

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

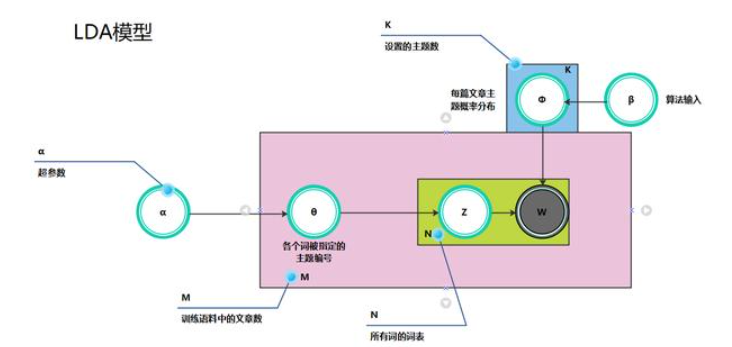
1.从Dirichlet分布中取样生成文档i的主题分布

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10\*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

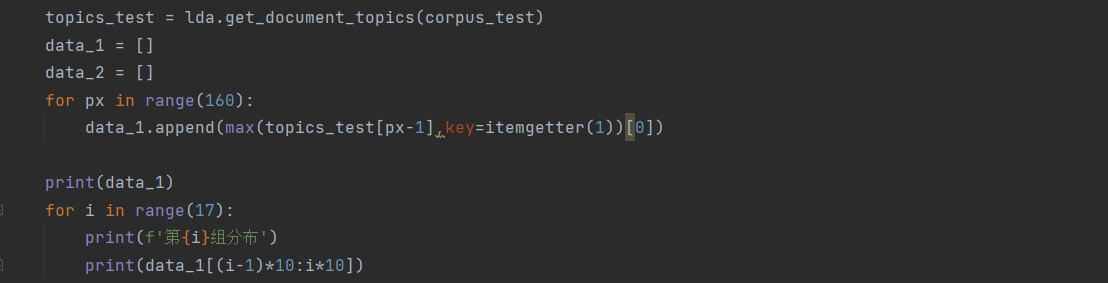
**注：最终每50行文本看作一个段落，每10个段落出自同一本小说，最后观察每10个段落在topic model下是否满足相同的分布**

****



利用gensim库进行LDA模型的训练，最后输出结果是以十六个主题的分布，以及选取的测试文本在训练出来的模型当中的适合程度

将试验结果分组输出，按每10个段落结果为一组



此时将每组中的topic最大值作为参考，认为当前最大值相同的段落拥有相同的topic分布，进而来自相同的小说。

2.试验结果

**16个主题的单词分布为：**

(0, '0.049\*"了" + 0.041\*"的" + 0.018\*"他" + 0.015\*"在" + 0.014\*"将" + 0.011\*"是" + 0.010\*"那" + 0.009\*"道" + 0.009\*"她" + 0.009\*"一声"')

(1, '0.027\*"任" + 0.011\*"六人" + 0.009\*"相救" + 0.009\*"招架" + 0.008\*"停" + 0.008\*"飞舞" + 0.008\*"仇人" + 0.008\*"力" + 0.008\*"凝视" + 0.007\*"怔怔"')

(2, '0.047\*"老者" + 0.032\*"崆峒" + 0.017\*"修为" + 0.011\*"不行" + 0.010\*"递" + 0.010\*"折" + 0.009\*"点中" + 0.008\*"插入" + 0.007\*"异状" + 0.007\*"挺剑"')

(3, '0.085\*"你" + 0.061\*"我" + 0.057\*"道" + 0.028\*"了" + 0.027\*"的" + 0.026\*"是" + 0.017\*"他" + 0.014\*"说" + 0.014\*"那" + 0.012\*"也"')

(4, '0.017\*"初时" + 0.015\*"奇" + 0.013\*"跪倒" + 0.011\*"睁开眼" + 0.011\*"剑锋" + 0.011\*"嗓子" + 0.009\*"双双" + 0.009\*"约莫" + 0.008\*"一撞" + 0.008\*"骇异"')

(5, '0.058\*"的" + 0.031\*"是" + 0.022\*"有" + 0.016\*"了" + 0.012\*"都" + 0.012\*"道" + 0.012\*"人" + 0.012\*"也" + 0.010\*"先生" + 0.010\*"那"')

(6, '0.023\*"挑" + 0.022\*"敌手" + 0.016\*"啷" + 0.014\*"又惊又喜" + 0.013\*"电子" + 0.013\*"新语丝" + 0.013\*"文库" + 0.013\*"得出" + 0.011\*"呛" + 0.011\*"国"')

