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。

二、实验原理

1. LDA 模型简介

LDA（Latent Dirichlet Allocation）由 Blei, David M.、吴恩达和 Jordan, Michael I 于 2003 年提出，用来推测文档的主题分布。它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题分布后，便可以根据主题分布进行主题聚类或文本分类。

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。LDA 模型通过概率模型建立了从主题到文档中每个词的关系，利用贝叶斯概率，能通过文档的词推导出文档的主题。

另外，正如 Beta 分布是二项式分布的共轭先验概率分布，狄利克雷分布作为多项式分布的共轭先验概率分布。因此正如 LDA 贝叶斯网络结构中所描述的，在 LDA 模型中一篇文档生成的方式如下：

从狄利克雷分布α中取样生成文档i的主题分布θ_i

从主题的多项式分布θ_i中取样生成文档i第j个词的主题$z_{i,j}$

从狄利克雷分布β中取样生成主题$z_{i,j}$的词语分布$\phi_{z_{i,j}}$

从词语的多项式分布$\phi_{z_{i,j}}$中采样最终生成词语$\omega_{i,j}$

因此整个模型中所有可见变量以及隐藏变量的联合分布是

$$p(\omega_i, z_i, \theta_i, \phi|\alpha, \beta) = \prod_{j=1}^{N} p(\theta_i|\alpha)p\left(z_{i,j}\middle|\theta_i\right)p(\phi|\beta)p\left(\omega_{i,j}\middle|\phi_{z_{i,j}}\right)$$

最终一篇文档的单词分布的最大似然估计可以通过将上式的θ_i以及ϕ进行积分和对z_i进行求和得到

$$p(\omega_i|\alpha, \beta) = \int_{\theta_i} \int_{\phi} \sum_{z_i} p(\omega_i, z_i, \theta_i, \phi|\alpha, \beta)$$

根据$p(\omega_i|\alpha,\beta)$的最大似然估计，最终可以通过吉布斯采样等方法估计出模型中的参数。

三、实验过程与结果

1. 实验过程

选取金庸的 16 本小说为数据集，依次为：白马啸西风，碧血剑，飞狐外传，连城诀，鹿鼎记，三十三剑客图，射雕英雄传，神雕侠侣，书剑恩仇录，天龙八部，侠客行，笑傲江湖，雪山飞狐，倚天屠龙记，鸳鸯刀，越女剑。

首先对文本进行预处理，同作业一中相似，去除特殊字符，但保留换行。

因为实验要求一共选取 200 个段落，每个段落大于 500 词。由于小说文件中有许多换行，每行至少有 10 个词，不妨取 50 行合为一段，视为大于 500 词的段落，在每部小说中抽取 13 个段落，作为测试文本。越女剑篇幅较短，仅能选取 2 个大于 500 词的段落，故段落总数约为 200。总集中除去测试文本外的部分为训练文本。

