一. 实验过程：

1.读取文档，以便进行预处理，为了便于遍历，按照文档原本的结构保存了文本。

2.预处理

首先要进行分词，利用 textblob 工具进行分词的时候还做了一部分预处理工作，主要进行了去除停用词，数字，以及 stemming。

考虑到会有低频词的出现，在后续构建字典时还去除了词频过低的单词，令字典更加精简。

3.统计词频，建立词典，生成 VSM

计算当前文档词频，返回一个字典 dict1，计算全局文档词频，返回一个字典 dict2。

基于 dict2 建立词典。

基于字典 dict1 和词典计算文档中单词的 tf-idf。

最终得到文档的 VSM 表示并进行存储。

实验得到的 VSM 比较稀疏，没有做进一步的处理，占用空间比较大。

4.取出文档的 VSM，划分训练集和测试集

划分时使用了 sklearn 的 train_test_split，可以更方便地在每个类别划分出相同比例的测试数据和训练数据。

5.KNN 分类

输入测试集中的文档，计算其与训练集中每个文档的 VSM 的相似度，这个距离度量采用了两种指标，cosine distance 和 euclidean distance。

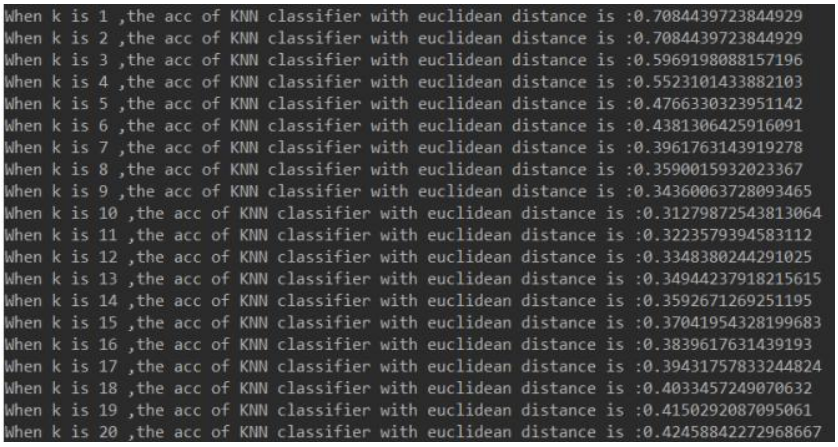
根据相似度找出与测试文档最近的 K 个文档，判断预测应得的类别标签。

计算分类准确度。

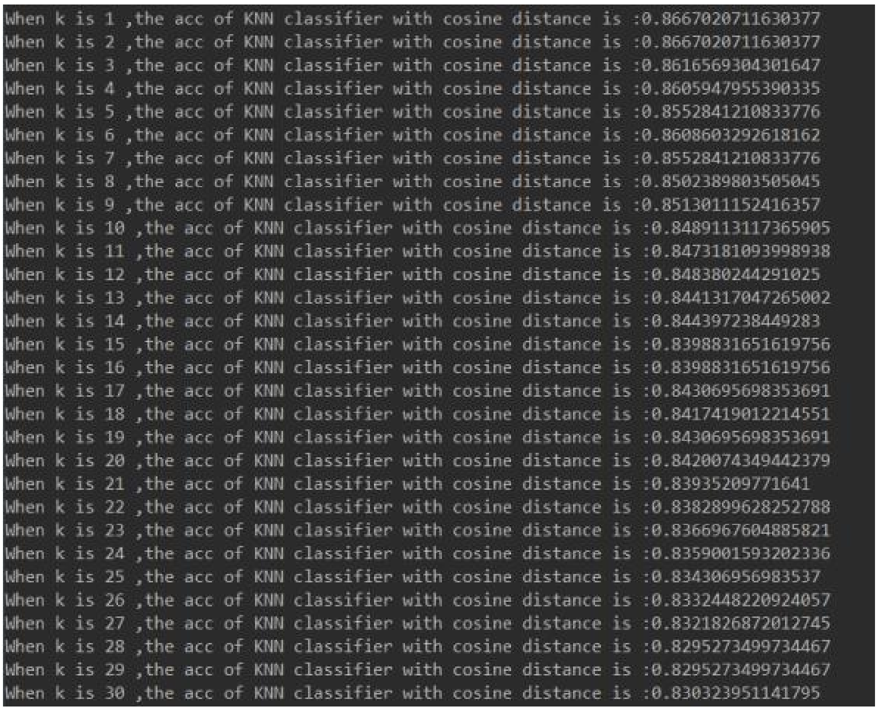
二. 解决问题：

1. 通过实验，掌握了 **python 对文档的读取和存储**等操作。
2. 最初距离度量函数是我自己写的，主要通过遍历数组完成相似度计算。测试过程中发现这个环节占用时间太长了，算一条测试数据大概需要 7s，3000 多个测试文档算完一轮大概需要 5，6 个小时。后来使用了 sklearn 的 **metrics.pairwise** 工具，测试集和训练集作为两个矩阵输入，可以很快得到结果，能够测试更多的 K 值。体会到了**矩阵操作以及工具库**的便利。

