Latent Dirichlet Allocation（LDA）主题文本建模算法

申海（ZY2203325）

## 一、摘要

从链接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。

本文研究了LDA主题模型在小说分类中的应用。我们从一个包含15个金庸小说的语料库中，随机抽取了200个段落，并将每个段落表示为主题分布。我们探究了在不同数量的主题个数下分类性能的变化，以及以"词"和以"字"为基本单元下分类结果的差异。实验结果表明，在适当的主题个数下，LDA主题模型可以有效地对小说段落进行分类，并且以"词"为基本单元的分类结果优于以"字"为基本单元的分类结果。

## 二、方法

我们采用LDA主题模型进行文本分析，得到段落的主题分布。然后使用段落的主题分布作为特征，使用朴素贝叶斯分类器进行分类。

#### 2.1LDA主题模型

**LDA（Latent Dirichlet Allocation）主题模型是一种基于概率图模型的文本分析方法，它可以用于发现文本中隐藏的主题。在LDA模型中，我们假设每个文档都由多个主题组成，每个主题又由多个单词组成。通过对文本中的单词进行统计分析，我们可以推断出每个文档中包含哪些主题以及每个主题中包含哪些单词。**

**假设我们有 个文档，每个文档由 个单词组成。我们用 表示文档 中的第 个单词，用 表示文档 中的第 个单词的主题。我们假设每个主题 都由一个多项分布 表示，每个文档 都由一个多项分布 表示。我们还需要定义一个先验分布 和一个先验分布 ，用来控制主题和单词的分布。**

**LDA模型的生成过程如下：**

**对于每个主题 ，从多项分布 中采样一个单词分布 。**

**对于每个文档 ，从多项分布 中采样一个主题分布 。**

**对于文档 中的每个单词 ：**

**从文档 的主题分布 中采样一个主题 。**

**从主题 的单词分布 中采样一个单词 。**

**LDA模型的目标是，对于给定的文本集合，推断出每个文档的主题分布 和每个主题的单词分布 。这个过程可以使用Gibbs采样或变分推断等方法来实现。**

**总体来说，LDA主题模型是一种非常有用的文本分析方法，可以帮助我们发现文本中隐藏的主题，从而更好地理解和分析文本内容。**

#### ****2.2朴素贝叶斯分类器****

**对段落进行分类，采用朴素贝叶斯分类器**是一种基于贝叶斯定理和特征独立假设的分类器，常用于文本分类等任务。在使用朴素贝叶斯分类器进行分类时，我们需要先将样本转换成数值特征，然后将这些特征输入到分类器中进行分类。

具体来说，设样本 包含 个特征 ，且每个特征都是离散的。我们假设每个特征在给定类别 的条件下是独立的，即：

根据贝叶斯定理，我们可以得到：

其中， 表示在给定样本 $x$ 的条件下，样本属于类别 的概率； 表示在给定类别 的条件下，样本 出现的概率； 表示类别 出现的概率； 表示样本 出现的概率。

在朴素贝叶斯分类器中，我们通常将 视为先验概率，将 视为似然函数。对于给定的样本 ，我们可以计算它属于每个类别的后验概率 ，然后将样本分配给具有最高后验概率的类别。

在文本分类等任务中，我们通常使用多项式朴素贝叶斯分类器。该分类器假设每个特征都是离散的，且特征值表示单词出现的次数。在使用多项式朴素贝叶斯分类器进行分类时，我们需要先将文本转换成数值特征，然后将这些特征输入到分类器中进行分类。

## ****三、实验****

从金庸的15本小说集语料库中均匀抽取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。进行LDA建模并进行分类，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。

首先读取文本，然后进行预处理，对小说集进行划分段落和贴标签，随后进行LDA训练，最后使用朴素贝叶斯分类器进行分类

1. 读取小说集（字体大小25）

使用os库读取指定文件下的所有txt文件，然后去除停词并进行jieba。Lcut分词等，读取为字典，键为小说名（文件名），值就是去除停词后的分词。15本小说的主要词如下图。

