# LDA主题模型对于文本的建模与分类

刘峻池
2252298@qq.com

## 摘要

本实验以作家金庸的16本中文小说作为语料库，分别对每本小说进行了数据预处理与片段选取，在实验中建立了LDA主题模型，对于语料库中的内容主题分类，并计算主题对应的词汇的概率分布，通过训练迭代更新LDA模型中的参数以得到最优模型，并与K-means算法结合，将K-means结果与初始标签进行比对，判断分类结果的优劣，并根据对比实验讨论了LDA模型中主题数、以“字”或“词”为最小单位对于分类结果的影响，并对于数据结果进行了图形化表示与分析。

## 一、介绍

在互联网时代，大量的文本数据不断产生，如何从这些海量文本中挖掘出有价值的信息，是一个重要的研究方向。传统的文本挖掘方法主要是基于关键词的检索和分类，但是这些方法都忽略了文本中隐藏的主题结构。因此，LDA主题模型应运而生，LDA（Latent Dirichlet Allocation）主题模型是一种概率模型，用于对文本集合进行主题建模。它是由Blei等人于2003年提出的，是文本挖掘领域的一个重要研究方向。

LDA主题模型是一种概率模型，用于发现文本集合中的主题，并将每个文档表示为主题的概率分布。它是一种无监督学习算法，它的基本思想是假设文档由多个主题组成，而每个主题又由多个单词组成。通过这种方式，LDA模型可以将每个文档表示为主题概率分布，将每个主题表示为单词概率分布。通过对文档集合中的单词进行观察，LDA模型可以通过反向推理估计主题和单词的分布。具体来说，LDA算法使用迭代的方法从初始的随机状态开始优化模型，找到最佳的主题和单词分布。

LDA主题模型的应用非常广泛，如文本分类、主题分析、信息检索、社交媒体分析等领域。其中，LDA在文本分析中的应用最为广泛，如新闻分类、主题分析、情感分析等。它还可以用于推荐系统中，如商品推荐、音乐推荐等。

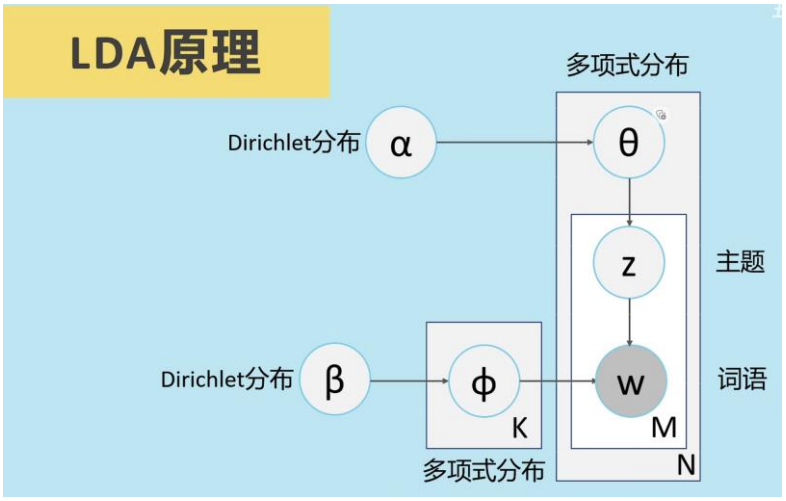
## 二、方法原理

### 1、LDA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，就是说，我们认为一篇文章的每个词都是通过“文章以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。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：对每一篇文档，从主题分布中抽取一个主题，从上述被抽到的主题所对应的单词分布中抽取一个单词，重复上述过程直至便利文档中的每一个单词。

LDA认为每篇文章是由多个主题混合而成的，而每个主题可以由多个词的概率表征，所以整个程序的输入是分词后的文章集（通常为一篇文章一行),主题数K，超参数α和β，输出是每篇文章的各个词被置顶的主题编号、每篇文章的主题概率分布、每个主题下的词概率分布、程序中词语的映射表、每个主题下概率从高到低的特征词。




由上图可知，LDA主题模型的生成原理大致如下[2]：

1、$\alpha$ 随机生成文档对应主题的狄利克雷多项式分布 $\theta$。

2、根据多项式分布 $\theta$ 随机生成一个主题z。

3、$\beta$ 随机生成主题对应词语的多项式分布 $\phi$。

4、综合主题z和主题对应词与的分布情况 $\phi$，共同生成词语w。

5、循环1-4步即可生成一个文档，其中包含M个词语。

6、 最终生成K个主题下的N篇文档。

在应用LDA模型时首先需要确定其中的参数，首先需要确定确定主题数(Topic Number)。主题数是LDA模型的一个重要参数，它指定了LDA模型中将要生成多少个主题。主题数的确定通常需要进行实验和评估，具体的步骤如下：

