**Report of Deep learning for Natural language Processing**

唐宗润 ZY2303211

**Abstract**

从语料库中均匀抽取1000个段落作为数据集（每个段落可以有 K 个 token, K取20，100，500, 1000, 3000）。利用LDA模型在给定的语料库上进行文本建模，主题数量为 T，并把每个段落表示为主题分布后进行分类（分类器自由选择），分类结果使用 10 次交叉验证（i.e. 900 做训练，剩余100 做测试循环十次）。讨论在设定不同的主题个数T的情况下，分类性能的变化变化；以"词"和以"字"为基本单元下分类结果的差异；不同的取值的K的文本，主题模型性能上的差异。

**Introduction**

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。

图示

描述已自动生成

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。LDA是一种非监督机器学习技术,可以用来识别大规模文档集(document collection）或语料库（corpus）中潜藏的主题信息。

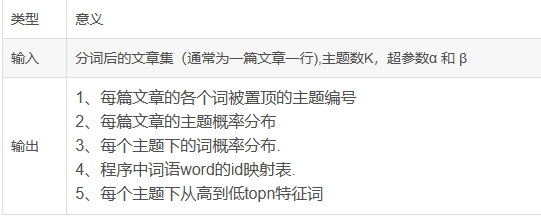
对于语料库中的每篇文档，LDA定义了如下的生成过程：

1对每一篇文档，从主题分布中抽取一个主题。

2从上述被抽到的主题所对应的单词分布中抽取一个单词。

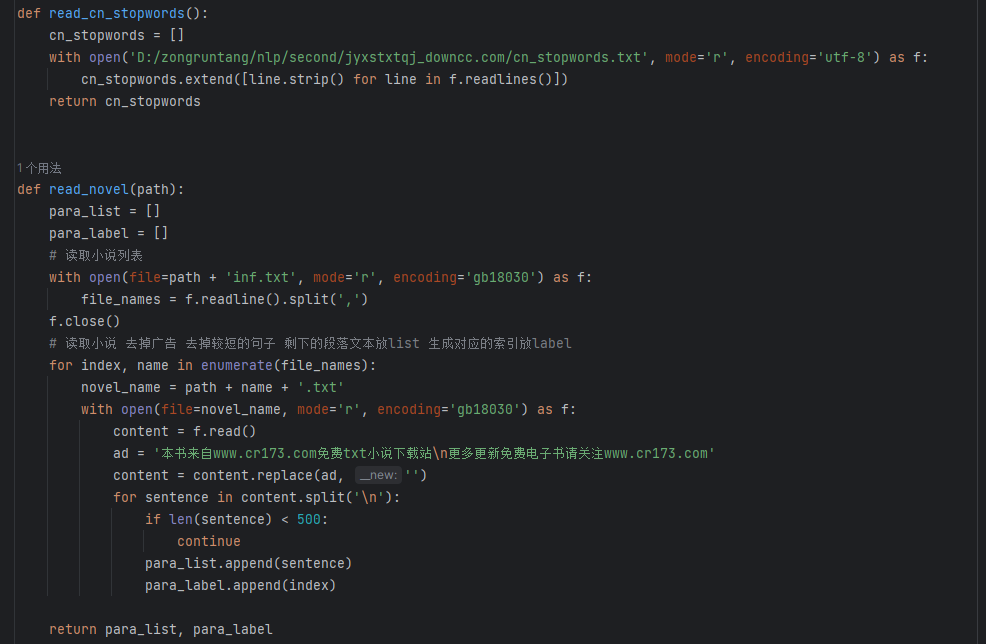
3重复上述过程直至便利文档中的每一个单词。

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和输出如下所示：



**Methodology**

首先对数据进行预处理，将文件中停词去掉，广告去除。



其次，对预处理后的段随机抽取1000个，书名为标签

文本

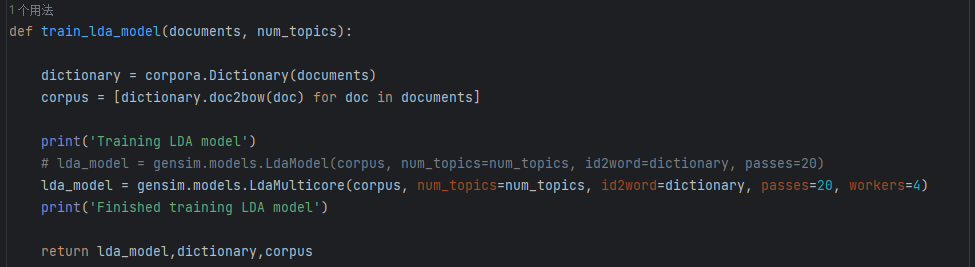
描述已自动生成

对选中段落进行分词，利用jieba分词，再此基础上加个循环实现以字为基本单元的数组

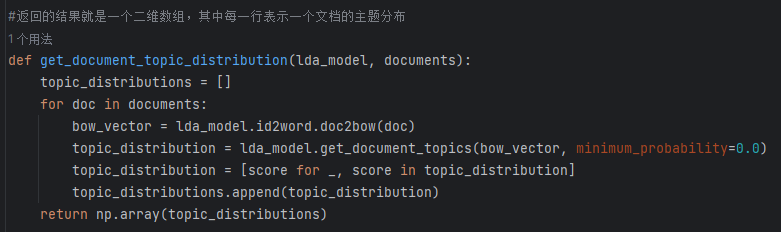
文本

描述已自动生成

构建字典和向量集后，进行LDA模型训练。



获得主题分布



文本

描述已自动生成划分测试集和数据集，并进行交叉验证

Conclusion

以字为单位和以词为单位的LDA模型测试集准确度如下。



1. 在设定不同的主题个数T的情况下，分类性能是否有变化？

当K值固定时，T在【20，60】内，准确度随着T值增大而增大，而当T在【60，80】时，准确度反而下降，说明Topics数量过多，可能导致分类过于细化而影响准确度

1. 以"词"和以"字"为基本单元下分类结果有什么差异？

当T和K过小时，两者差距不大，当T和K取适当值时（如T=40，K=3000），以字为基本单位的准确度低于以词为基本单位的准确度。

1. 不同的取值的K的短文本和长文本，主题模型性能上是否有差异？

当T值固定时（如取T=40）,以字为基本单位或者以词为基本单位 ，k值越大，取得的token数越多，准确度越高

最后，为了评价LDA模型的性能，引入困惑度评价指标，计算如图 