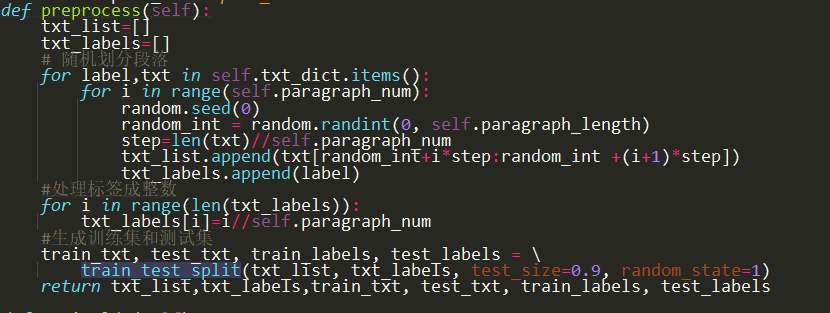




1. 预处理

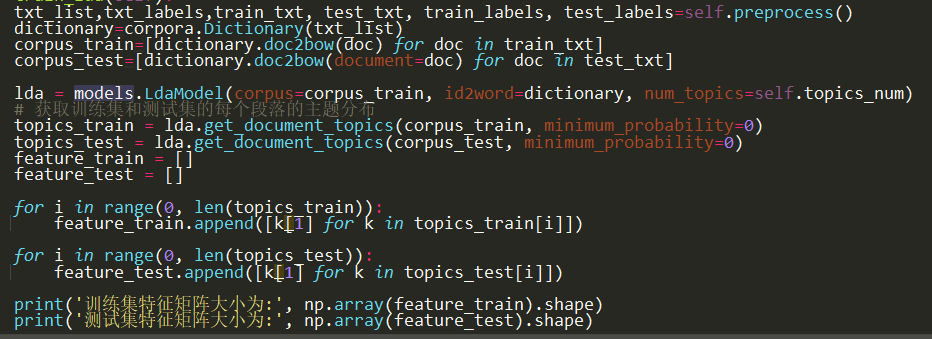
预处理主要是生成训练和测试数据，首先进行段落划分，然后以小说名为依据生成段落的标签。最后使用sklearn.model\_selection.

