机器学习与自然语言处理实验三

院（系）名称	自动化科学与电气工程学院
学生姓名	赵怡然
学生学号	ZY2103208
指导老师	秦曾昌

2022 年 05 月 06 日

一、 实验背景介绍

隐含狄利克雷分布（Latent Dirichlet Allocation，LDA），是一种主题模型（topic model），它可以将文档集中每篇文档的主题按照概率分布的形式给出。

二、 实验目标

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(Latent Dirichlet Allocation)主题模型进行文本建模,并把每个段落表示为主题分布后进行分类。

三、 LDA 模型与 Gibbs 采样

3.1 LDA 模型

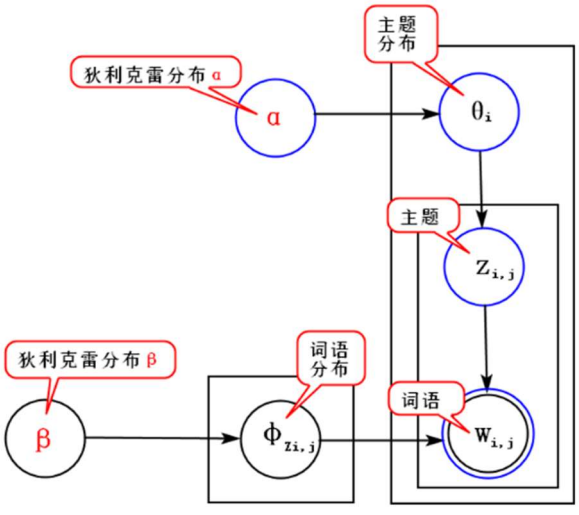
在 LDA 模型中，一篇文档生成的方式如下：

1．从狄利克雷分布α中取样生成文档 i 的主题分布θ_i

2．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$z_{i,j}$

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

4．从词语的多项式分布$\Phi_{z_{i,j}}$中采样最终生成词语$w_{i,j}$



根据上述生成文档方式，生成一篇文档的联合概率分布为：

$$p(\vec{w}_m, \vec{z}_m, \vec{\vartheta}_m, \underline{\Phi}|\vec{\alpha}, \vec{\beta}) = \prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}}) p(z_{m,n}|\vec{\vartheta}_m) \cdot p(\vec{\vartheta}_m|\vec{\alpha}) \cdot p(\underline{\Phi}|\vec{\beta})$$

3.2 Gibbs 采样

LDA 主题模型完成了文档的建模后需要对模型进行参数估计。参数估计根据联合概率求得给定观测量 w 下的隐含量 z 的条件分布并构造马尔科夫链通过 Gibbs 采样完成样本的主题分布和词分布的训练。

四、 实验结果与分析

4.1 实验流程

根据上述 LDA 原理完成给定语料库的文本建模后，均匀采样来自不同小说的 200 个段落（每个段落词语个数大于 500），进而得到各个段落的主题分布估计。

由 200 个段落的主题分布以及其对应的小说名称，将数据按 7:3 比例分为训练集与测试集。通过支持向量机（SVM）完成由各段落主题分布归类于各小说名称的分类器的训练与测试。

4.2 实验结果

（1）实验参数：100 个主题特征，200 个段落，每个段落不少于 500 字

训练集结果如下：

训练集预测结果如下：

	precision	recall	f1-score	support
三十三剑客图	1.00	1.00	1.00	10
书剑恩仇录	1.00	1.00	1.00	8
侠客行	1.00	1.00	1.00	10
倚天屠龙记	1.00	1.00	1.00	9
天龙八部	1.00	1.00	1.00	11
射雕英雄传	1.00	1.00	1.00	12
白马啸西风	1.00	1.00	1.00	10
碧血剑	1.00	1.00	1.00	10
神雕侠侣	1.00	1.00	1.00	6
笑傲江湖	1.00	0.88	0.93	8
越女剑	1.00	1.00	1.00	6
连城诀	0.89	1.00	0.94	8
雪山飞狐	1.00	1.00	1.00	6
飞狐外传	1.00	1.00	1.00	9
鸳鸯刀	1.00	1.00	1.00	8
鹿鼎记	1.00	1.00	1.00	9
accuracy			0.99	140
macro avg	0.99	0.99	0.99	140
weighted avg	0.99	0.99	0.99	140

测试集结果如下：

测试集预测结果如下：

	precision	recall	f1-score	support
三十三剑客图	0.33	0.67	0.44	3
书剑恩仇录	1.00	1.00	1.00	5
侠客行	1.00	0.33	0.50	3
倚天屠龙记	0.60	0.75	0.67	4
天龙八部	0.67	1.00	0.80	2
射雕英雄传	1.00	1.00	1.00	1
白马啸西风	1.00	1.00	1.00	3
碧血剑	1.00	1.00	1.00	3
神雕侠侣	1.00	0.33	0.50	6
笑傲江湖	0.67	0.50	0.57	4
越女剑	1.00	0.83	0.91	6
连城诀	0.67	0.50	0.57	4
雪山飞狐	0.75	0.50	0.60	6
飞狐外传	1.00	1.00	1.00	3
鸳鸯刀	0.43	0.75	0.55	4
鹿鼎记	0.50	1.00	0.67	3
accuracy			0.72	60
macro avg	0.79	0.76	0.74	60
weighted avg	0.80	0.72	0.72	60

