实验一：KNN+VSM

学号：201834878 姓名：王润琦

实验任务：

1. 对原始文档进行处理，得到文档的 VSM，并划分训练集和测试集。
2. 使用 KNN 对得到的 VSM 中的测试集进行分类，统计正确率。

数据集：20Newsgroups 链接：http://qwone.com/~jason/20Newsgroups/

实验步骤：

1. VSM

a) 将所有文档划分为单词，并对每个单词进行预处理，如去除非单词、大写变小写、动词变原型，去除 StopWords。

b) 对处理好的单词去重、统计词频，构建词典。

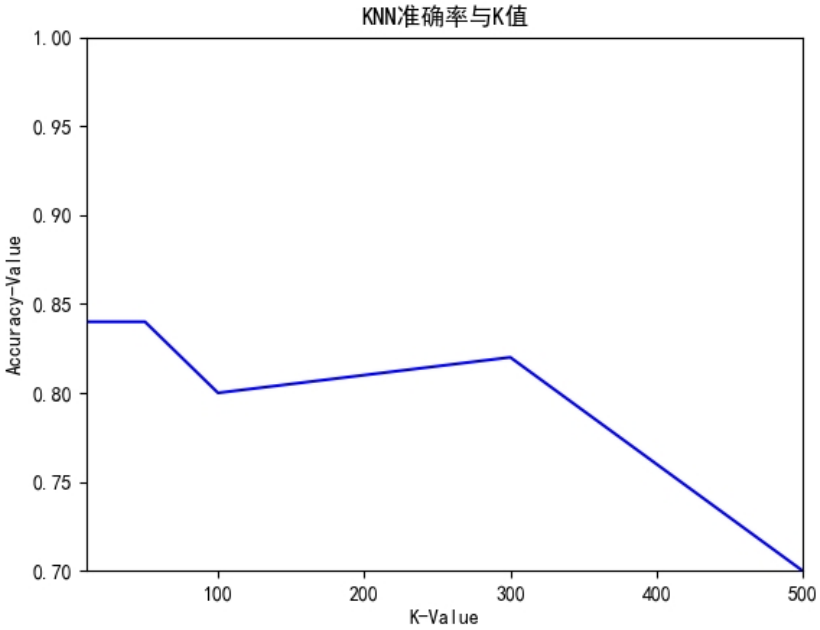
c) 对每个文本中的单词计算 tf-idf，通过建立好的字典生成 VSM。

2. KNN

a) 在已经获取到的 VSM 中按 4:1 的比例划分训练集和测试集。

b) 对测试集中的向量计算其与训练集向量的余弦相似度，选取前 K 个相似度高的向量，找其中相同 label 最多的那一项。统计分类正确的数目，得到正确率

实验结果：



实验总结：

a) 在实验中注意 K 值的选取，从图中我们可以看出 K 的值对准确率的影响是蛮大的。
b) 在数据处理中，要注意代码的优化，未优化好的代码对文档处理速度差异较大，例如在计算 idf-tf 时代码对速度的影响是很大的，优化之前基本十几秒才能算出一个文档，总时长预计 40 个小时以上，而优化后总时长仅仅 20 分钟。
c) 总的来看，KNN 算法在处理文档分类上还是蛮有效的，能够得到不错的准确率。