VSM + KNN 实验报告

一、项目描述

此次实验主要是实现 VSM 和 KNN，过程简介如下：首先预处理文本数据集并获取每个文本的 VSM 表示；其次实施 KNN 分类器并测试其对 20Newsgroups 的影响。

二、过程实现

分为三个文件实现具体的过程：

VSM.py:

划分数据集，训练集为 80%，数据集为 20%；对 80%的数据集执行 5 次交叉验证，并将其分为 5 份，然后按顺序执行；对文档内容进行分词，去标点，小写化，去停用词，词干化等处理；遍历文档统计词频，删除小于 4 的低频词，并获取字典；根据字典清理文件并删除字典中未包含的单词。

TFIDFcompute.py:

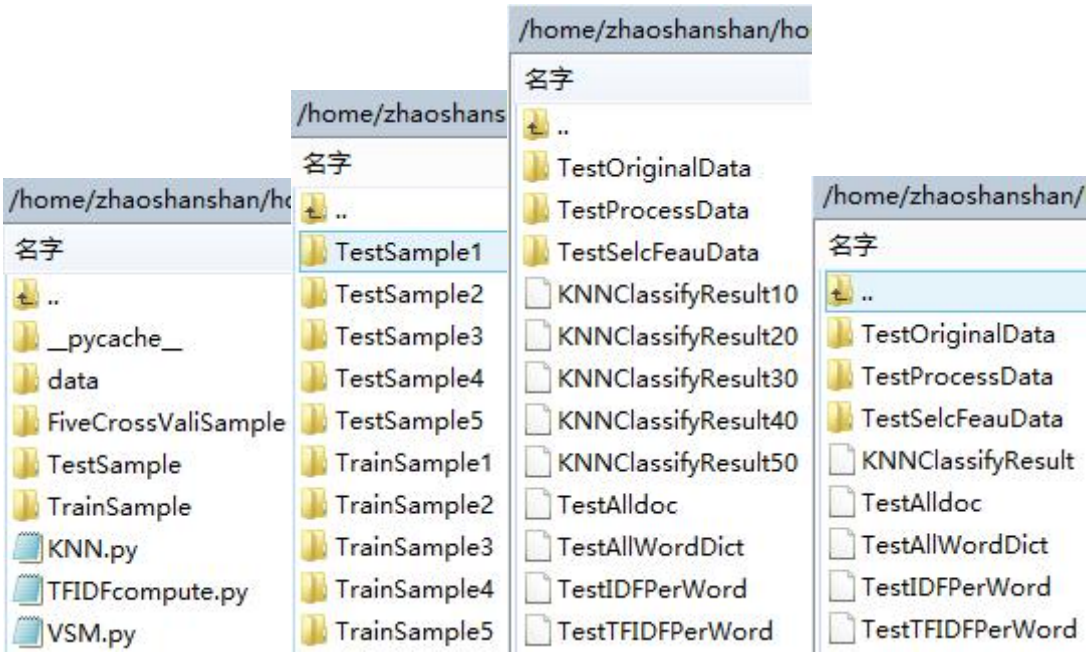
计算单词的 TF-IDF 值；其中为了简便计算，先计算的 IDF，之后根据求得的 TF 一起计算 TF*IDF。

KNN.py:

加载训练字典的 TFIDF 并处理测试数据的 TFIDF；根据 5 折交叉验证的结果，选择最好的 K 值。使用余弦相似度计算测试文档与所有训练样本之间的向量距离；找到距离最大的类并计算测试数据集的正确率。

三、运行结果

5 折交叉验证中，依次取 K=10,20,30,40,50，分别对训练集中的每个子集进行处理，得到最小错误率为 0.160227，此时 K=50（因运行速度慢，仅尝试了这些 K 值），再以 K=50 对训练集进行处理，得到最后的错误率为 0.150848，即准确率为 85%。




```
errorCount : 569 , Count : 3772 , errorRate : 0.150848
```

四、问题分析及解决

1.去除停用词时要下载 nltk data。

2.读取文件时使用路径出错，具体如下：os.mkdir 与 os.makedirs 的区别，os.makedirs 会递归的建立输入的路径，即使是上层的路径不存在，它也会建立这个路径，而 os.mkdir 父级路径不存在，那么就会报错。写的时候不注意会混淆出错。

3. 在初始划分训练集和测试集合时，如果每次执行不清空已存在的文件夹，会多写入文件，总数目不符合实际。所以每次运行都要从零开始。

4. 文件分类数目错误导致词典大小和单词的 TF-IDF 计算错误。

5. 划分数据集时尝试使用 shutil.copy()函数，而不是读取每行文件写入新文件。

6. 计算 IDF 与 TF 的先后问题。先计算 IDF，这样可以在计算 TF 后一起计算 TF*IDF。

7. 对处理好的数据计算 IDF，分别执行训练数据和测试数据，写入文档后，继续执行 TF，发现除 0 错误，且在执行 KNN 后发现训练单词的 TF-IDF 为空。原因是在使用 if 语句判别写入训练数据文档还是测试数据文档时，先打开了文件，再进行选择执行，这样会出现句柄使用错误。所以改为打开路径的方式。

8. m = len(lineSplit) 产生越界错误，原因是 split(' ')按空格分割后最后一位是空串，应当-1 防止产生越界。

9. 统计前 K 个值时出现问题。原因是应当统计的是前 K 个中具有相同类的距离之和，再选取最大的值，得到最终的分类。

10. 在计算余弦相似度时出错。解决：应当使用 float()将字符型数据转换成数值型数据，同时要将列表转矩阵，便于下面向量相乘运算和使用 Numpy 模块的范式函数计算。

11. 进行 5 折交叉验证时，按顺序将训练集分为 5 份，执行 KNN 发现总文件数据不符合实际。原因是在划分数据时采用含 y 的语句应当放在一层 for 循环里面。

12. 进行 5 折交叉验证时，修改路径出现很多问题，同时主函数中的路径中含有 str(i)报错，其中 i 是第 i 份训练集与测试集。先 m=str(i)，再将 m 放入路径中，同时清除其他函数中存在的 i，防止影响其他运算。