**K-NN文档分类实验报告**

余 平

**一、重点知识：**

1、**向量空间模型VSM**（Vector Space Model，简称VSM）：表示通过向量的方式来表征文本。 将一个文档抽象为一系列关键词的向量，该向量由n个关键词组成，每个词都有一个权重（Term Weight），不同的词根据自己在文档中的权重来影响文档相关性的重要程度。

2、**TF-IDF**：表示TF（词频）和IDF（逆文档频率）的乘积。特征抽取完后，因为每个词语对实体的贡献度不同，所以需要对这些词语赋予不同的权重。一个文档的TF-IDF与一个词在该文档中的出现次数成正比，与该词在整个语言环境（训练集）中的出现次数成反比。TF-IDF计算权重越大表示该词条对这个文本的重要性越大。

3、**余弦相似度：**两篇文章间的相似度通过两个向量的余弦夹角cos来描述。Cosine值越大，文章相似性越高。

**二、实验步骤：**

1、读出数据集文档，按照训练集与测试集4：1的比例进行分类，并为每篇文档打上分类标签。

2、对数据集进行预处理：去特殊符号、分词、转换成小写字母、去掉非英语单词、词形还原、去停用词。

分词使用TextBlob()分词工具；

词形还原只是将名词还原成单数形式，将动词转换成原型，其余词性的词未作变换。

3、计算的词频：使用collections.Counter()类对每篇文档进行词频统计。

4、使用训练集构建字典：根据统计词频的情况去除一些低频词和高频词，以减小文档向量空间的维度。

5、用训练集的词频统计计算IDF：

为避免，在公式分母加1，即：

6、分别计算训练集和测试集文档的TF：

7、计算训练集和测试集文档的TF\*IDF值：

𝑤(𝑡,𝑑)=𝑇𝐹(𝑡,𝑑)×𝐼𝐷𝐹(𝑡)

8、计算测试集文档与训练集文档向量的cosine值，用两个文档向量的cosine值表示其similarity。

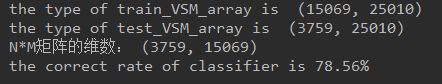
9、设置K值，找到与测试文档相似文档最多的类，判断分类是否正确。

**三、实验结果：**

实验中所有的数学运算使用矩阵进行计算。

将数据集的80作为训练集并构建词典，20%作为测试集测试模型，在该数据集上，K取值在区间[5,35]上滑动时，对结果的影响不大；经过词频统计，训练集低频词数量较多。

当K=35，字典词频为2-5000时，结果如下：



图中第一行为训练集生成的向量空间的维数；

第二行为测试集生成的向量空间的维数；

第三行为测试集与训练集的cosine值的维数；

第四行最终分类的正确率。

四、实验总结

分类结果很大程度上取决于数据集的预处理和词典的构建，K值在一定范围内对结果的影响相对较小。在预处理阶段，判断一个字符串是否为英语单词、去停用词的效果取决于字典的大小，在实验中也发现了词典不够用的情况；词形还原时效果不是很好。预处理函数的选取很大程度上影响了实验结果，完善数据集的预处理能够进一步提高正确率。

使用VSM进行文本分类，运算量很大，在本次实验中一个向量的维度达到25010。