



数据库专题训练

数据库新型检索技术

作业一 近似查询

助教 李浩达 lihaoda9@163.com



实验框架



- 请参考框架代码，实现SimSearcher类的方法：
 - createIndex() 函数
 - searchJaccard() 函数
 - searchED() 函数
- 请不要修改这三个方法的声明，可以根据需要自行添加其他方法。



createIndex函数



- 函数声明: `int createIndex(const char *filename, unsigned q);`
 - filename: 输入文件:
 - 每行一个字符串, 代表一条记录
 - 每条记录的id为其行号, 从0开始
 - q: 如使用qgram方法, 该参数为q值
 - 作用: 读取指定的输入文件并建立索引
 - 创建成功请返回SUCCESS
 - 创建失败请返回FAILURE



searchJaccard函数



- 函数声明: `int searchJaccard(const char *query, double threshold, vector<pair<unsigned, double>> &result);`
 - query: 查询串
 - threshold: Jaccard阈值
 - vector<pair<unsigned, double>> &result, 返回的结果, 每个pair是<字符串id, 与query之间的Jaccard相似度>, 需按照id从小到大排序, 且无重复结果
 - 返回值同createIndex



searchED函数



- 函数声明: `int searchED(const char *query, unsigned threshold, vector<pair<unsigned, unsigned> > &result);`
 - query: 查询串
 - threshold: ED阈值
 - `vector<pair<unsigned, unsigned>> &result`, 返回的结果, 每个pair是<字符串id, 与query之间的编辑距离>, 需按照id从小到大排序, 且无重复结果
 - 返回值同createIndex



实验要求



- 实验平台：Ubuntu, g++ 4.8
- 评测标准：
 - ✓ 正确性：
 - ✓ 返回的结果均满足查询要求
 - ✓ 满足查询要求的结果全部被返回
 - ✓ 时间：跑的越快越好，最终以速度排名为依据给分
 - ✓ 空间：要求能够跑动最终评测数据集（一般不需考虑）
- 提交材料：
 - 所有的源代码，不包括main.cpp
 - 设计文档：描述算法设计思路以及实现难点
 - 网络学堂提交
- 截止时间：以网络学堂为准



评测说明



- 最终编译会采用给定的makefile，大家可以自行测试自己的代码是否能通过编译
- 可以使用c++11中的特性来简化代码，可以使用stl标准库
- 请不要使用多线程等手段来加速程序
- 最终提交文件中请不要包含main函数，以避免链接失败。最终评测流程为：
 - 将提交的代码压缩包解压缩
 - 将评测用的main.cpp，makefile复制到同一目录
 - 编译，运行得到的程序
- 请不要尝试攻击实验室服务器☺



Thanks, Questions?