# NLP大作业——LDA模型

**学院：** 自动化科学与电气工程学院 **姓名：**王明贤 **学号：**ZY2103526

## 一、LDA模型简介

在文本挖掘领域中大量的数据都是非结构化的，难以从信息中直接获取相关和期望的信息。主题模型（Topic Model）能够识别在文档里的主题，并且挖掘语料里隐藏信息，在主题聚合、特征选择等场景有广泛的用途。

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“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。

LDA采用了“词袋”的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是没有考虑词与词之间的顺序，这简化了问题的复杂性。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

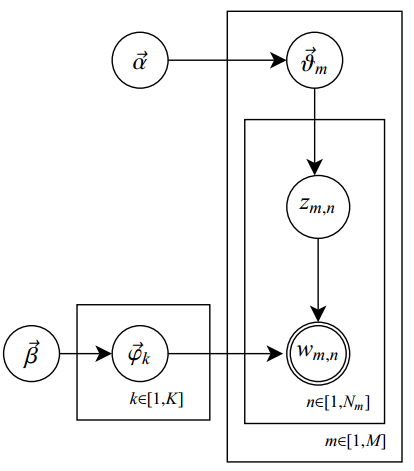


图 1 LDA模型

1.按照先验概率选择一篇文档

2.从狄利克雷分布中取样生成文档的主题分布

3.从主题的多项式分布中取样生成文档第j个词的主题

4.从狄利克雷分布β中取样生成主题对应的词语分布

5.从词语的多项式分布，中采样最终生成词语

## 二、问题描述与分析

### 1.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。截至日期：5月6日晚12点前

### 2.问题分析

整个问题包含三个部分：段落抽取、LDA模型训练、分类器训练。

段落抽取通过均匀抽取的办法可以实现，这里选取五本小说各抽取四十个段落。LDA模型训练采取初始化各参数后，吉布斯采样逼近所求分布的办法。

吉布斯采样（Gibbs Sampling）首先选取概率向量的一个维度，给定其他维度的变量值当前维度的值，不断收敛来输出待估计的参数。具体地

1.随机给每一篇文档的每一个词w，随机分配主题编号z

2.统计每个主题下出现字w的数量，以及每个文档n中出现主题中的词w的数量

3.每次排除当前词w的主题分布，根据其他所有词的主题分类，来估计当前词w分配到各个主题的概率，即计算（Gibbs updating rule）。得到当前词属于所有主题的概率分布后，重新为词采样一个新的主题。用同样的方法不断更新的下一个词的主题，直到每个文档下的主题分布和每个主题下的词分布收敛。

