深度学习与自然语言处理第三次作业

姓名：郑梓岳　　学号：19231191

一、作业描述

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。

（1）在不同数量的主题个数下分类性能的变化；（2）以"词"和以"字"为基本单元下分类结果有什么差异？

二、LDA(Latent Dirichlet Allocation)主题模型

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 从主题分布中抽取一个主题，然后从该主题对应的单词分布中抽取一个单词，重复上述过程直到遍历文档中的每一个单词。具体流程如下图所示：

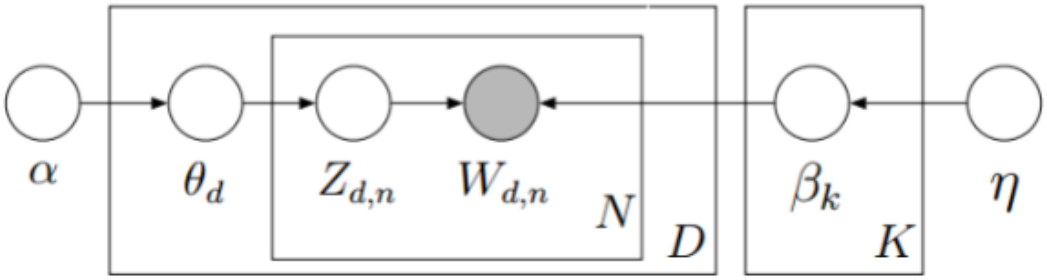

图 1 LDA 模型生成过程

具体步骤为：

1. 从 Dirichlet 分布α中取样生成文档d的主题分布θ_d；
2. 从主题的多项式分布θ_d中取样生成文档d的第n个词的主题$Z_{d,n}$；
3. 从 Dirichlet 分布η中取样生成主题$Z_{d,n}$对应的词语分布β_k；
4. 从词语的多项式分布β_k中采样最终生成词语$W_{d,n}$。

其中，θ_d主题文档-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α，θ_d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；β_k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η；β_k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；$Z_{d,n}$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）；$W_{d,n}$是从$Z_{d,n}$这个主题及其对应的词语中抽取出来的一个词语（观测值）。

三、实验过程设计

3.1 预料预处理

以金庸的 16 本小说作为语料库，进行预处理。预处理包括文本清洗和使用 jieba 分词并创建训练集和测试集。文本清洗使用百度停用词表 baidu_stopwords.txt。

3.2 LDA 模型构建及训练

本次作业使用了 gensim 库中的 corpora 和其中的 lda 模型进行训练。利用 corpora.Dictionary()构建词典，利用 doc2bow()计算词向量。利用 gensim.,odels.LdaModel()构建模型，选择构建的主题为 16。主题分别对应语料库中的 16 本小说。

```
dictionary=corpora.Dictionary(content)
dictionary.filter_n_most_frequent(100)
corpus = [dictionary.doc2bow(text) for text in content ]
lda = models.LdaModel(corpus=corpus, id2word=dictionary, num_topics=16)
```

图 2 LDA 模型创建及训练代码

3.3 预测文本主题

使用测试集测试训练好的 LDA 模型

```
corpus_test = [dictionary.doc2bow(text) for text in test]
topics_test = lda.get_document_topics(corpus_test)
labels=list(names)
for i in range(208):
    label=labels[int(i//13)].replace('.txt','')
    print(label+'的段落的主题分布为: \n')
    print(topics_test[int(i//13)],'\n')
```

图 3 最终测试训练好的 LDA 代码

四、实验结果及分析

4.1 实验结果

下表展示前 5 个主题的主要词概率分布：

Topic	Top 10 Keywords
0	麽，韦小宝，一声，爹爹，李文秀，姑娘，请，剑，心想，问
1	一声，韦小宝，麽，后，张无忌，心想，身子，少女，派，令狐冲
2	麽，一声，著，剑，剑士，心想，汉子，手中，范蠡，爹爹
3	麽，韦小宝，一声，李文秀，心想，后，著，爹爹，陈家洛，便是
4	一声，韦小宝，麽，后，剑，剑士，心想，爹爹，姑娘，范蠡
5	麽，范蠡，剑士，一声，李文秀，著，心想，剑，韦小宝，姑娘

当以 16 个主题数进行分类时，正确率为

```
0.9230769230769231
```

当以 5 个主题数进行分类时，正确率为

```
0.9134615384615384
```

当以 20 个主题数进行分类时，正确率为

```
0.9182692307692307
```

当以字为单位进行分类时，正确率为

```
0.9086538461538461
```

4.2 分析及总结

由图可得，当以 16 个主题且以词为单位进行分类时，正确率最高，当改变主题数时，无论是增加还是减少主题都会使正确率下降，但是正确的下降不多，可以认为主题数对于分类的结果影响不大。当以字为单位进行分类时， 也会导致正确率下降，初步认为是因为金庸先生的小说中有许多重复用字，写作风格相近，使得模型难以分辨主题，当以词进行分类时便能够避免一部分的重复用字，使正确率上升。

通过本次作业，我对 LDA 模型有了初步的理解和认识，初步掌握了 LDA 建模的过程和训练过程。