201834890ZhangZhao

Homework of Data Mining

Final product description

HomeWork 1：KNN

> 1. 预处理过程：token->normalization（去特殊字符、小写、判断是否英语单词）->Stemming->Stopword
> 2. 构造词典：去除频率4以下token。最终词典大小：18708
> 3. 实现01型与tf-idf权重型space vector
> 4. 实现通过计数法、1/d*d 权重法进行KNN分类
> 5. KNN分类过程中保存的中间文件：
> 6. 词典：data/knn-out/dictionary.csv
> 7. 预测结果：data/knn-out/prediction.csv
> 8. 训练集预处理结果：data/knn-out/train_X.csv ; data/knn-out/train_Y.csv
> 9. 测试集预处理结果：data/knn-out/test_X.csv ; data/knn-out/test_Y.csv
> 10. 测试集准确率：（测试集按层次划分，占数据20%）

实现	准确率	K值
01型+计数法	0.729	30
01型+权重法	0.753	42
tf-idf+计数法	0.733	40
tf-idf+权重法	0.742	40

HomeWork 2：NBC

> 1. 调用homework1的vsm.py读取数据、生成词典。并使用knn分出的train set、test set。（8 2分）
> 2. 构造词典：过滤词频大于2000的token
> 3. 采用多项式模型实现，并进行平滑处理。
> 4. 测试集准确率：0.802

HomeWork 3：Cluster

> 1. 调用homework1的vsm.py生成词典。并调用knn.py生成tf-idf型vsm。
> 2. 使用NMI(Normalized Mutual Information)评估聚类效果
> 3. score:

实现	score
K-Means	0.771
AffinityPropagation	0.733
MeanShift	0.110
SpectralClustering	0.759
DBSCAN	0.733

AgglomerativeClustering：

linkages	ward	average	complete
score	0.792	0.167	0.463

GaussianMixture：

cov_types	spherical	diag	tied	full
score	0.663	0.716	0.727	MemoryError
