实验报告

姓名：刘兵　　学号：201814819

任务一：预处理文本数据集，并且得到每个文本的 VSM 表示。

工具平台：python3.7.0

实现方法：

1.准备数据集（18828 个文档）；

2.依次**读取**每个文件夹内 80%文档；

3.对文件进行预处理（Txt2wordLst 函数），处理内容包括全部变为小写字母、去掉小于 3 个字母的单词、单词词型还原、去除非英语词、去除停用词。

4.去除词频低的词（判断标准为大于 3 或大于最大词频的一半的值），为选出来的每个词构建字典（键为单词，值为该单词在该文档中的词频即 tf 值，所建词典大小为 15677）。

5.每个文本处理后形成一个列表 topicLst

（三维,格式为：[[[文本]，……,[文本]],……, [[文本]，……,[文本]]]）。

6.根据上述列表构建词典（dict(Counter(tmp)）），为每个文本列表的每个单词计算出其在所有文档中出现的文档个数即 df 值，根据公式计算出每个文本对应到字典各单词的 tfidf 值（tf>0,tf=1+math.log(tf),idf=math.log(docNum/df)）。

7.将值表示为 list 并输出保存到磁盘为 txt 文本。

实验结果：

完成了 15076 个文档的 VSM 表示，其中每个文档使用 15677 长度的字典表示，每

个值为 TF_IDF 值。

运行结果数据详见压缩文件 vsmResult.rar（压缩文件大小为 875KB，解压后为 678MB）

任务二：实现 KNN 分类器，测试其在 20Newsgroups 上的效果

工具平台：python3.7.0

实现方法：

训练过程：

1）将该测试文档的向量与任务一中 60%文档的每一个文档向量计算距离。

2）使用每个文件夹内文件总数的第 60%~80%个文档进行训练（与前 60%数据依次求距离，根据前 k 个结果中哪个类占比最多得出属于哪个类，通过对比准确率调出最佳 k 值），具体过程与测试过程 2-6 项相同。

测试过程：

1. 逐个读取任务一数据集中未使用的 20%的文档。
2. 对每个文档进行预处理，调用 VSM.py 模块中的 Txt2wordLst 函数，处理内容与任务一处理方法一样。
3. 调用 VSM.py 中的 txtReduce 函数，去除各个文档词频低的词，方法与任务一一致。
4. 计算出根据任务一提供的字典的相关词在本文档中的 tf 值，计算出各词与训练集文档的 df 值，最后计算出该文档的 tf_idf 值的向量表示(向量长度为字典长度，对应字典中每个词的 ti_idf 值)。
5. 与前 60%数据依次求距离，每个距离选取前 k（根据训练结果，k 取 10）个结

果中占比最多的类别得出属于哪个类。

6. 根据该文档实际类别和判断出来的类别确定正确率，输出准确率。