基于Topic Model的中文文本分类

题目：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 模型进行文本建模，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。

# LDA模型

LDA(Latent Dirichlet Allocation)是一种文档主题生成模型，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。所谓生成模型，我们认为一篇文章的每一个词都是通过“文章以一定的概率选择了某一主题，并从这个主题中以一定的概率选择某一词语”这个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

对于语料库中的每篇文档，LDA定义了如下的生成过程：

1. 对于每一篇文档，从主题分布中抽取一个主题。
2. 从上述被抽到的主题所对应的单词分布中抽取一个单词。
3. 重复上述过程直至遍历文档中的每一个单词。

图示, 示意图

描述已自动生成

1. Theta文档-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alpha，theta的每一行数据是一个K维向量（语料库共有K个主题），比如（1，0，0，1，0，1），表示该文档包含那些主题以及对应的概率。
2. Beta主题-词语分布是多项式分布，该多项式分布的参数服从Dirichlet分布，该Dirichlet分布的参数为nita；beta的每一行是一个V维的向量，表示给主题包含哪些词语以及对应的概率。
3. Z是从theta中抽取出了的一个主题，是一个k维向量，比如（0，0，0，1，0，0）.
4. W是从z这个主题及其对应的词语中抽取出来的一个词语（观测值）
5. 重复3-4步骤直至遍历该文档中的每一个单词，然后遍历下一个文档直至m篇文档全部完成。
6. 选择m个待分类文档，再逐个抽完n个词后，还原观测值此时将图中最高的柱状图对应的主题找出来分析这些主题中出现最高频的词语，根据这些词语，认为定义主题分类名称。

## LDA训练算法

1. 随机初始化alpha、beta
2. 以下步骤迭代直至收敛：
3. 对训练集中的每篇文档利用当前的alpha和beta值计算每篇文档的主题分布、每个词所属的主题分布。
4. 积累所有文档中，属于主题K的词的个数，得到gammas;以及词i属于主题k的次数，得到矩阵batas；
5. 根据当前的gammas，利用Newton-Raphson迭代方法求得当前的最优α值；
6. 对矩阵betas的列归一化，直接得到当前的β值，即每个主题的词的分布；
7. 输出达到收敛时的α和β的值

## LDA预测算法

1. 以平均分布初始化K维向量nt，ntk是当前文档中属于类别k的词的个数，nt可视为未归一化的文档的主题分布；
2. 以下步骤迭代直到nt达到稳定：
3. 根据当前的α值(决定主题的先验分布)，以及当前的nt值(当前文档的主题分布)，以及当前的β值(主题的词的分布)，计算文档中的各个词的主题分布，得到矩阵q， qij=文档中的第i个词属于主题k的概率。
4. 利用矩阵q的值更新向量nt的值。
5. 将nt归一化作为文档的主题分布，矩阵q则为文档中每个词的主题分布。

# 文本分类

文本分类是指在给定分类体系，根据文本内容自动确定文本类别的过程。最基础的分类是归到两个类别中，称为二分类问题，例如电影评论分类，只需要分为“好评”和“差评”。分到多个类别中的称为多分类问题，例如，把名字分类为法语名字、英语名字、西班牙语名字。

## 步骤

一般来说文本分类大致分为以下几个步骤：

1. 定义阶段：定义数据以及分类体系，具体分为哪些类别，需要哪些数据。
2. 数据预处理：对文档做分词、去停用词等准备工作。
3. 数据提取特征：对文档矩阵进行降维，提取训练集中最有用的特征。
4. 模型训练阶段：选择具体的分类模型以及算法，训练出文本分类器。
5. 评测阶段：在测试集上测试并评价分类器的性能。
6. 应用阶段：应用性能最高的分类模型对待分类文档进行分类。