(1) 初步估计主题数：在确定主题数之前，可以首先对数据集进行一些探索性分析，比如通过可视化等方法来初步估计主题数的范围。

(2) 应用质量度量：通过主题的质量度量来评估不同主题数下的LDA模型效果，比如一些通用的评估指标有：困惑度（Perplexity）、主题一致性（Topic Coherence）、主题独立性（Topic Independence）等。

(3) 选择最优主题数：通过比较不同主题数下的质量度量指标，选择最优的主题数作为LDA模型的参数。

LDA模型中还包含一些超参数，比如α和β，它们分别控制着主题分布和单词分布的平滑程度。超参数的选择通常也需要经验和实验，常见的方法包括：

(1) 采用默认值：LDA模型库通常会提供一些默认值，可以先采用这些默认值进行建模。

(2) 交叉验证：通过交叉验证等方法来确定超参数的最优值，具体来说，可以将数据集分为训练集和测试集，利用训练集训练LDA模型，然后在测试集上通过一些评估指标来评估模型的性能，最后选择最优的超参数。

(3) 先验知识：对于某些特定的任务或数据集，可以利用先验知识来确定超参数的值。

### 2、K-means算法

K-means算法是一种常见的聚类算法，用于将数据点分成K个不同的组或簇。聚类是一种无监督学习方法，其目标是将数据点划分为具有相似特征的组或簇，使得组内差异尽可能小，组间差异尽可能大。

K-means算法的基本思想是：首先随机选取K个点作为簇中心，然后将每个数据点分配到距离最近的簇中心所在的簇中。接着，重新计算每个簇的中心，直到簇中心不再改变或达到最大迭代次数为止。

具体来说，K-means算法的步骤如下：

(1) 选择簇的个数K。

(2) 随机初始化K个簇中心。

(3) 对于每个数据点，计算其与每个簇中心的距离，将其分配到距离最近的簇中心所在的簇中。

(4) 对于每个簇，重新计算其中心点。

(5) 如果中心点不再改变或者达到最大迭代次数，则停止迭代；否则，返回第3步继续迭代。

K-means算法的优点是易于理解和实现，计算效率高。但是，K-means算法的结果可能受到初始簇中心的影响，并且对于非凸数据集的聚类效果可能不好。因此，常常需要多次运行算法，并从多个聚类结果中选择最优的结果。

## 三、实验结果与分析

本实验使用Python进行编程实现，主要的实验内容如下：

1、 数据预处理

以作家金庸的16本中文小说作为语料库，分别对每本小说进行了数据预处理，即删除了停词、英文大小写字母、广告等冗余内容，并在初步运行程序后根据实验结果更新了停词内容，降低非法字符或无意义字符对于算法本身的影响

```
with open(path, 'r', encoding='ANSI') as f:
    data = f.read()
    data = data.replace('本书来自 www.cr173.com 免费 txt 小说下载站\n '
                        '更多更新免费电子书请关注www.cr173.com', '')
    data = data.replace('本书来自www.cr173.com免费txt小说下载站\n'
                        '更多更新免费电子书请关注www.cr173.com', '')
    data = re.sub(u'[a-zA-Z0-9’!"#$%&\'()*+,-./:: ;<=>?@, 。?'
                  u'★、…【】《》？“”‘’！[\\]^_`{|}~「」『』（）]+', '', data)
    data = data.replace(' ', '')
    data = data.replace('\n', '')
    data = data.replace('\t', '')
    data = data.replace('\u3000', '')
    for item in stopwords:
        data = data.replace(item, '')
    f.close()
```

2、 数据分割处理

从上述16本金庸小说中均匀选取段落，选取方式为从在每本小说中以5%为单位等间距地取用全文中20个段落，每个段落500字，并为每个段落加上标签，标签即为对应段落所属的小说，共计320个段落，用以后续LDA主题模型训练。

3、 建立LDA模型与K-means模型

调用Python的库lda与sklearn其中的函数，建立LDA模型将语料库中提取的数据集以“词”为基本单元，根据输入的主题数来建立概率分布，并计算每一个主题数对应的词汇的概率分布，通过训练迭代更新LDA模型中的参数以得到最优模型，之后利用上述LDA模型的结果，适用K-means算法对语料库中的320个段落进行聚类，将K-means结果与初始标签进行比对，即可判断该模型的分类结果。（LDA模型参数中，主题数为13，迭代次数为3000次）

```
INFO:lda:n_documents: 320
INFO:lda:vocab_size: 45734
INFO:lda:n_words: 159626
INFO:lda:n_topics: 13
INFO:lda:n_iter: 3000
```

