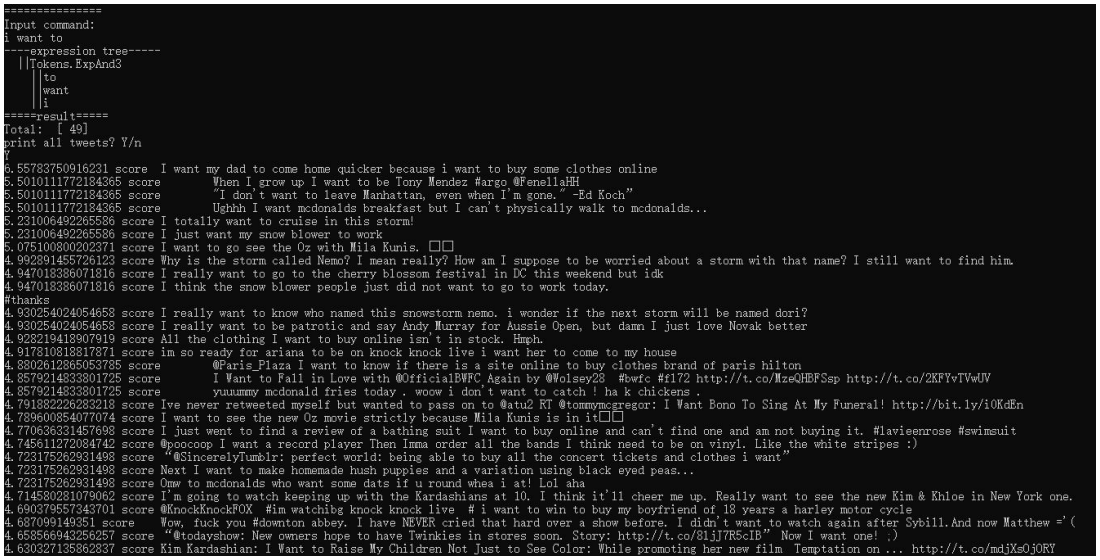
山东大学 计算机科学与技术 学院

信息检索与数据挖掘 课程实验报告

学号：201600130032	姓名： 陆禹承	班级： 智能 16
实验题目：BM25		
实验内容： 实现 BM25 模型		
实验过程中遇到和解决的问题： 使用和上次相似的代码，主要更改的是增加 rating 功能 算法过程： 首先根据用户查询筛选出文档，然后对每个文档进行 rating（这些文档数据已经被分过词） rating 的式子大致如下： • Pivoted Length Normalization VSM [Singhal et al 96] $f(q,d)=\sum_{w\in q\cap d} c(w,q)\frac{\ln[1+\ln[1+c(w,d)]]}{1-b+b\frac{\lvert d\rvert}{avdl}}\log\frac{M+1}{df(w)}$ 其中常数值根据多次尝试自行选择了一个合适的数值，这样排名出来的效果会好很多 此外还对运算过程进行了优化，计算速度都有大幅度的提升。多个 AND/OR 运算可以同时运算（使用一个小根堆进行排序），对于 a&b&c 这种情况效率有小幅度提升。 为了方便 rating 的计算， 做 inverted index list 时统计了词汇频率（tf），并统计所有文档长度 结果出来以后，再对每个文档进行评价，最后根据得分排序。 评析时对每个查询词计算结果求和 文档处理要花费几分钟的时间，为了方便调试，将处理过的文档信息暂存下来（cache），方便之后读入使用（只需要几秒的时间） 查询花费时间较小，简单查询基本<1s 运行效果：		



结论分析与体会：

对模型优化有所了解