实验报告

Homework1 VSM+KNN 将文档分为训练集和测试集，比例为=8:2 1、VSM 处理 20news-18828 数据集，生成文本对应的 VSM 表示。 （1）根据文件夹顺序读取数据集。 （2）依次通过去特殊字符、统一小写字母、分词、词干提取、去停用词等步骤进行数据预处理。 （3）选取训练集中词频大于 9 且小于 10000 的词创建词典。 （4）计算 TF-IDF 值以及生成文档的向量表示。 2、KNN 计算每个测试集文档与训练集之间的相似度，利用 knn 算法选取不同的 K 值进行分类，最终选取 K 值为 40，准确率为 74%。