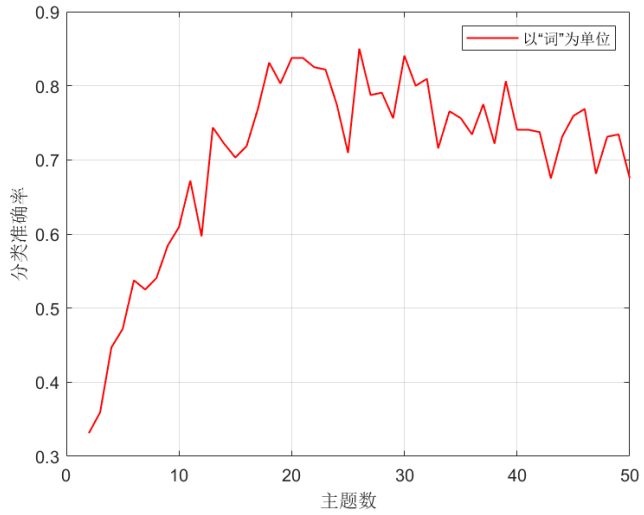
```
主题 0 : 听 甚 见 笑 话 想 兄弟 师父 武功 日
主题 1 : 袁承志 陈家洛 见 李沅芷 霍青桐 徐天宏 群雄 陆菲青 清兵 张召重
主题 2 : 范蠡 剑士 剑 勾践 青衣 长剑 少女 吴国 薛烛 竹棒
主题 3 : 爹爹 狄云 万圭 万震山 丁典 僧 朱 剑 水笙 紫
主题 4 : 著 卓天雄 周威信 爹爹 刀 袁冠南 於 孩子 瞎子 书生
主题 5 : 石破天 帮主 派 雪山 武功 白万剑 女子 老僧 绣 慕容复
主题 6 : 身子 胡斐 功夫 出 武功 左手 招 眼见 长剑 胸口
主题 7 : 杨 洪七公 郭靖 黄蓉 欧阳锋 蒙古 法王 周伯通 甚 黄药师
主题 8 : 韦宝 皇帝 做 乾隆 侍卫 康熙 海老公 会 闯王 心想
主题 9 : 曰 虱髯 见 做 杀 毛文龙 妻子 妾 袁崇焕 秦桧
主题 10 : 李文秀 著 苏鲁克 陈达海 曹云奇 白马 後 强盗 车库 於
主题 11 : 令狐 张忌 教主 盈盈 谢逊 张翠山 派 忌 弟子 华山派
主题 12 : 听 见 里 走 心想 瞧 姑娘 杀 突然 想
```

```
------------------------节选片段聚类结果-------------------------

白马啸西风 ：
14 14 14 14 14 14 14 14 14 14 14 14 14 5 14 14 14 14 14 14

碧血剑 ：
7 8 8 8 8 8 8 9 8 8 8 3 8 8 6 8 8 8 7 7

飞狐外传 ：
6 5 5 8 9 9 9 9 5 5 9 9 6 9 6 6 0 9 9 12
```

4、 研究主题数对于分类结果的影响

由3中的图像可知，《白马啸西风》其中的20个段落在聚类有19个段落分为了“14”号类别，只有1个段落分为了“5”号类别，故此处认为对于《白马啸西风》的分类准确率为95%。而最终LDA模型在某一主题数下对于16本小说的分类准确率为：16本小说的分类准确率的平均值。

在本次验证中，LDA模型的主题数从2变化至50，迭代次数设为3000，语料库中提取的内容“词”为基本单元，记录每一主题数最终对应的分类准确率，绘制出如下图像。




可以看出，在主题数位于[2, 20)区间时，分类准确率与主题数呈正相关，从0.33上升至0.83。其原因是，当主题数较少时，并不能根据不足的主题类别对于语料库中所有的内容进行分类判别，故最终分类准确率较低，而随着主题数的增加，对于语料库中的内容的分辨能力增加了，故分类准确率也提高了。

