Data Mining
实验报告

姓名：徐万龙

学号：201844908

Homework1　VSM+KNN

一、实验任务

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、实验数据

20 Newsgroups

三、实验过程

按照训练集：测试集=8:2 的比例划分数据集。

1、处理 20news-18828 数据集，生成文本对应的 VSM 表示。

（1）根据文件夹顺序读取数据集。

（2）依次通过去特殊字符、统一小写字母、分词、词干提取、去停用词等步骤进行数据预处理。

（3）选取训练集中词频大于 9 且小于 10000 的词创建词典。

（4）计算 TF-IDF 值以及生成文档的向量表示。

2、计算每个测试集文档与训练集之间的相似度，利用 KNN 算法选取不同的 K 值进行分类，最终选取 K 值为 30，准确率为 79.2%。

Homework2　Naive Bayes classifier

一、实验任务

实现朴素贝叶斯分类器，测试其在 20 Newsgroups 数据集上的效果

二、实验数据

20 Newsgroups

三、实验过程

按照训练集：测试集=8:2 的比例划分 20news-18828 数据集。

1、分别读取训练集和测试集并进行数据预处理，选取训练集中词频大于 9 且小于 10000 的词创建词典。

2、采用通过平滑处理的多项式朴素贝叶斯分类器对测试集数据进行分类，准确率为 83.3%。

Homework3　Cluster

一、实验任务

测试 sklearn 中聚类算法在 tweets 数据集上的聚类效果，并使用 NMI 作为评价指标。

二、实验数据

Tweets.txt

三、实验过程

使用 sklearn 提供的 TfidfVectorizer 工具将文本转化为 TF-IDF 向量形式表示，分别调用 8 种聚类方法，调整参数，使用 NMI 评价指标分别评价 8 次不同聚类方法所得到的聚类效果。

K_Means: 0.8023213665437697

Affinity Propagation: 0.7831387602380028

Mean_Shift: 0.7265625

Spectral Clustering: 0.8161466348167306

Ward Hierarchical Clustering: 0.7843154591464186

Agglomerative Clustering: 0.7843154591464186

DBSCAN: 0.6987909501721541

Gaussian Mixture: 0.782922639478339