4.最后输出待估计参数，每个单词的主题也可以得到。

在大量的迭代后，主题分布和字分布都比较稳定也比较好了，LDA模型收敛。

最终根据训练得到段落主题分布的向量训练分类器，本文中采用SVM进行分类。Sklearn中OvR为每一个类别配备一个分类器，是目前最常用的一种多类分类策略。

## 三、算法设计

该问题是一个混合高斯模型求解问题，可以使用EM算法估计隐含量。



图 2 方案流程图

如图1所示，该算法分为以下3步：

（1）段落生成：文本读取、分词预处理、均匀提取段落

（2）LDA模型训练：参数随机初始化、吉布斯采样更新迭代参数、计算困惑度

（3）分类器训练：划分训练集测试集、训练SVM分类器、评估模型

## 四、运行结果

### 1.运行结果

本实验选取倚天屠龙记0、天龙八部1、射雕英雄传2、神雕侠侣3、笑傲江湖4进行实验，每本小说各均匀抽取40段，每段500词。

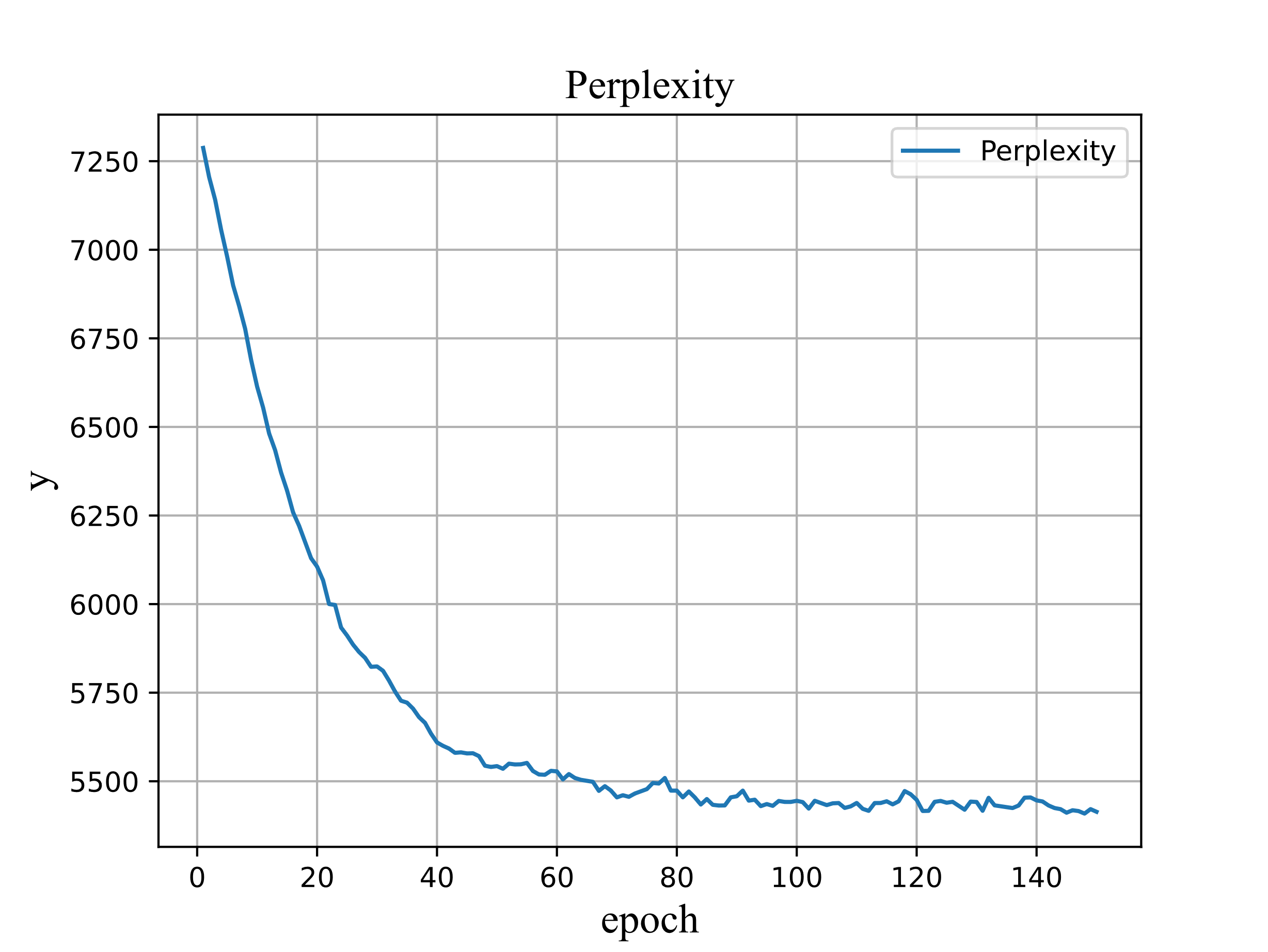


图 3 困惑度下降趋势

可以看出，在迭代100次后训练基本收敛。

表 1 各主题前十高频词

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Topic1 | 弟子 | 令狐冲 | 长剑 | 剑法 | 甚么 | 教主 | 众人 | 岳不群 | 兵刃 | 跟着 |
| Topic2 | 张无忌 | 说道 | 张翠山 | 今日 | 少林 | 谢逊 | 大师 | 少林寺 | 咱们 | 冷笑 |
| Topic3 | 说到 | 不是 | 只是 | 师傅 | 心中 | 知道 | 出来 | 一个 | 如此 | 怎么 |
| Topic4 | 汉子 | 萧峰 | 兄弟 | 一个 | 向问天 | 少年 | 之中 | 向来 | 无法 | 大哥 |
| Topic5 | 杨过 | 小龙女 | 武功 | 此时 | 却是 | 如何 | 少女 | 两个 | 弟子 | 只是 |
| Topic6 | 郭靖 | 黄蓉 | 师父 | 黄药师 | 洪七公 | 周伯通 | 甚么 | 欧阳锋 | 欧阳克 | 梅超风 |
| Topic7 | 令狐冲 | 一个 | 你们 | 我们 | 左冷禅 | 自己 | 二人 | 正是 | 说道 | 当下 |
| Topic8 | 只见 | 一个 | 声响 | 段正淳 | 不见 | 下来 | 老者 | 几个 | 登时 | 段誉 |
| Topic9 | 自己 | 一声 | 身子 | 突然 | 右手 | 之下 | 内力 | 胸口 | 对方 | 敌人 |
| Topic10 | 蒙古 | 英雄 | 郭靖 | 王语嫣 | 一条 | 突然 | 慕容复 | 襄阳 | 星宿 | 今日 |

从最终训练得到的主题高频词中可以看出，对部分主题学习的效果较好，但是也出现了很多常用词，不能体现具体主题内容，如果对这类没有特点的词进行限制将得到更好的结果。

表 2 SVM分类器结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 小说 | 0 | 1 | 2 | 3 | 4 |
| 正确率 | 1 | 0.57 | 0.83 | 0．81 | 0.8 |

最终分类器对倚天屠龙记0、射雕英雄传2、神雕侠侣3、笑傲江湖4有较好的分辨，对天龙八部1则不能区分。

### 2. 结果分析

首先结果体现了LDA模型的有效性。对于部分较差的分类结果是由于语料库不全导致的。同时如果能够增加停用词将进一步提升分类效果。

## 五、总结体会

该问题是一个文本模型构建的分类问题，通过LDA主题模型学习文本内部关联，并通过SVM分类器进行分类，充分说明了统计方法在文本分析等自然语言处理中的作用。