Data Mining Report

一、作业题目

运用 20 Newsgroups dataset,处理文本数据集,并且得到每个文本的 VSM 表示;实现 KNN 分类器,测试其在 20Newsgroup 上的效果。

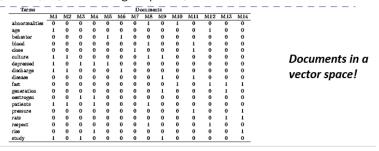


图 1.1 VSM 表示方法

$$- \ \text{Cosine Similarity} \qquad \qquad cosine \left(d_i, d_j\right) = \frac{V_{d_i}^T V_{d_j}}{\left|V_{d_i}\right|_2 \times \left|V_{d_j}\right|_2}$$

图 1.2 KNN 计算方法

二、实现思路

- 1、下载数据集 20 Newsgroups dataset, 在数据集处理过程中, 发现文本编码有问题, 采用自动转编码方式读取数据集。
- 2、处理数据集(运用 nltk 包进行处理): 对数据集中单词进行词形还原; 去掉所有词中不带字母的单词, 剩下的分词存到 list 中; 运用 nltk 的停用词包, 将 list 中单词的停用词去掉,并且将所有词转化为小写; 将分词结果以 list 的格式按原文件名存到磁盘中以备后续计算使用。(其中拼写检查部分及词干提取本次不进行设计)
- 3、VSM 的生成: 读取所有处理过的数据文件拼接成一个字符串生成词典, 并且运用 sklearn 包中的 TfidfVectorizer 函数进行 vsm 的 tfidf 值生成(之前写过 tfidf 的 java 版生成, 在 tfidf_java 中, 含报告, 此处不再进行 tfidf 的相关工作)。
- 4、KNN 计算:编写计算向量的 cos 函数,对 vsm 进行 cos 距离计算,排序,得到最终的匹配结果。