# 实验过程

本次实验继续使用金庸先生的16本武侠小说作为数据集，利用LDA进行文本分类。

## 数据预处理

在这一环节中，删除文本的所有隐藏符号，删除所有的非中文字符，不考虑上下文关系的前提下删去所有的标点符号。以jieba库对中文语料进行分词。得到训练集。

训练集内容如下，共59202行：

日历

描述已自动生成

## 运行LDA进行训练并验证

本次实验使用了gensim中的corpora和其中自带的lda模型进行训练。

文本

描述已自动生成

使用验证集进行验证，测试训练结果：

文本

描述已自动生成

## 实验结果

16个主题的词语分布结果如下：

图片包含 背景图案

描述已自动生成

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 主题 | 词及其对应的概率 | | | | | | | | | |
| 0 | 0.038言语 | 0.024面前 | 0.024及 | 0.019觉 | 0.019或 | 0.015她们 | 0.013应道 | 0.012修为 | 0.011自幼 | 0.010服饰 |
| 1 | 0.067两位 | 0.020发作 | 0.018衣袖 | 0.015诀 | 0.013门户 | 0.013窜 | 0.013迅速 | 0.011无异 | 0.010偏 | 0.010大殿 |
| 2 | 0.084剑法 | 0.041无法 | 0.025未 | 0.024多半 | 0.018撞 | 0.016之色 | 0.010大门 | 0.010传授 | 0.010眼前 | 0.009倒退 |
| 3 | 0.068罢 | 0.050兄弟 | 0.025否则 | 0.020即 | 0.015显然 | 0.012别说 | 0.012按 | 0.011不如 | 0.010一道 | 0.010从此 |
| 4 | 0.042的 | 0.041了 | 0.039我 | 0.038道 | 0.032是 | 0.025你 | 0.024也 | 0.020说 | 0.019他 | 0.016那 |
| 5 | 0.023少女 | 0.018了 | 0.018那 | 0.015抢 | 0.013老人家 | 0.013正 | 0.013出 | 0.012一位 | 0.009在 | 0.009不及 |
| 6 | 0.022此刻 | 0.014在 | 0.014人物 | 0.014不得 | 0.014其中 | 0.013如此 | 0.012的 | 0.012是 | 0.012出 | 0.010抓 |
| 7 | 0.032写 | 0.024皇帝 | 0.021一日 | 0.018相见 | 0.017怪 | 0.016右掌 | 0.016位 | 0.015那姓 | 0.015旁观 | 0.014般的 |
| 8 | 0.017出 | 0.014在 | 0.013武林中 | 0.012的 | 0.009便 | 0.008下 | 0.008得 | 0.008往 | 0.008与 | 0.008若 |
| 9 | 0.060的 | 0.042了 | 0.032他 | 0.025是 | 0.023在 | 0.014这 | 0.013她 | 0.011又 | 0.011便 | 0.010将 |
| 10 | 0.030\*"此处 | 0.026\*"房中 | 0.023\*"惊 | 0.021\*"伤势 | 0.020\*"生气 | 0.017\*"转念 | 0.016\*"见识 | 0.014\*"一颗 | 0.012\*"不管 | 0.011\*"连声 |
| 11 | 0.155你 | 0.096我 | 0.080道 | 0.028了 | 0.016叫 | 0.012好 | 0.011不 | 0.011说道 | 0.010他 | 0.010笑 |
| 12 | 0.033不料 | 0.017决不能 | 0.016落入 | 0.015猛 | 0.012发 | 0.011有意 | 0.011那边 | 0.011一层 | 0.010一年 | 0.009臂 |
| 13 | 0.025得 | 0.025左手 | 0.020长剑 | 0.018一声 | 0.018右手 | 0.015已 | 0.015向 | 0.015听 | 0.011之声 | 0.010只 |
| 14 | 0.060师哥 | 0.043武林 | 0.039张 | 0.026怎能 | 0.019纷纷 | 0.019而已 | 0.016西域 | 0.014劈 | 0.012四下 | 0.011通红 |
| 15 | 0.029尽数 | 0.026也好 | 0.018老子 | 0.017断 | 0.016抵挡 | 0.015足 | 0.014剩下 | 0.012呸 | 0.012是从 | 0.012一副 |

