实验报告

一、实验要求

1.预处理文本数据集，并且得到每个文本的 VSM 表示。

2.实现 KNN 分类器，测试其在 20Newsgroups 上的效果。

3.20%作为测试数据集，保证测试数据中各个类的文档均匀分布。

二、实验过程

1.要对每个类别的每个文件的文本进行预处理，使之转化为更容易处理的格式。将准备用于输入的文档进行去噪、分词、编码格式转换、以及去除停用词等等。然后通过文件读写操作将处理后的数据进行保存，处理后每个单词占一行。

2.按照要求比例划分训练集和测试集，在遍历文件时，先统计文件总数，然后将占该类别总数的 80%文件拷贝到训练集指定的路径，剩下的拷贝到测试集路径下。

3.VSM：空间向量模型

首先为数据建一个字典，遍历每一个文档得到该文档的每个词的词频字典从而得到整个数据集的词频字典，利用词频减小字典长度，只保留词频大于某个值的词，并且计算 idf 值。

然后每个文档建立 tf_idf 字典，每个文档建立一个词频字典。遍历文档中的所有单词，首先判断该单词是否出现在 idf 字典中，如果有，则统计该词的词频，并计算其 tf_idf 值，最终得到一个字典。

4.KNN 算法的实现步骤：首先将训练集和测试集的每个文档都表示成向量，计算测试集向量与训练向量的相似度并存放在列表中，取列表中相似度最大的 K 组，统计其中类型出现次数，选取最多的作为该测试向量的分类结果。分类结果与测试向量类名进行比较，统计分类正确和错误次数，算出正确率。

三、实验结果

K=10,实验的准确率最高

总测试次数：3654

预测成功次数：2967

预测准确率：0.811986863711002

四、实验心得

在实验的过程中我了解到了实验的整个步骤，VSM 和 KNN 的实现，都要一步步地搞清楚其原理才能进行下去，所以我对其有了更多的了解。

由于实验数据过大，我只选择了其中的一部分进行验证，并且 KNN 分类有待优化，实验准确率较差。