

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。

其次，遍历词典，计算每个单词的 IDF，从而根据 $w=TF*IDF$ 得到单词的权重。

最后，设置权重阈值 w_min 重新建立词典，单词权重大于 w_min 的放入词典中，从而达到缩小词典规模的效果，再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	110000
5	50586
15	31472
20	6184
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。 K 值设定为 9，得到分类正确率 0.763， K 值最好选奇数，为了避免出现“平局”情况。

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。我采用的方法是缩小测试集的规模，从 $2/8$ 缩小到 $1/9$ 。