三. 实验结果：



```
When k is 1 ,the acc of KNN classifier with euclidean distance is :0.7084439723844929
When k is 2 ,the acc of KNN classifier with euclidean distance is :0.7084439723844929
When k is 3 ,the acc of KNN classifier with euclidean distance is :0.5969198088157196
When k is 4 ,the acc of KNN classifier with euclidean distance is :0.5523101433882103
When k is 5 ,the acc of KNN classifier with euclidean distance is :0.4766330323951142
When k is 6 ,the acc of KNN classifier with euclidean distance is :0.4381306425916091
When k is 7 ,the acc of KNN classifier with euclidean distance is :0.3961763143919278
When k is 8 ,the acc of KNN classifier with euclidean distance is :0.3590015932023367
When k is 9 ,the acc of KNN classifier with euclidean distance is :0.34360063728093465
When k is 10 ,the acc of KNN classifier with euclidean distance is :0.31279872543813064
When k is 11 ,the acc of KNN classifier with euclidean distance is :0.3223579394583112
When k is 12 ,the acc of KNN classifier with euclidean distance is :0.3348380244291025
When k is 13 ,the acc of KNN classifier with euclidean distance is :0.34944237918215615
When k is 14 ,the acc of KNN classifier with euclidean distance is :0.3592671269251195
When k is 15 ,the acc of KNN classifier with euclidean distance is :0.37041954328199683
When k is 16 ,the acc of KNN classifier with euclidean distance is :0.3839617631439193
When k is 17 ,the acc of KNN classifier with euclidean distance is :0.39431757833244824
When k is 18 ,the acc of KNN classifier with euclidean distance is :0.4033457249070632
When k is 19 ,the acc of KNN classifier with euclidean distance is :0.4150292087095061
When k is 20 ,the acc of KNN classifier with euclidean distance is :0.42458842272968667
```



```
When k is 1 ,the acc of KNN classifier with cosine distance is :0.8667020711630377
When k is 2 ,the acc of KNN classifier with cosine distance is :0.8667020711630377
When k is 3 ,the acc of KNN classifier with cosine distance is :0.8616569304301647
When k is 4 ,the acc of KNN classifier with cosine distance is :0.8605947955390335
When k is 5 ,the acc of KNN classifier with cosine distance is :0.8552841210833776
When k is 6 ,the acc of KNN classifier with cosine distance is :0.8608603292618162
When k is 7 ,the acc of KNN classifier with cosine distance is :0.8552841210833776
When k is 8 ,the acc of KNN classifier with cosine distance is :0.8502389803505045
When k is 9 ,the acc of KNN classifier with cosine distance is :0.8513011152416357
When k is 10 ,the acc of KNN classifier with cosine distance is :0.8489113117365905
When k is 11 ,the acc of KNN classifier with cosine distance is :0.8473181093998938
When k is 12 ,the acc of KNN classifier with cosine distance is :0.848380244291025
When k is 13 ,the acc of KNN classifier with cosine distance is :0.8441317047265002
When k is 14 ,the acc of KNN classifier with cosine distance is :0.844397238449283
When k is 15 ,the acc of KNN classifier with cosine distance is :0.8398831651619756
When k is 16 ,the acc of KNN classifier with cosine distance is :0.8398831651619756
When k is 17 ,the acc of KNN classifier with cosine distance is :0.8430695698353691
When k is 18 ,the acc of KNN classifier with cosine distance is :0.8417419012214551
When k is 19 ,the acc of KNN classifier with cosine distance is :0.8430695698353691
When k is 20 ,the acc of KNN classifier with cosine distance is :0.8420074349442379
When k is 21 ,the acc of KNN classifier with cosine distance is :0.83935209771641
When k is 22 ,the acc of KNN classifier with cosine distance is :0.8382899628252788
When k is 23 ,the acc of KNN classifier with cosine distance is :0.8366967604885821
When k is 24 ,the acc of KNN classifier with cosine distance is :0.8359001593202336
When k is 25 ,the acc of KNN classifier with cosine distance is :0.834306956983537
When k is 26 ,the acc of KNN classifier with cosine distance is :0.8332448220924057
When k is 27 ,the acc of KNN classifier with cosine distance is :0.8321826872012745
When k is 28 ,the acc of KNN classifier with cosine distance is :0.8295273499734467
When k is 29 ,the acc of KNN classifier with cosine distance is :0.8295273499734467
When k is 30 ,the acc of KNN classifier with cosine distance is :0.830323951141795
```

1. 可以看到两种度量方式，均是 K=1 时精确度最高。对此，我的理解是，k=1 时，会找到训练集中与测试文档最为相似的一个文档，由此判断二者的标签相同。

2. 对于 euclidean distance 度量方式，当 k = 10 时精确度降到了最低，随后精确度开始慢慢的回涨，实验过程中我测试到了 k=60，发现精确度涨到了 0.55 左右。可以理解为范围越大，容错性也在慢慢变好，对比较难检测分类（可供分类的特征单词比较少）的文档比较好。