train\_test\_split生成训练数据和测试数据。



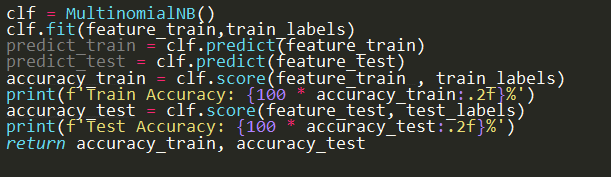
1. LDA训练

直接使用gensim.models.LdaMode进行训练



1. 朴素贝叶斯分类

调用sklearn.naive\_bayes.MultinomialNB进行分类，具体代码如下。



#### 3.1不同主题个数下分类性能的变化

在一个段落500词下，不同主题个数下的分类准确率如下图所示。

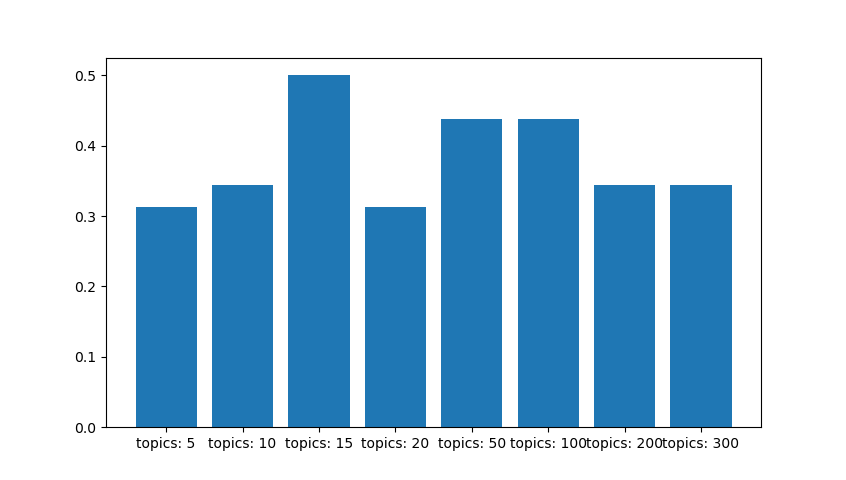


图 1训练集不同主题个数准确率图

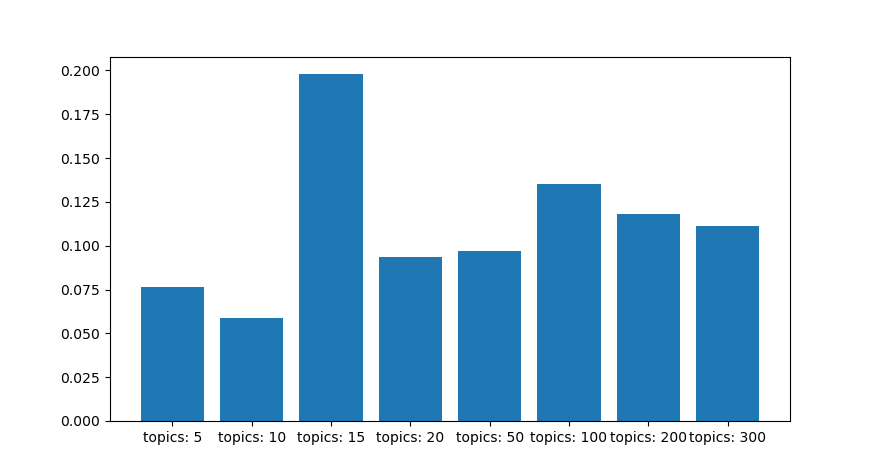


图 2测试集不同主题个数准确率图

从结果来看，不管是测试集还是训练集，主题个数并不是越多越好，大致在15个时，分类效果最佳。但是可以看到，不管是测试集还是训练集，分类效果都比较差。

#### 3.2以"词"和以"字"为基本单元下分类结果分析。

以每篇段落数为50，每个段落1000字，一共50个主题，n-gram模型下训练集和测试集的准确率如下。

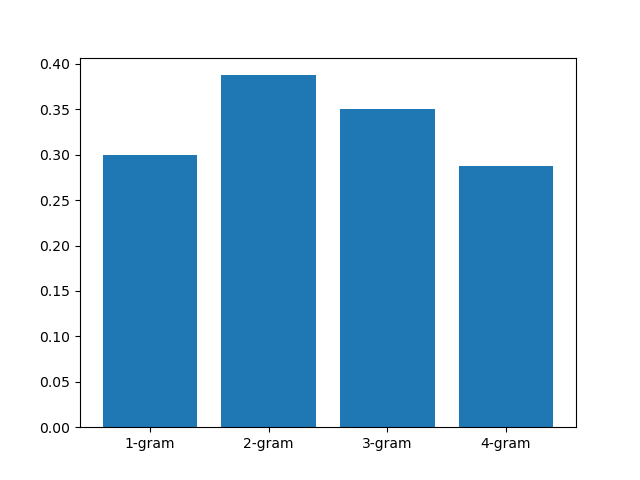


图 3不同n-gram训练集准确率

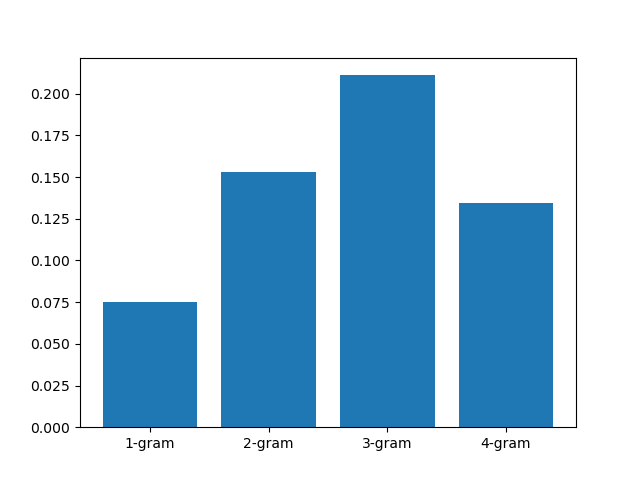


图 4不同n-gram测试集准确率

从结果图看来，不管是测试集还是训练集，词的段落分类准确率都要高于字的段落分类准确率。