姓名：付志杰　　学号：201800109

实验一

本次实验是处理文本数据集，得出 VSM 空间向量，knn 分类器计算测试集相似度

一、Shujuchili.py 文件是对源数据 20news-group 进行预处理，提取干净的数据集，用读写操作将处理后的数据储存下来。遍历所有预处理文件，统计文件数目，将百分之八十数据作为训练集，将百分之二十作为测试集进行分类。

二、vsm.py 文件是将生成的文档表示为向量，建立词典，统计词项出现的频率，去除掉频率太小的词项。最后将所有文档表示成向量列表。计算出所有的 tf-idf 值。

三、knn.py 文件是对测试集的每一个向量，计算出它与训练向量的相似度，将训练向量的类型和相似度作为二元组存储在列表中，取出列表中相似度最大的元组，统计元组中类型出现次数，选取出现次数最多的便是分类出来的结果。

实验二

实验目的：

使用朴素贝叶斯分类器，测试其在 20Newsgroups 数据集上的效果。

类别概率=类条件概率 * 先验概率。

实验步骤：

(1)首先对文本进行处理并将处理的文字和标签分别存入字典中，将其向量化；

(2)统计类的总数量；

(3)调用几种不同的聚类算法；

(4)测试聚类后的效果。

实验结论：

贝叶斯分类器根据测试集每个词属于每个类的概率，计算出每个文档属于每个类的概率，并选择概率最大的，将文档归类；根据实验测试，贝叶斯分类器的分类效果明显，简单好用，是个很好的分类工具。

实验三

实验目的：

使用 Tweets 数据集测试各种聚类算法。

实验步骤：

对数据进行划分，80%作为 training data，20%作为 testing data，作为测试集和训练集；

统计每个类中单词总数及出现次数；

计算类条件概率和先验概率；

计算样本属于类别的概率，对文档进行分类。

实验结果：

KMeans：0.721292248386

Affinity：0.734286617625

Spectral：0.689574578741

Agglomerative：0.757178787482

Gaussian：0.683066495726

MeanShift：0.690981587324

Dbscan：0.738217267295

实验心得：

第一次使用 scikit-learn，虽然有些函数不太了解，过程中也会出现一系列的错误，不过 scikit-learn 对于聚类确实是很好用，学到了很多新的东西，收获了很多，以后还会加强练习，有更新的认知。