Leihai Nie

USTC  PB16080377

HW6

Problem of the Week

|  |
| --- |
| **Web Content Mining: Classification**  kNN with inverted index  1) 在kNN算法中，对一个document分类可以搜索所有的documents，并找出其中距离最近的k个。试分析该过程的时间复杂度。  2) 如果我们使用document的tf.idf向量的余弦作为相似度，请思考是否可以使用倒排索引来查询一个document的k近邻。如果可以，请讨论该方法适用于何种情况，描述详细过程，并分析该方法和1中方法的区别。 |

Final Answer Box: Put your final answer here. Describe or explain it in one complete sentence. (Do not just write a number.)

|  |
| --- |
| 1. 需要O(nd+kn) or O(ndk).(d为数据维数,n为测试数据的势)   O(nd+kn):对训练集每个数据，计算和指定点的距离dist，复杂度为O(nd).随后遍历1到k，每次都选取尚未选中中的相关度最大值，这里复杂度为O(kn).  O(ndk):对于j=1 to k ,每次都循环遍历计算dist，并取出最大者。由于为乘法操作，所以复杂度为O(ndk)   1. 我觉得可以。适用条件：要求相关性假设(需要知道在文档集中使用哪些词项来表达，相关文档中出现的词项类似)   倒排索引主要由处理人员根据相关问题、内容提取出来关键词，(根据document相关类型、内容，事先确定关键词)然后计算出各关键词tf-idf值，根据权重组成词项向量，最后可以不同向量间的余弦相似度。从而可以遍历1 to k，每次取走相似度最大值，即可获得k近邻。  区别：第二种方法依赖于词频和权重，计算相似度在于余弦，与向量方向有关，不同词项的存在会影响相似度。Knn也是设置词项向量，通过距离来衡量相似度，但距离以及词项向量的设置就有更多可能性。 |

|  |
| --- |
|  |

Final Answer Box: Put your final answer here. Describe or explain it in one complete sentence. (Do not just write a number.)

|  |
| --- |
|  |