作业一

实验过程：

##step 1：VSM 向量空间模型 （在VSM.py中实现）

##每个单词在所有文本的出现次数，以下简称大词典

##每个文本中单词词频，以下简称小词典

首先把文本进行预处理，将标点，数字等删除，然后把处理后的文本进行单词分割保存。保存单词用的是三维dict来存储文件路径的关系

首先建立好三维dict 遍历所有文本：一边统计单词在每个文本的词频和在每个文本的出现次数，计算TFIDF 并对词典进行筛选（将IFIDF大于15的加入到词典中），最后大词典大约有9000+个维度。将大词典保存在文本中，计算每个文本的向量，存为一行，每个文件夹存为一个文件，多少个文本即为多少行

###step2：KNN进行分类 （在KNN.py中实现）

首先把数据集分为测试集和训练集，其中20%做为测试 用VSM中的算法把测试集和训练集的向量求出再保存在文件中，然后用测试集的文本和训练集的文本向量求余弦相似度 把相似度最高的前k个结果进行统计，占比较大的就是结果

实验结果：0.695