向量空间模型与 K 紧邻分类器实验报告

1. 实验任务

预处理文本数据集，并且得到每个文本的 VSM 表示;

实现 KNN 分类器，测试其在 20Newsgroups 上的效果。

2. 实验程序框架

实验程序大致框架示意见图 1.1：

图 1.1 程序结构示意

3. 实验遇到的主要问题

①一开始使用 python 的 sklearn 库内置函数求 tf-idf，得出的向量空间矩阵非常大，计算较慢，因此自己写了一个计算 tf-idf 的方法，为每一类文档生成一个词典，即所有类文档词典的并集为整个语料库的词典。在每一个小词典下可生成每类文档的向量空间矩阵。

②在使用 20 个小词典的前提下，计算相似度时需要找到测试文档中单词在训练集中的位置，比较耗费时间，故使用词典和向量空间矩阵直接将每个文档转换成 key 为单词，value 为 if-idf 值的形式，以此计算文档与文档之间的相似度。

③使用训练集训练结果对测试集数据进行分类准确度低，一开始去低频词时低于 8 的词不加入词典，后经实验发现，降低低频词下限可提高准确率，故使用 2 作为低频词下限，这增大了词典与计算的时间代价。且分析实验结果，

comp.sys.ibm.pc.hardware 与 comp.sys.mac.hardware 常分类错误，comp.graphics 与 comp.windows.x 常分类错误，我认为有可能是两类文档之间共同词较多可能造成分类错误，故每个小词典与另外 19 个词典的并集求差集（以后称此为小词典之间去共同词）以去除不同类之间过多重复词对分类造成的影响，但这种方法对准确率的提高并几乎没有帮助，但可以缩小词典，故本实验中采取了这种方式。

4. 关于实验参数

由于程序运行时间较慢，在有限时间内难以多次实验不同参数的值来寻找一组相对较优的参数，故选取一些认为可能使得分类效果好的参数进行实验。如 KNN 中的 K 分别取 5，10，15，20，30，以及生成词典时去低频词的下限分别为 2，5，8，10，15，分析比较发现 K 值在一个合理范围内的大小并不太影响分类结果（10-30），低频词的下限取的较低时分类准确率较高（缺点为词典太大），也曾尝试按照比例去高频词，但没有找到一个合适的上限。

5. 实验结果

在多次调整参数的实验中，分类最好的结果是：总数为 7532 的测试文档，4645 个分类正确，2887 个分类错误。此时，查准率为 61.67%

观察分类结果发现，有些类文档分类效果较差，如 alt.atheism，sci.mec，sci.electronics 等，有些类文档分类效果很好，如 rec.motorcycles，rec.sport.hockey，sci.space 等。

小词典之间去共同词与否的区别:

当其他参数相同的前提下，不去共同词时，词典较大，此时 alt.atheism 类文档分类效果较好，comp.graphics 与 comp.os.ms-windows.misc 类文档分类效果不好，推测 comp.graphics 与 comp.os.ms-windows.misc 类文档的词典之间的共同词影响了文档分类效果，综合运行时间与整体正确率考虑，去掉小词典之间的共同词进行实验较好。