(7, '0.015\*"一闪" + 0.014\*"招数" + 0.012\*"意料之外" + 0.012\*"一言不发" + 0.011\*"原是" + 0.009\*"大出" + 0.009\*"装" + 0.008\*"已久" + 0.008\*"徐达" + 0.008\*"威震"')

(8, '0.053\*"的" + 0.031\*"他" + 0.025\*"是" + 0.022\*"了" + 0.014\*"在" + 0.013\*"这" + 0.013\*"但" + 0.012\*"又" + 0.011\*"便" + 0.009\*"也"')

(9, '0.028\*"之" + 0.023\*"的" + 0.012\*"蒙古" + 0.010\*"招数" + 0.010\*"而" + 0.010\*"於" + 0.009\*"但" + 0.009\*"是" + 0.009\*"为" + 0.008\*"在"')

(10, '0.049\*"了" + 0.034\*"的" + 0.027\*"我" + 0.021\*"他" + 0.019\*"是" + 0.019\*"也" + 0.015\*"她" + 0.014\*"这" + 0.013\*"在" + 0.013\*"道"')

(11, '0.035\*"的" + 0.026\*"了" + 0.025\*"在" + 0.017\*"他" + 0.014\*"上" + 0.013\*"是" + 0.011\*"那" + 0.010\*"得" + 0.010\*"已" + 0.010\*"中"')

(12, '0.032\*"六" + 0.020\*"小兄弟" + 0.016\*"呸" + 0.013\*"数里" + 0.012\*"呻吟" + 0.011\*"山道" + 0.009\*"瞪视" + 0.008\*"多年" + 0.008\*"学过" + 0.007\*"重誓"')

(13, '0.042\*"头陀" + 0.031\*"中土" + 0.026\*"她们" + 0.023\*"力道" + 0.019\*"二来" + 0.014\*"颤" + 0.014\*"馀" + 0.013\*"打了个" + 0.012\*"声道" + 0.010\*"者"')

(14, '0.044\*"的" + 0.032\*"是" + 0.029\*"了" + 0.029\*"道" + 0.018\*"他" + 0.013\*"那" + 0.012\*"这" + 0.012\*"说道" + 0.011\*"教主" + 0.010\*"在"')

(15, '0.038\*"他" + 0.032\*"的" + 0.016\*"了" + 0.012\*"将" + 0.011\*"一剑" + 0.010\*"剑" + 0.009\*"这" + 0.009\*"长剑" + 0.009\*"在" + 0.009\*"一"')**对选取并且组合后的160个段落进行测试结果大致如下：**

将每个文本的最大topic统及如下，由于每本小说选择了10个500词以上的段落，所以将每一组统及在一起进行对比，全部topic分布可以见附录

第1组分布

[0, 10, 10, 10, 10, 10, 10, 10, 10, 10]

第2组分布

[10, 0, 10, 10, 10, 10, 10, 11, 11, 11]

第3组分布

[11, 11, 0, 11, 11, 11, 11, 11, 11, 11]

第4组分布

[11, 11, 11, 0, 10, 10, 10, 10, 10, 10]

第5组分布

[10, 10, 10, 10, 0, 14, 14, 14, 14, 14]

第6组分布

[14, 14, 14, 14, 14, 0, 14, 14, 14, 14]

第7组分布

[14, 14, 14, 14, 14, 14, 0, 10, 11, 11]

第8组分布

[11, 10, 10, 10, 10, 11, 10, 0, 10, 10]

第9组分布

[10, 10, 10, 10, 10, 10, 10, 10, 0, 11]

第10组分布

[11, 11, 11, 11, 11, 11, 11, 11, 10, 0]

第11组分布

[10, 10, 10, 10, 10, 10, 10, 10, 10, 10]

第12组分布

[0, 10, 10, 10, 10, 10, 10, 10, 10, 10]

第13组分布

[10, 0, 10, 10, 10, 10, 10, 10, 10, 10]

第14组分布

[10, 10, 0, 10, 10, 10, 10, 10, 10, 10]

第15组分布

[10, 10, 10, 0, 10, 10, 10, 10, 10, 10]

第16组分布

[10, 10, 10, 10, 0, 11, 11, 11, 11, 11]

上面的结果可以看出，有大量的结果都分布在0，10，11，回到16个主题分布处看出，这三个主题的词语大多是虚词介词等，这样的词语大量存在于金庸的小说中，所以文本分类的效果一般。

## 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，比如试验结果中的10，11主题，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，或者在初始对金庸小说处理时，对的，了，等词语进行删除处理，更多的保留小说本身的一些特征词语，从而提升模型判断的准确性。**

**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