（2）实验参数：100 个主题特征，200 个段落，每个段落不少于 700 字

训练集结果如下：

训练集预测结果如下：

	precision	recall	f1-score	support
三十三剑客图	1.00	0.67	0.80	9
书剑恩仇录	1.00	1.00	1.00	8
侠客行	1.00	1.00	1.00	8
倚天屠龙记	1.00	1.00	1.00	8
天龙八部	1.00	1.00	1.00	8
射雕英雄传	1.00	1.00	1.00	9
白马啸西风	1.00	1.00	1.00	7
碧血剑	0.89	0.80	0.84	10
神雕侠侣	1.00	1.00	1.00	11
笑傲江湖	0.55	0.67	0.60	9
越女剑	1.00	1.00	1.00	8
连城诀	1.00	0.60	0.75	10
雪山飞狐	0.85	1.00	0.92	11
飞狐外传	1.00	1.00	1.00	9
鸳鸯刀	1.00	0.71	0.83	7
鹿鼎记	0.50	0.88	0.64	8
accuracy			0.89	140
macro avg	0.92	0.90	0.90	140
weighted avg	0.92	0.89	0.90	140

测试集结果如下：

测试集预测结果如下：

	precision	recall	f1-score	support
三十三剑客图	1.00	0.50	0.67	4
书剑恩仇录	1.00	1.00	1.00	5
侠客行	1.00	1.00	1.00	5
倚天屠龙记	1.00	1.00	1.00	5
天龙八部	1.00	1.00	1.00	5
射雕英雄传	1.00	1.00	1.00	4
白马啸西风	1.00	1.00	1.00	6
碧血剑	1.00	0.67	0.80	3
神雕侠侣	0.50	1.00	0.67	1
笑傲江湖	0.33	0.33	0.33	3
越女剑	1.00	1.00	1.00	4
连城诀	0.33	0.50	0.40	2
雪山飞狐	0.50	1.00	0.67	1
飞狐外传	1.00	1.00	1.00	3
鸳鸯刀	1.00	0.60	0.75	5
鹿鼎记	0.67	1.00	0.80	4
accuracy			0.87	60
macro avg	0.83	0.85	0.82	60
weighted avg	0.91	0.87	0.87	60

（3）实验参数：20 个主题特征，200 个段落，每个段落不少于 500 字

训练集结果如下：

训练集预测结果如下：

	precision	recall	f1-score	support
三十三剑客图	0.90	0.90	0.90	10
书剑恩仇录	0.53	0.89	0.67	9
侠客行	1.00	1.00	1.00	8
倚天屠龙记	0.91	0.83	0.87	12
天龙八部	0.69	0.82	0.75	11
射雕英雄传	1.00	1.00	1.00	13
白马啸西风	0.60	0.43	0.50	7
碧血剑	0.75	0.75	0.75	8
神雕侠侣	0.70	0.70	0.70	10
笑傲江湖	1.00	1.00	1.00	8
越女剑	1.00	1.00	1.00	8
连城诀	0.75	0.67	0.71	9
雪山飞狐	0.50	0.17	0.25	6
飞狐外传	0.50	0.50	0.50	8
鸳鸯刀	1.00	1.00	1.00	6
鹿鼎记	1.00	1.00	1.00	7
accuracy			0.81	140
macro avg	0.80	0.79	0.79	140
weighted avg	0.81	0.81	0.80	140

测试集结果如下：

测试集预测结果如下：

	precision	recall	f1-score	support
三十三剑客图	0.50	0.67	0.57	3
书剑恩仇录	0.20	0.25	0.22	4
侠客行	1.00	0.80	0.89	5
倚天屠龙记	0.25	1.00	0.40	1
天龙八部	0.40	1.00	0.57	2
白马啸西风	0.33	0.17	0.22	6
碧血剑	0.50	0.40	0.44	5
神雕侠侣	0.33	0.50	0.40	2
笑傲江湖	1.00	1.00	1.00	4
越女剑	0.67	0.50	0.57	4
连城诀	0.00	0.00	0.00	3
雪山飞狐	1.00	0.33	0.50	6
飞狐外传	0.43	0.75	0.55	4
鸳鸯刀	0.80	0.67	0.73	6
鹿鼎记	0.83	1.00	0.91	5
accuracy			0.57	60
macro avg	0.55	0.60	0.53	60
weighted avg	0.61	0.57	0.56	60

4.3 结果分析

分类器准确率随 LDA 主题模型主题特征个数增加而增加，100 主题特征>20 主题特征；随每段文本的长度增加而增加，700 词以上>500 词以上。