## 测试结果

认为选取几本小说内的一些段落，将其文本进行预处理后作为测试集，测试LDA模型对于文本的分类效果。得到不同测试段落的主题分布。

这里随机选取几段，将其分词后改成一行，结果如下：（这里仅展示了一段中的部分片段）





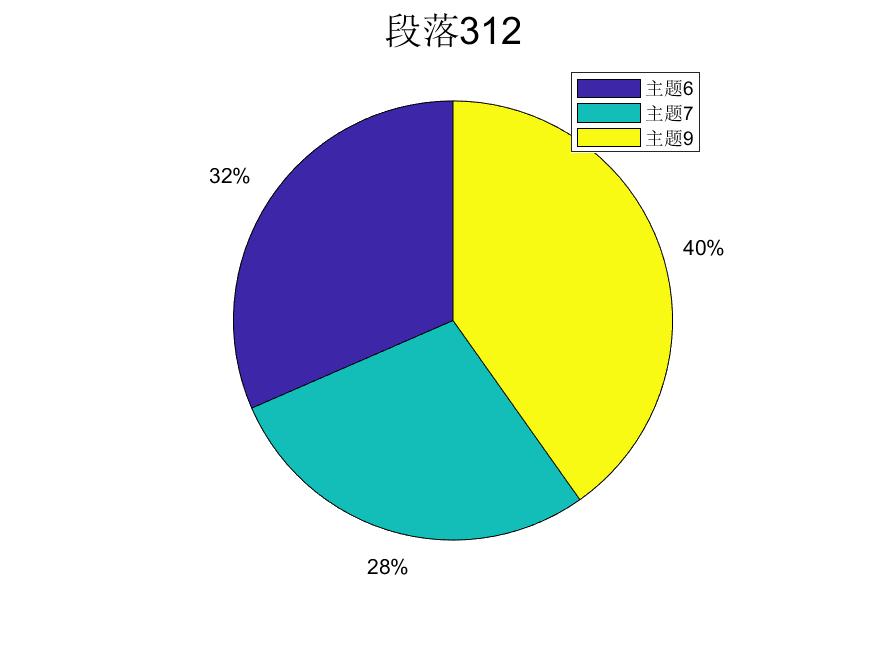
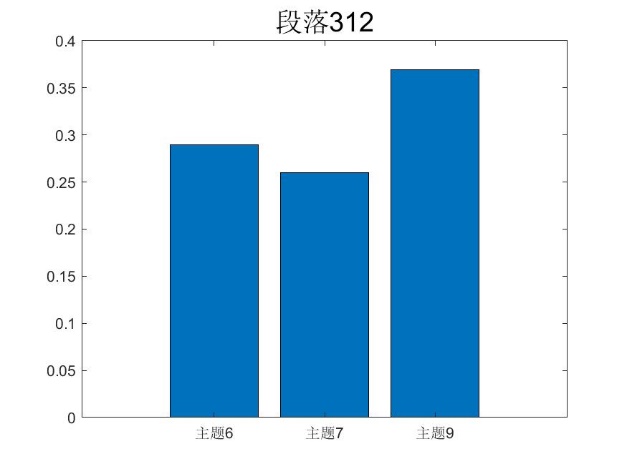


