DataMining 实验报告

姓名：李聪聪 学号：201844900

作业一：VSM/KNN

一、实验目的

1. 学会 github 的使用，建立并管理自己的项目

2. 掌握一定的预处理文本的方法

3. 理解并掌握 Vector Space Model

4. 掌握 knn 算法并能利用它对文档进行分类

二、实验要求

1. 对原始数据进行预处理，得到每个文档的 VSM 表示。

2. 实现 KNN 分类器，测试其在 20Newsgroups 数据集上的效果。

三、文件夹目录说明:

1. vector:导入数据->数据预处理（大小写转化，去除特殊字符，去除停用词，词干提取等操作)->创建词典->创建 0-1 型向量

2. knn:将数据利用 sklearn.model_selection 中的 train_test_split 函数划分为测试训练集和测试训练集->计算余弦相似度->>计算准确率

3. 关于数据（开始全部的数据是放在 data 目录下,但是一直没法将其传到 gitup 上，所以项目原本是在DataMining 下有个 data 文件夹存放数据 ）

4. out 文件夹下存放数据处理过程中的结果（其中 Input_data,labels,dict,vsm 是预处理阶段产生的，dictionary,text_x,text_y 是 knn 阶段产生的）

三、 实验结果:

knn 准确率：当 k=30 时，准确率为 0.79

心得体会：由于之前没使用过 gitup，没学过机器学习和 python 语言，此次作业虽然过程很艰难但是收获也很大。

作业二：NBC

一、实验要求

实现朴素贝叶斯分类器分类文档，测试其在 20Newsgroups 数据集上的效果。

二、实验步骤:

1. 调用 sklearn.model_selection 中的 train_test_split 函数划分为 80%的测试训练集和 20%的测试训练集。

2. 调用作业一的函数读取数据、生成词典，构建词典。

3. 采用贝叶斯的多项式模型，并进行平滑处理。

4. 实验结果：准确率：0.84

作业三： Clustering with sklearn

一、实验目的

1. 了解 sklearn 工具中的各种聚类算法;

2. 掌握 NMI(Normalized Mutual Information)评价指标;

3. 掌握 sklearn 中的七种算法并能利用它对文档进行分类

二、实验任务

测试 sklearn 中聚类算法在 tweets 数据集上的聚类效果。

三、实验数据：

Tweets.txt

四、实验过程：

1. 文本预处理：

利用 sklearn 中 tfidfvectorizer 中的 fit_transform 函数将获取到的文本向量化，并用 toarray()函数转化为稀疏矩阵。

2. 调用各种聚类方法

分别调用 8 种聚类方法，并利用 normalized_mutual_score()函数进行结果正确率预测

K-means accuracy:0.79

AffinityPropagation accuracy: 0.78

MeanShift accuracy: 0.75

SpectralClustering accuracy: 0.83

AgglomerativeClustering accuracy: 0.78

DBSCAN accuracy: 0.70

GaussianMixture accuracy: 0.78