Data Mining 实验报告

Homework 1：VSM and KNN

学号：201844906　　姓名：宋春娇

一、 实验要求

1、对于 20 个新闻组数据集，进行必要的预处理之后，建立每个文本的空间向量模型。

2、按照 8:2 的比例划分训练集和测试集（要求测试数据均匀覆盖各个类别），应用 KNN 算法对测试数据进行分类，统计分类准确率。

二、 实验步骤

1、 实验数据预处理

需要对每个类别里的每个文件的文本进行预处理，使之转换为更容易处理的格式。需要进行分词操作，将句子划分为单词组成的列表，实验中使用 python 的 re 模块，来进行分词；并且需要删除无实际意义的停用词、字母转换为小写。然后通过文件读写操作将处理后的数据进行保存。处理后每个单词占一行，格式如图 1 所示。

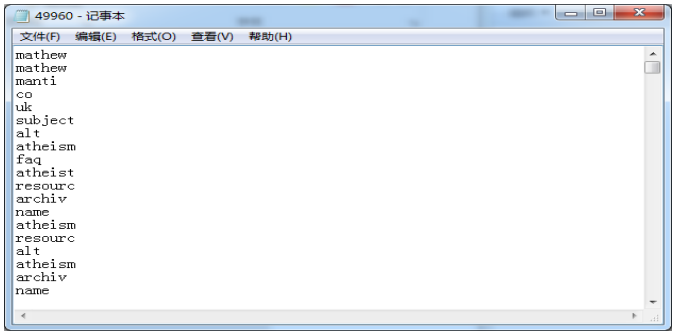

图 1 预处理后的文档格式

2、 数据集的划分

要求按照 8:2 的比例划分训练集和测试集且测试数据均匀覆盖

各个类别，实验中在遍历各个类别的文件时，首先计算该类别内的文件总数，遍历各文件时进行计数，使用 python 中的 shutil 模块进行文件的复制，将占该类别总数的 80%文件拷贝到训练集的指定路径下，20%拷贝到测试集路径下。

3、 VSM

空间向量模型基本思想是将一个文档看做向量空间中的一个点。文档中出现的词的重要程度是不同的，用词权重来度量。实验中用 TF_IDF 算法计来算词权重。

首先需要为数据集建立词典（用字典存储，key 为 word，value 为 idf 值）：遍历每一个文档时，使用 Counter 函数得到该文档中每个词的词频字典，从而得到整个数据集的词频字典以及 df 字典，利用词频减小字典长度，只保留词频大于某个值的词，并计算这些词的 idf 值。

然后为每一个文档建立 tf_idf 字典：需要注意的是词频（tf）是针对一个文档而言的，所以为每个文档建立一个词频字典。遍历某个文档中所有单词，首先判断该单词是否出现在 idf 字典中（如果不出现，则没有计算其 tf 值的必要），若出现，则统计该词的词频，并计算该词的 tf_idf 值，最终得到一个文档的 tf_idf 字典（{ ‘word1’ : tf_idf1， ‘word2’ : tf_idf2，……’ }）

Tf_idf 算法取一个文档中 tf_idf 值较大的前 50 个词作为该文档的关键词。由于 KNN 算法需要使用每个文档的类别属性，所以最终创建的向量形式为[‘label’ ，{ ‘word1’ : tf_idf1， ‘word2’ :

tf_idf2，……}]

4、 KNN

KNN 算法的实现步骤：首先将训练集和测试集的每个文档都表示成向量，对于测试集的每一个向量，计算其与每一个训练向量的相似度，将训练向量的类型和相似度作为二元组存储在列表中，取列表中相似度最大的 K 个元组（需要排序操作），统计元组中类型出现次数，选取出现次数最多的类别作为该测试向量的分类结果。分类结果与测试向量的类名进行比较，统计分类正确的次数与分类错误的次数，计算分类正确率。

三、 实验结果分析

1、多次修改 k 值观察其对分类准确率的影响，发现 k 取 10 时，实验准确率最高，约为 83.697%。如图 2 所示。

```
the performance of KNN:
('total test numbers:', 3754)
('number of success:', 3142)
('number of failure:', 612)
('the success rate:', 0.836973894512 52)
```

图 2 实验结果（k=10）

2、字典建立的规模也会影响准确率和运行时间，字典规模太小会使得分类准确率下降，规模太大则运行时间会大大增加。建立字典时，通过更改过滤词频大小来观察字典的规模，选择能得到合适规模的设置。

四、 感受体会

1、 整个实验使用 python 语言进行编程，由于之前没有使用 python 进行编程的经验，所以花费了一些时间去学习 python 的语法以及数

据结构。实验中切实感受到了 python 语言的简洁性以及许多模块和函数的方便，使用 python 的 nltk、os、math 模块在预处理等步骤上极大地减少了代码编写量，编程中要合理地选择列表、元组、字典等数据结构实现目的。

2、 VSM 以及 KNN 的实现，首先要根据思想确定每一步的步骤，弄清楚每一个步骤要达到的效果，分模块进行编程。在建立空间向量模型时，一开始对 tf 的理解出现了一些偏差。后来明白为数据集建立词典时，我们需要统计整个数据集内单词的出现次数来对出现在词典的单词进行过滤。而为每一个文档建立向量时，需要得到该文档中的每一单词的 tf_idf 值，此时的 tf 指的是该单词在该文档中的标准化词频，它的值等于某个词在该文档中出现的次数/该文出现次数最多的词出现的次数。

3、 实验的过程是不断发现问题与解决问题的过程，遇到问题要结合网上相关资料进行思考，弄清楚问题出现的原因以及解决办法。