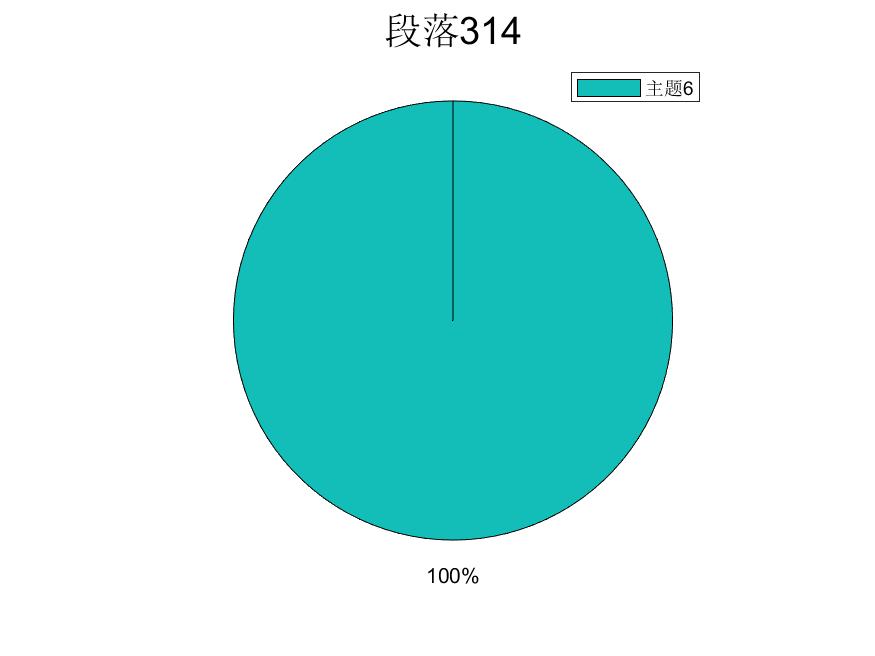
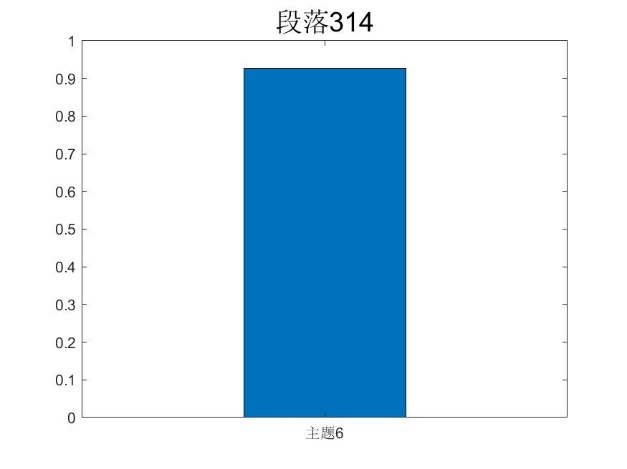
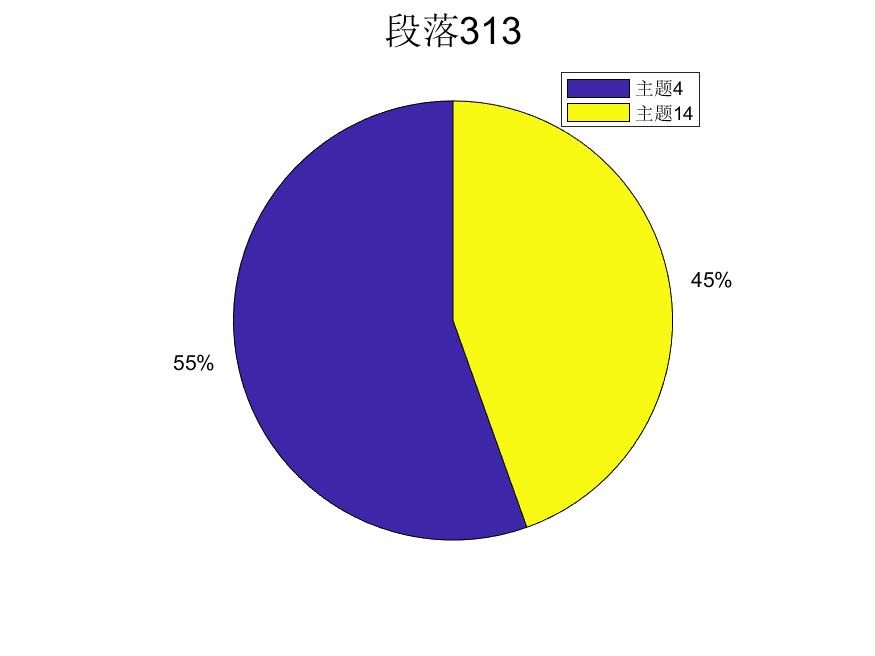
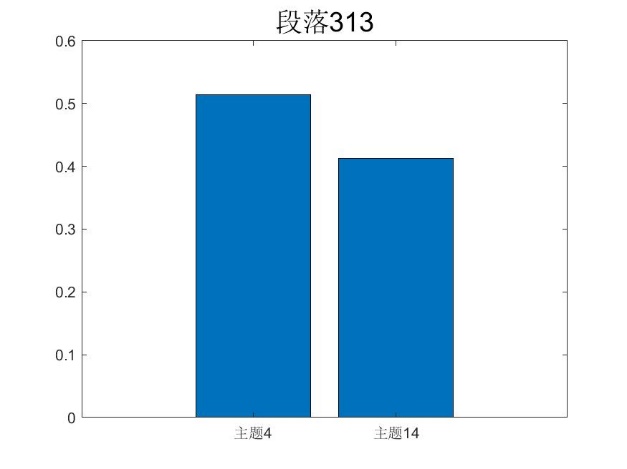
文本

