*"喝" + 0.010\*"两名" + 0.010\*"得" + 0.009\*"酒" + 0.008\*"逼"')

(11, '0.084\*"的" + 0.041\*"了" + 0.032\*"是" + 0.031\*"他" + 0.024\*"在" + 0.024\*"这" + 0.013\*"有" + 0.013\*"那" + 0.013\*"也" + 0.012\*"和"')

(12, '0.024\*"无法" + 0.017\*"了" + 0.015\*"的" + 0.013\*"多半" + 0.012\*"今日" + 0.012\*"嘿嘿" + 0.012\*"报仇" + 0.011\*"抢" + 0.010\*"皇帝" + 0.010\*"非"')

(13, '0.037\*"了" + 0.023\*"的" + 0.018\*"便" + 0.016\*"也" + 0.014\*"即" + 0.012\*"人" + 0.012\*"他" + 0.011\*"但" + 0.010\*"是" + 0.010\*"这"')

(14, '0.034\*"不成" + 0.022\*"胡说八道" + 0.017\*"此言" + 0.015\*"凶险" + 0.015\*"说来" + 0.014\*"身分" + 0.013\*"推开" + 0.012\*"骗" + 0.011\*"怎会" + 0.011\*"霎时间"')

(15, '0.108\*"我" + 0.102\*"你" + 0.048\*"道" + 0.035\*"了" + 0.031\*"的" + 0.028\*"是" + 0.022\*"说" + 0.017\*"不" + 0.016\*"她" + 0.015\*"也"')

根据16个主题的单词分布可以看出，存在着大量的常用单词比如“的“，”了“，”是“等等，这样的单词分布对后续测试文本的主题判断存在影响。

**对选取并且组合后的16个段落进行测试：**

0的主题分布为：

[(8, 0.16621971), (11, 0.7083162)]

1的主题分布为：

[(3, 0.09229115), (5, 0.47637045), (6, 0.1613137), (10, 0.08385552), (13, 0.1402859)]

2的主题分布为：

[(2, 0.13807033), (3, 0.4987275), (5, 0.15067177), (8, 0.13744242)]

3的主题分布为：

[(3, 0.14343326), (11, 0.2623876), (15, 0.5264334)]

4的主题分布为：

[(0, 0.010420149), (1, 0.010420123), (2, 0.010420119), (3, 0.010420116), (4, 0.010420111), (5, 0.010420133), (6, 0.010420123), (7, 0.010420115), (8, 0.010420122), (9, 0.010420111), (10, 0.22392856), (11, 0.0104201315), (12, 0.6301897), (13, 0.0104201455), (14, 0.010420112), (15, 0.010420137)]

5的主题分布为：

[(3, 0.15108675), (11, 0.7815863)]

6的主题分布为：

[(8, 0.12873177), (11, 0.5041883), (15, 0.30903953)]

7的主题分布为：

[(0, 0.21781944), (12, 0.07103713), (15, 0.6633336)]

8的主题分布为：

[(0, 0.015633238), (1, 0.015633246), (2, 0.015633244), (3, 0.2624978), (4, 0.015633238), (5, 0.01563324), (6, 0.51863676), (7, 0.015633248), (8, 0.015633238), (9, 0.015633238), (10, 0.015633242), (11, 0.01563325), (12, 0.01563324), (13, 0.01563324), (14, 0.015633238), (15, 0.015633255)]

9的主题分布为：

[(3, 0.11295732), (5, 0.20832662), (10, 0.24682207), (14, 0.06882141), (15, 0.31706068)]

10的主题分布为：

[(3, 0.1458553), (6, 0.272264), (11, 0.4375059), (15, 0.100216225)]

11的主题分布为：

[(10, 0.1935914), (14, 0.13883993), (15, 0.61338145)]

12的主题分布为：

[(5, 0.25131044), (11, 0.6903397)]

13的主题分布为：

[(0, 0.29163375), (3, 0.33957842), (6, 0.17546047), (8, 0.14328936)]

14的主题分布为：

[(11, 0.16789418), (15, 0.77741444)]

15的主题分布为：

[(11, 0.73602265), (15, 0.17644764)]

## 四、试验总结和体会

**从试验结果上可以看出，测试文本的分类效果有限，分析其原因由于所有的文本小说均为金庸的武侠小说，所有客观上他们的主题相近，对分类的效果有一定的影响。而根据16个主题的单词分布可以看出，存在大量的常用单词比如“了“，”的“之类，这样的常用单词并不能很好的反应当前主题的特殊性，并且在不同的主题中存在大量的重复单词，这样的主题导致分类的效果存在局限，要更好地测试该模型，可以选取差异更大的文本进行试验。**

**经过这次试验，提升了我对LDA模型的进一步理解，并且在试验过程中也提升了我对python代码的熟悉程度，便于以后进一步运用python解决其他问题。**

## 五、参考资料

https://blog.csdn.net/weixin\_42663984/article/details/116264233?spm=1001.2014.3001.5502

https://blog.csdn.net/weixin\_39842475/article/details/112187203