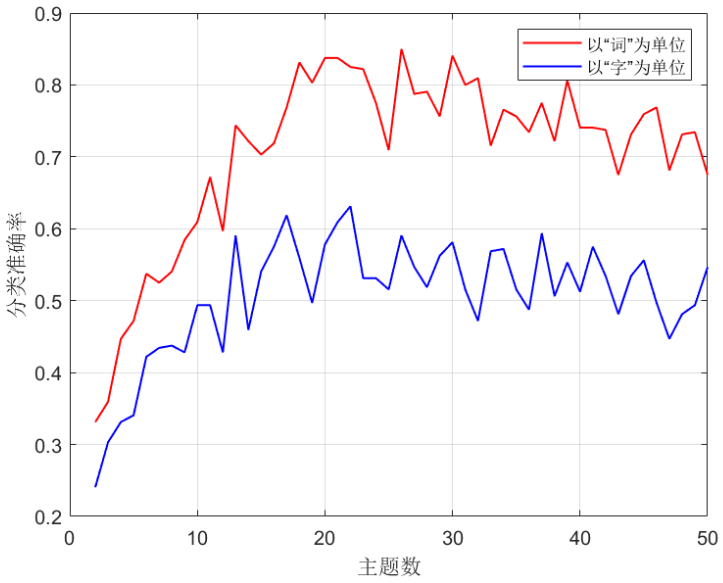
当主题数位于[20, 30]区间时，模型的整体分类效果较好，模型的分类准确率在[0.71,0.85]区间内波动，此区间内主题数的设定对于模型的分类结果来说影响较小。

当主题数位于(30, 50]区间时，模型的分类准确率呈现波动式下降，其主要原因是主题数过多，使得LDA模型对于语料库的分类过于细致，导致了过拟合，使得最终分类的效果呈现了下降的趋势，而随着主题数的进一步增加，过拟合的情况也更严重，分类准确率也更低。

5、 研究“字”和“词”对于分类结果的影响

相对于以“词”为基本单元的LDA模型进行对比实验：以“字”为基本单元进行模型训练，LDA模型的主题数从2变化至50，迭代次数设为3000，实验结果如下：

```
主题 0 : 师 子 知 事 三 相 武 十 年 二
主题 1 : 刀 萧 林 著 子 镖 夫 周 信 威
主题 2 : 袁 陈 承 志 家 青 见 洛 周 弟
主题 3 : 石 张 主 帮 教 天 谢 破 忌 爷
主题 4 : 马 十 回 书 少 城 行 女 字 生
主题 5 : 子 声 心 见 身 头 想 里 听 出
主题 6 : 杨 郭 黄 靖 蓉 婆 姊 师 洪 陆
主题 7 : 韦 宝 皇 公 国 官 王 帝 兵 曰
主题 8 : 胡 刀 斐 苗 宝 云 爹 家 狄 心
主题 9 : 剑 士 范 青 吴 王 国 蠡 名 越
主题 10 : 李 秀 文 著 苏 克 孩 计 汉 针
主题 11 : 令 狐 僧 师 段 仙 林 派 岳 山
主题 12 : 手 剑 出 声 招 身 力 掌 法 左
```




从上图可知，以“字”为基本单元时，尽管分类的结果更加细致，但是由于“字”相对于“词”所包含的信息量较少，且信息较为分散，不如“词”所包含的信息完整，故LDA模型在分类时可用的信息量变少，导致分类效果不佳，最终以“字”为基本单位的分类的准确率相较于以“词”为单位的下降了20%左右。

# 四、结论

本实验以作家金庸的16本中文小说作为语料库，分别对每本小说进行了数据预处理与片段选取，在实验中建立了LDA主题模型，对于语料库中的内容主题分类，并计算主题对应的词汇的概率分布，通过训练迭代更新LDA模型中的参数以得到最优模型，并与K-means算法结合，将K-means结果与初始标签进行比对，判断分类结果的优劣，并根据对比实验讨论了LDA模型中主题数、以“字”或“词”为最小单位对于分类结果的影响，并对于数据结果进行了图形化表示与分析。

对实验结果分析得知，当使用不同数量的主题个数进行分类时，我们会发现分类性能的表现会有所不同。如果我们使用较少的主题个数，那么分类结果可能会比较模糊，难以区分不同小说之间的差异。相反，如果我们使用较多的主题个数，那么分类结果可能会过于细致，难以将不同小说归为同一类别。因此，在选择主题个数时，我们需要考虑具体情况，选择合适的数量。

此外，如果我们以不同的基本单元（即“词”和“字”）进行分类，分类结果也会有所不同。以“字”为基本单元时，尽管分类的结果更加细致，但是由于“字”相对于“词”所包含的信息量较少，且信息较为分散，不如“词”所包含的信息完整，故LDA模型在分类时可用的信息量变少，导致分类效果不佳，最终以“字”为基本单位的分类的准确率较差于以“词”为单位的结果。

# 参考资料

[1] LDA模型介绍 https://blog.csdn.net/qfikh/article/details/103043630

[2] 主题模型分析-LDA (Latent Dirichlet Allocation)【python-sklearn】

[3] 一文详解LDA主题模型 https://zhuanlan.zhihu.com/p/31470216