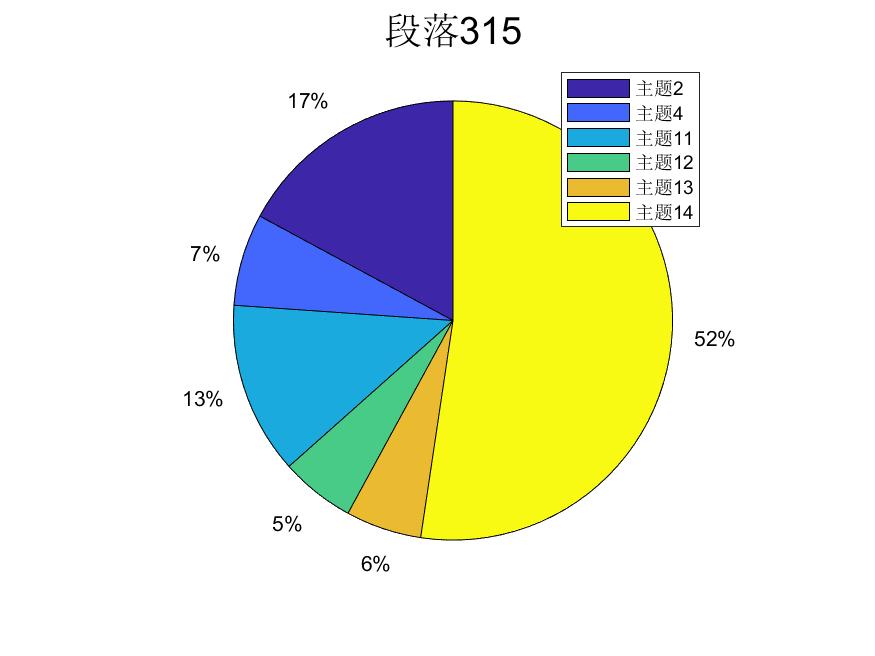
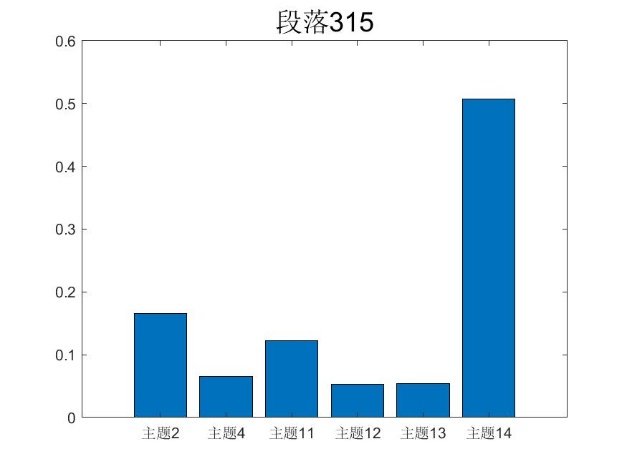
描述已自动生成

