Homework 1: VSM and KNN

姓名：史金婉　　学号：201834874

教师：尹建华　　项目：KNN 进行文本分类

一、实验目的

1. 学会 github 的使用，建立并管理自己的项目
2. 掌握一定的预处理文本的方法
3. 理解并掌握 Vector Space Model
4. 掌握 knn 算法并能利用它对文档进行分类

二、实验任务

1. 预处理文本数据集，并且得到每个文本的 VSM 表示
2. 实现 KNN 分类器，测试其在 20Newsgroups 上的效果。

三、实验数据

20 Newsgroups dataset

四、实验步骤

1. 文本预处理

（1）分词

使用 TextBlob 这个分词工具对 20 类文本进行分词。将文档划分成单词，并对单词做一些处理：大写字母变成小写字母，名词复数变单数，只保留由英文字母和-组成的单词，各种时态和形式的动词变成原形；

（2）划分数据集

将数据集划分成训练集和测试集，其中训练集占 80%，测试集占 20%；

（3）创建词典

从训练集中读取所有的文档，统计所有的单词及词频，去掉停用词后创建字典；

（4）计算 tf 和 idf

计算词典中所有单词的 tf 和 idf，tf（t，d）是单词 t 在文档中出现的次数，$\text{idf} = \log（\frac{\text{N}}{\text{df}（\text{t}）}）$，df（t）是含有单词 t 的文档数量；

（5）得到文本的向量表示

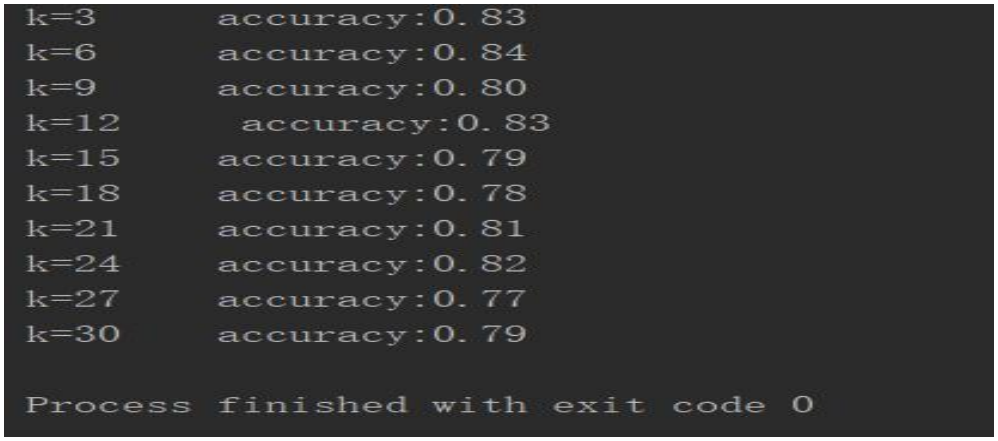
tf-idf(t, d) = tf(t, d)*idf(t)

2. KNN 文本分类

从测试集中读取一个文本的向量，然后计算它与所有训练集中所有文本的向量的 cos 距离，再把所有距离排序，前 k 个文档中类别最多的就是该文档的类型。本实验中的 k 设置为 3,6,9,12,15,18,21,24,27,30 这 10 个不同的值。

五、实验结果

结果如下图：



六、实验分析

1. 本实验中文本分类用到的 KNN 算法，，该方法的思路非常直观：如果一个样本在特征空间中的 k 个最相似的样本中的大多数属于某一个类别，则该样本也属于这个类别，该方法在类决策上只依据最邻近的一个或者多个样本的类别来决定带分样本所属类别；
2. 本实验中，k 值的选取不同，准确率也是不同的，为了得到最好的分类结果，可以进一步尝试更多的 k 值；
3. KNN 进行文本分类的这个方法的一个不足之处就是计算量比较大，因为对每一个待分类的文本都要计算它到全体已知样本的距离，才能求得它的 k 个最近邻点。