Data Mining Report

姓名：王睿　　学号：201834877　　专业：计算机技术

一、 VSM+KNN

1、下载数据集 20 Newsgroups dataset，在数据集处理过程中，发现文本编码有问题，采用自动转编码方式读取数据集。

2、处理数据集（运用 nltk 包进行处理）：对数据集中单词进行词形还原； 去掉所有词中不带字母的单词，剩下的分词存到 list 中；运用 nltk 的停用词包，将 list 中单词的停用词去掉，并且将所有词转化为小写；将分词结果以 list 的格式按原文件名存到磁盘中以备后续计算使用。（其中拼写检查部分及词干提取本次不进行设计）

3、 VSM 的生成： 读取所有处理过的数据文件拼接成一个字符串生成词典，并且运用 sklearn 包中的 TfidfVectorizer 函数进行 vsm 的 tfidf 值生成（之前写过 tfidf 的 java 版生成，在 tfidf_java 中，含报告，此处不再进行 tfidf 的相关工作）。

4、 KNN 计算：编写计算向量的 cos 函数，对 vsm 进行 cos 距离计算，排序，得到最终的匹配结果。

二、NBC

1、 获取所有的文本类别；

2、 处理数据集数据，获取每一个文本数据及对应的类别标签；

3、 按照 VSM 的方法得到文本的 TFIDF 值；

4、 调用 sklearn 库中的 GaussianNB 方法获取文本的朴素贝叶斯的高斯模型（训练集过大，可用 partial_fit 分批训练模型）；

5、 输入测试集， 进行评估获取预测；

三、Clustering with sklearn

1、 获取 Tweets 中的 text；

2、 处理数据集，调用 TfidfVectorizer 得到 text 的 tfidf 矩阵；

3、 调用 sklearn 中的聚类方法进行聚类；

4、 结果比对。