使用 gensim 包中的 corpora 和其中自带的 lda 模型来进行训练，输出通过训练文本得到的 16 个主题，以及测试文本隶属于上述主题的概率。

2. 实验结果

初次进行实验时，由于小说中存在过多“的”、“了”之类的虚词或助词，严重影响了主题的区分度，导致生成的主题如下图所示：

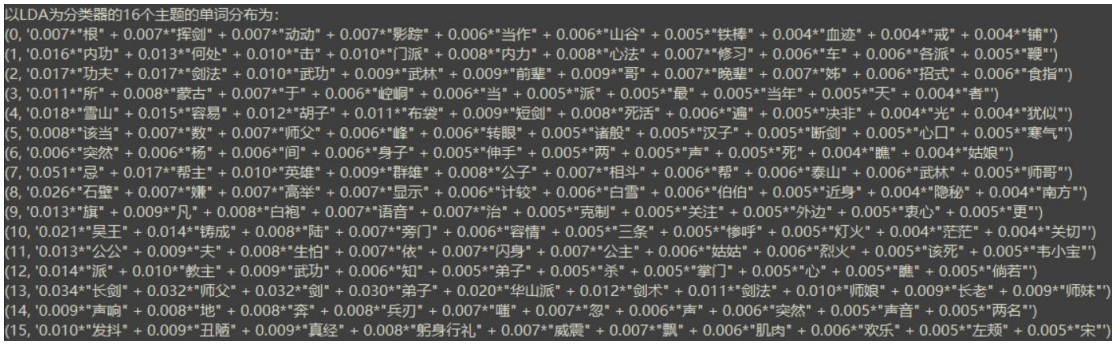
```
以LDA为分类器的16个主题的单词分布为:
(0, '0.046*"的" + 0.025*"了" + 0.024*"是" + 0.013*"在" + 0.012*"和" + 0.012*"他" + 0.011*"道" + 0.011*"有" + 0.011*"都" + 0.011*"也"')
(1, '0.022*"於" + 0.014*"深" + 0.013*"胡子" + 0.011*"心口" + 0.011*"至尊" + 0.009*"犹似" + 0.007*"守势" + 0.007*"终" + 0.007*"曦里" + 0.007*"石头"')
(2, '0.049*"的" + 0.031*"他" + 0.019*"是" + 0.018*"这" + 0.015*"但" + 0.012*"在" + 0.012*"武功" + 0.011*"了" + 0.010*"却" + 0.009*"剑法"')
(3, '0.019*"青衣" + 0.012*"四下里" + 0.012*"一辈子" + 0.010*"猛地" + 0.010*"自有" + 0.010*"良久" + 0.010*"岩石" + 0.009*"倾听" + 0.009*"这一掌" + 0.009*"刺入"')
(4, '0.017*"双眼" + 0.016*"莫非" + 0.015*"娶" + 0.015*"修为" + 0.013*"这一" + 0.012*"女" + 0.011*"并肩" + 0.010*"广场" + 0.009*"直至" + 0.008*"醒"')
(5, '0.046*"了" + 0.028*"的" + 0.024*"她" + 0.024*"他" + 0.021*"在" + 0.017*"我" + 0.015*"是" + 0.014*"道" + 0.013*"你" + 0.009*"去"')
(6, '0.070*"我" + 0.058*"你" + 0.032*"的" + 0.029*"道" + 0.028*"了" + 0.027*"是" + 0.024*"他" + 0.019*"也" + 0.016*"说" + 0.013*"这"')
(7, '0.010*"徐" + 0.008*"教主" + 0.007*"惨呼" + 0.007*"教众" + 0.007*"哼哼" + 0.007*"响" + 0.007*"停留" + 0.006*"境界" + 0.006*"传出" + 0.006*"引"')
(8, '0.026*"她" + 0.017*"的" + 0.014*"是" + 0.011*"他" + 0.010*"身受" + 0.009*"父母" + 0.007*"了" + 0.007*"在" + 0.006*"心神" + 0.006*"能够"')
(9, '0.053*"先生" + 0.015*"什" + 0.012*"修习" + 0.011*"曰" + 0.010*"尴尬" + 0.010*"其" + 0.009*"的" + 0.008*"亦" + 0.008*"之" + 0.007*"为"')
(10, '0.037*"的" + 0.029*"了" + 0.020*"在" + 0.019*"他" + 0.013*"将" + 0.011*"得" + 0.010*"上" + 0.010*"便" + 0.010*"向" + 0.009*"那"')
(11, '0.021*"心念" + 0.019*"臭" + 0.016*"射" + 0.014*"一闪" + 0.012*"谢" + 0.009*"七八" + 0.007*"一痛" + 0.007*"关注" + 0.007*"之气" + 0.006*"埋"')
(12, '0.028*"石壁" + 0.020*"破绽" + 0.017*"正好" + 0.016*"挑" + 0.010*"喜" + 0.009*"几日" + 0.008*"胡" + 0.007*"壁上" + 0.007*"占上风" + 0.007*"毒"')
(13, '0.034*"不料" + 0.022*"何处" + 0.014*"松树" + 0.011*"铺" + 0.010*"得出" + 0.010*"烈火" + 0.009*"不约而同" + 0.009*"剑锋" + 0.009*"马蹄声" + 0.008*"绝"')
(14, '0.034*"长老" + 0.014*"答话" + 0.013*"意料之外" + 0.013*"那人道" + 0.011*"瞪" + 0.011*"同道" + 0.010*"大出" + 0.009*"数里" + 0.008*"数十丈" + 0.008*"钻入"')
(15, '0.047*"道" + 0.044*"的" + 0.037*"了" + 0.030*"你" + 0.029*"是" + 0.022*"那" + 0.015*"我" + 0.012*"这" + 0.011*"说" + 0.010*"他"')
```

最终分类大多集中在虚词较多的主题上，分类效果十分不好。

对文本中的虚词进行删除处理，具体如下图：

```
word_to_be_replaced = ['的', '了', '是', '在', '和', '他', '道', '有', '都', '也', '於', '这', '但', '却',
                       '她', '我', '你', '说', '曰', '得', '其', '亦', '为', '之', '那', '便', '将', '同',
                       '不', '去', '来', '好', '着', '要', '甚', '么', '又', '不', '人', '打', '只', '来',
                       '去', '给', '与', '以', '到', '中', '而', '可', '等', '上', '下', '向', '已', '还',
                       '就', '等', '大', '过', '无', '跟', '咱', '一', '个', '没', '谁', '问', '众', '被',
                       '叫', '见', '听', '再', '们', '起', '话', '走', '后', '从', '对', '请', '时', '或',
                       '能', '麽', '什', '做', '想', '啊', '出', '笑', '想', '啦', '才', '会', '正', '看',
                       '自己', "如何", "如此"]
```

处理后生成的主题如下图所示：



此时对测试文本的分类效果略好。例如：对于来自鹿鼎记的 13 个段落，其属于主题 11 的概率显著大于来自其他小说的测试文本。但整体效果还是较弱，分类结果大多集中在主题 6 与主题 12 上，见“result.txt”，没有明显的分类产生。

四、结果分析

此次实验中，即使剔除了部分虚词的影响，LDA 模型对金庸小说的主题划分和文本分类效果也较为有限。原因是实验文本类型均为武侠小说，结果分布较为集中的主题 6 与主题 12 中，都是武侠小说中常见的词语，如“派”、“武功”、“弟子”等，并不具有特殊性，分类效果不甚显著。

五、参考资料

[1] 基于 Topic model 的中文文本分类

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