文本相似性检测与抄袭判断

1.算法流程解析

对于检测一篇文章是不是新华社所发布的，并对一篇文章进行抄袭判断的问题，可将其分成两个问题，第一个是文本分类，第二个是文本距离计算。

对于文本分类，首先想到的是通过使用tf-idf将文本生成向量空间，然后再通过机器学习模型将文本进行分类；模型选择方面，由于文本中已有target label,因此只需使用监督学习模型即可，在数据量较大的情况下，SVM和KNN模型的结果比较好，但是由于电脑性能较差，25000个文本是上限，NB因为比较简单选择使用；模型衡量标准，由于是分类问题，所以使用了精准率、召回率、f1-score和准确率四种标准参考。还有将数据集进行80%训练集，20%测试集分类。

算法流程为：

输入数据——数据分集，去除stopWord——tfidfg构建向量 —— 使用NB机器模型进行训练——使用editdistance定位错误位置.

2.遇到的问题及思考

1）电脑配置仅为苹果Mac air 8g 内存，i5处理器，由于电脑运算速度慢，经测试发现25000个文本向量是nb算法的上限，KNN和SVM的上限更小。曾考虑使用GPU服务器解决问题，但是由于价格较贵以及时间问题没有尝试，考虑以后换台电脑；

2）由于对于模型搭建流程不熟练，第一次测试时得到的准确率时0，当时信心大受打击，2天没有更进一步学习。后来发现时数据集只使用了500导致的。思考以后遇到难题需要调整心态；

3）使用editdistance定位错误位置没有完成，因为时间和知识所限，只能得到每一个疑似抄袭文章和每一篇新华社文章的distance组成的列表。

4)最后得到结果遇到了`AttributeError: 'float' object has no attribute 'decode'`这个问题，需要解决。