结果分析报告

一、数据分析

训练集	样本数	占比（总样本）	类数	占比（总类）
样本总数	738793	100%	3150	100%
出现次数小于 2	474	0.06%	474	15%
出现次数小于 10	4062	0.50%	1269	40.20%
出现次数小于 20	9865	1.33%	1683	53.40%
测试集	样本数	占比（总样本）	类数	占比（总类）
样本总数	13421	100%	1670	100%
出现次数小于 2	513	3.80%	513	30.70%
出现次数小于 10	4016	29.20%	1394	83.50%
出现次数小于 20	5992	44.60%	1541	92.30%

训练集总词数（去重）	232165
测试集总词数（去重）	9987

二、构建数据集

方法一：

1. 用 jieba（精确模式）对训练集和测试进行分词；
2. 使用 collection 中的 Counter 包进行词频统计；
3. 分别在完整训练集和处理训练集（去掉出现次数小于 2 的样本）中，通过选择出现频率最高的前 1、3、5、7 万个词作为特征维度，然后构建 TF-IDF 词频矩阵；

方法二：

1. 将训练集和测试集中的样本，按单个字符分开；
2. 去掉样本中的数字，构建 TF-IDF 词频矩阵；(738793,4884)
3. 利用随机森林对词频矩阵中的特征按重要性进行排序，构造一个 1000 维度的数据集；

三、预测

主要采用了三个模型：随机森林、支持向量机（批量拟合）、Kmeans 聚类思想（计算相似性）。

（1）如下为利用方法一构建的数据集，在三个模型上的表现：

model	dimension	accuracy	traing time
RandomForest	10000	0.368	15 min
	10000_2	0.368	20 min
	30000	0.422	25 min
	30000_2	0.419	24 min
	50000	0.434	16 min
	50000_2	0.432	30 min
	70000	0.4367	18 min
	70000_2	0.438	33 min
SVM_partial_fit	10000	0.32	
	10000_2	0.32	
	30000	0.37	
	30000_2	0.38	
	50000	0.4	
	50000_2	0.4	
	70000	0.39	
	70000_2	0.39	
Kmeans(计算相似性)	10000	0.248	3h 37min
	30000	0.318	1h 36min
	70000	0.37	3h

注：‘10000_2’表示去掉训练集中，出现次数少于 2 次的样本点后，选取词频最高的前 10000 个构造词频矩阵。

错误样本原因：

1. 从下表可知，测试集中有大约 30%的词不存在于训练集中；
2. 在使用随机森林训练时，受限于内存容量，决策树的个数只能设置为 10，常见一般要设置 100 以上
3. 模型超参数设置可能并未达到最优

训练集词频前 K 万个与测试集中重复的词，占测试集总词数（去重）的比重

词频	占比（测试集）
训练集词频前 1 万与测试集的重复词	41.30%
训练集词频前 2 万与测试集的重复词	54.30%
训练集词频前 3 万与测试集的重复词	60.72%
训练集词频前 4 万与测试集的重复词	64.65%
训练集词频前 5 万与测试集的重复词	67.53%

训练集词频前 6 万与测试集的重复词	69.06%
训练集词频前 7 万与测试集的重复词	70.41%
训练集词频前 8 万与测试集的重复词	71.27%
训练集词频前 9 万与测试集的重复词	72.76%
训练集词频前 10 万与测试集的重复词	72.89%
训练集词频前 12 万与测试集的重复词	73.30%
训练集词频前 14 万与测试集的重复词	73.61%
训练集词频前 16 万与测试集的重复词	73.94%
训练集词频前 18 万与测试集的重复词	74.28%
训练集词频前 20 万与测试集的重复词	74.77%
训练集词频前 22 万与测试集的重复词	75.23%
训练集所以词与测试集的重复词	74.44%

（2）如下为利用方法二构建的数据集，在三个模型上的表现：

RandomForest	4884	n_estimator = 12	max_depth = NONE	0.467	10 min 05
	1000	n_estimator = 15	max_depth = NONE	0.459	8 min 46
SVM_partial_fit	4884			0.35	
	1000			0.33	
Kmean(计算相似性)	4884			0.34	
	1000			0.27	

代码说明：

任务 1\Task1\code\data　该目录对应初始数据集
任务 1\Task1\code\Task1_1\class_analyse.py　作用为数据分析
任务 1\Task1\code\Task1_1\main.py 里面包含如下 5 个方法 用于方法一构建的数据集

Load_Original_Traindata_Testdata_Cut_and_Save() # 分词并保存

Make_Stop_Words_and_Save(dimensions=70000)　# 选择停用词（总词减去前 n 个最高频率的词）

Load_Stop_Words() # 载入停用词

Make_Train_and_Test_Tf_Idf_and_Save() # 构建词频矩阵并且保存

Load_Traindata_Testdata_with_Tfidf(filename) # 载入 TF-IDF

train_by_RandForest(filename) # 随机森林 模型
train_by_partial_SGD(filename)# 支持向量机 模型

任务 1\Task1\code\Task1_2\kmeans.py # 计算相似性 模型
任务 1\Task1\code\Task1_1_1|single_word.py 用于方法二构建的数据集