

## 实验报告

### 一、 实验内容

#### 1、 构建 vsm

首先读取每个文本，对其分词、steming、去停用词等处理，初步建立词典。同时计算每个文本中各单词出现的频率。因为频率受文本大小的影响，所以使用 Sub-linear TF scaling 公式对其标准化。把上述过程生成的结果存放再 data\_process 文件夹中。

其次，读取上一步处理过的文本，计算词语的 IDF，从而根据  $w=IF*IDF$  得到词语的权重。

最后，设置权重阈值  $w\_min$  过滤词典，根据词典构建文本的向量。

阈值  $w\_min$  与词典规模的对应关系

阈值 $w\_min$	词典规模（个）
0	221587
5	220586
15	31472
20	8685
25	2200

#### 2、 KNN

按照一定 9/1 的比例划分训练集和测试集，使用向量间

的余弦值度量相似性。计算测试集样本与训练集每个样本的相似度，按相似度从大到小排序后，看排在前 k 个的训练集样本属于哪个类型的比较多，就把测试集样本分到哪个类型。设置不同的 K 值，选择分类准确率高的 K 值，最后输出分类的准确率。

K 值与 acc 的对应关系

K	Acc
6	0.65
8	0.72
10	0.81

## 二、实验日志

### 1、遇到的困难

(1)、文本数量太多，构建 vsm 时生成的词典规模过大，算起来非常慢。

(2)、KNN 分类时，运行时间特别长。

### 2、解决方案

(1)、对于词典规模过大的问题，设置最小权重阈值，从全局考虑来过滤词典，可以大大缩小词典规模。

(2)、对于 KNN 分类运行时间过长的的问题，从两方面来优化，一是划分训练集和测试集的时候，缩小测试集的规模；二是使用部分训练数据来估算分类的准确率。