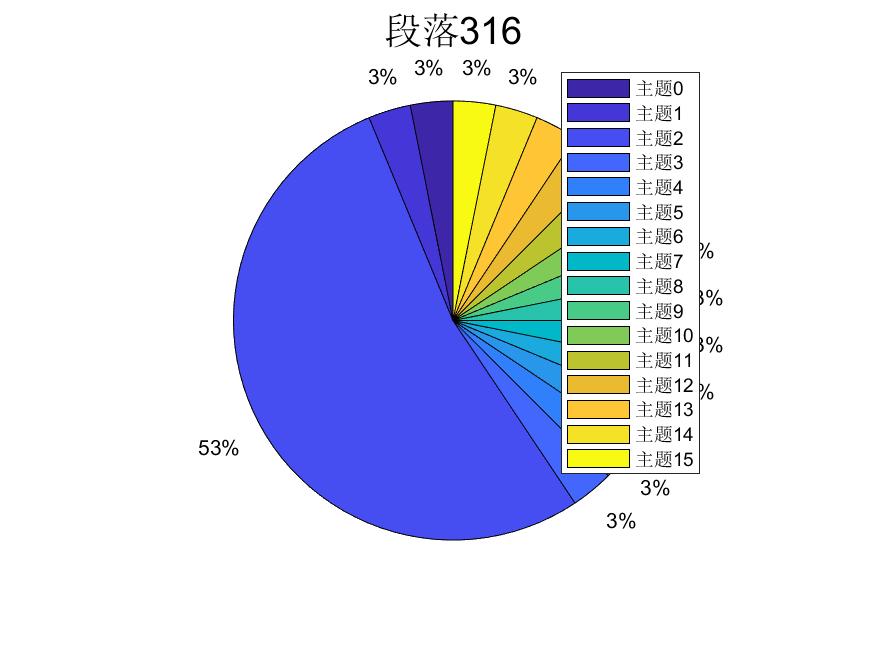
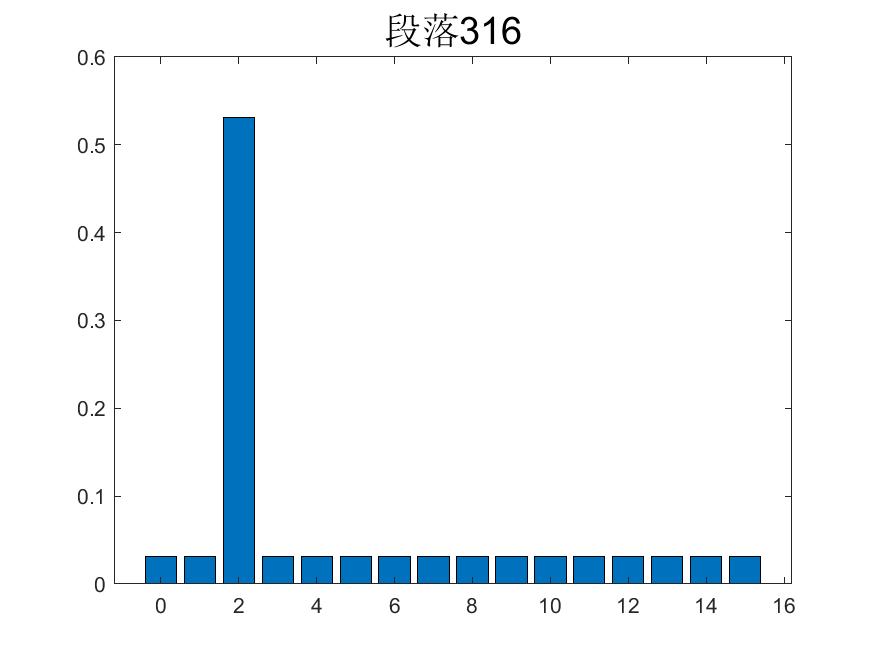
本次实验共选取了317个段落进行测试，测试结果如上图所示：（这里选取最后5个进行分析）。

为了直观表示，将以上主题分布的预测结果数据导入MATLAB进行绘图分析。









## 结果分析

从段落316的结果我们可以看出，该段落由主题2生成的概率最大，占到了总概率的百分之50以上。而段落314几乎全部由主题6所生成；段落312，和段落313由其中几种主题所生成的概率比较平均。

# 参考文献

<https://blog.csdn.net/weixin_42663984/article/details/116264233>