实验报告一

本次实验是处理文本数据集，得出 VSM 空间向量，knn 分类器计算测试集相似度

一、Shujuchili.py 文件是对源数据 20news-group 进行预处理，提取干净的数据集，用读写操作将处理后的数据储存下来。遍历所有预处理文件，统计文件数目，将百分之八十数据作为训练集，将百分之二十作为测试集进行分类。

二、vsm.py 文件是将生成的文档表示为向量，建立词典，统计词项出现的频率，去除掉频率太小的词项。最后将所有文档表示成向量列表。计算出所有的 tf-idf 值。

三、knn.py 文件是对测试集的每一个向量，计算出它与训练向量的相似度，将训练向量的类型和相似度作为二元组存储在列表中，取出列表中相似度最大的元组，统计元组中类型出现次数，选取出现次数最多的便是分类出来的结果。