Homework1：vsm and knn

学号：201814810
姓名：段新朋

一、 实验内容

本实验包括两部分内容：
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

二、Vsm

在这个部分，需要对数据集进行预处理，并得到每个文本的 vsm 表示，该过程主要包括以下部分：
1、分词：采用正则表达式以及 python 的 split 函数进行分词，以所有的非字母符号作为分割符号('[^a-z]*'，便可得到分词的结果
2、对单词进行处理：用 nltk，textblob 等包对单词进行以下处理，以减小词典大小：去掉单词长度小于 3 的词，去掉停用词，复数变单数，动词形式变为一般形式，去掉文档中出现频率小于某个阈值的单词。
3、统计单词的 tf，idf 等数据，由于用到了倒排索引的思想加速 knn 的运行，所以需要以单词为 key 值，统计 word_idf,word_doc_tf,doc_word_tf 等三个变量。
4、以某个测试数据文本中出现的单词为基础，通过 word_idf,word_doc_tf,doc_word_tf 三个字典，建立该测试文本，以及所有出现该文本中单词的训练文本的 vsm 表示。

三、Knn

按照上面 vsm 过程中建立的测试文本以及相应训练文本的 vsm 表示，即可按照 knn 的思想为测试文本分类，并得到计算准确率，做的工作主要有：
1、写了两种计算向量相似性的方法：欧几里得距离计算相似性，cos 值计算相似性。
2、采用 inverse index 的方式加速了 knn 的过程。
3、实现了 knn 算法。

四、实验结果

通过本次实验，得到了以下的结论：
1、当数据量巨大的时候，knn 方式进行分类的计算速度非常慢，很长时间都无法得到最终结果。
2、使用 inverse index 可以加速该计算过程。
3、最终得到的准确率是：0.79，