Lab5 - Hash

1. 问题描述

哈希表有多种实现方法，性能表现也各不相同。你需要实现不同的哈希表，构造测试数据，进行性能测试，并分析测试结果。同时，你还需要处理utf-8编码的字符串（包含中文和其他一些特殊字符），体会对它们做哈希和ascii字符串有何不同。

2. 背景知识——utf-8编码

utf-8编码是一种“变长编码”，即不同字符所占的字节数目可能不同。一个utf-8字符可能是1-6个字节，但1-3个字节的utf-8字符已经覆盖了大部分会用到的字符。我们给出的数据中**只包含1-3个字节的utf-8字符**。

- 长度1字节的utf-8字符，和ascii码相同，且最高位一定是0。长度不为1字节的utf-8字符，最高位一定是1。
- 长度2字节的utf-8字符，第一个字节最高三位一定是110，第二个字节最高两位一定是10。
- 长度3字节的utf-8字符，第一个字节最高四位一定是1110，第二、三个字节的最高两位一定是10。

按照这种编码方式，在将字符按照utf-8的规则转换为一串字节后，我们还可以将这一串字节中的字符区分开来。
utf-8是Mac OS X和linux下的默认编码方式，但并不是windows默认的编码。所以建议在**linux或Mac系统下进行本实验**。

3. 实验步骤

3.1. 编写数据生成器

我们提供了 `poj.txt` 和 `hdu.txt` 两个数据集，是poj.org和acm.hdu.edu.cn两个做题网站上的用户ID、排名和做题量。
你需要根据这两个数据集，编写程序作为数据生成器，以 `poj.txt` 或 `hdu.txt` 为输入时，能够输出格式类似于 `1.in` 的测试数据。**需要保证每条映射关系只进行一次插入操作。**

生成的输入文件格式如下：
每一行一个操作，第一个数字表示操作类型。

- 0表示给出一条需要插入哈希表的“字符串-数字”映射关系
- 1表示查询某个字符串对应的数字（不存在时输出-1）

- 2表示输入结束

你应当编写一个程序，以 `poj.txt` 或 `hdu.txt` 为输入，分别用于生成ascii编码和utf-8编码的测试数据。注意，这些测试数据应当具有适当的性质，可以用来比较之后实现的各种哈希策略。
不同数据规模、不同的插入/查询操作比例、插入和查询的不同分布方式，都可以构造出不同的测试数据。你总共需要构造6组不同的数据，3组来自 `poj.txt`，3组来自 `hdu.txt` 。

3.2. 不同哈希策略的实现

你需要实现不同的哈希策略并在3.1.生成的数据上比较不同哈希策略的性能。具体要求如下：

- 针对ascii编码和utf-8编码分别设计哈希函数
- 设计**至少三种**冲突处理策略，其中**必须包含链地址法和二次探测再散列法(即双向平方试探)**

温馨提示：针对utf-8编码的字符串，请不要将其当成一串字节来处理，而是当成“一串utf-8字符”来处理，从字符串中把一个一个的utf-8字符提取出来， 以utf-8字符为哈希函数处理的基本元素。

3.3. 进行测试

现在你至少完成了有2种哈希函数、X(X>=3)种冲突处理策略，两两组合，共有2X种不同的哈希表和6组测试数据。

然后，你需要进行 2X * 6 次测试，获取运行时间数据，汇总成表格。如果某些程序运行完毕需要的时间过长，不必将其运行完，直接杀死程序，在结果中用“大于Y秒”表示即可（Y是你自己设定的一个门槛）。

3.4. 回答问题

在实验报告中用**简短的语言**回答这些问题，每个问题的回答**不得超过300字**：

- 将utf-8字符串“当作”ascii字符串进行处理，使用针对ascii字符串的哈希函数，实际效果如何（相比“针对utf-8字符串设计的哈希函数）？可能的原因是什么？
- 在你的测试中，不同的冲突处理策略性能如何？可能的原因是什么？
- 设计哈希函数时，我们往往假定字符串每个位置上出现字符集内每个字符的概率都是相等的，但实际的数据集往往并不满足这一点。这可能造成什么影响？
- 对于“字符串到数字映射”问题(给定一组字符串以及它们各自对应的数字，然后多次查询某个字符串对应的数字，可能会在中途更新某个字符串对应的数字)，哈希表并不总是最优方案。请描述一种输入数据，再举出一种哈希表之外的数据结构，对这种数据，这种数据结构能比哈希表更高效地解决“字符串到数字映射”问题。

4. 实验要求

1. 撰写实验报告。在实验报告中，应当包括：1)6组测试数据的构造方法、数据特征；2)哈希函数的实现思路、冲突处理策略的实现方法；3)测试结果表格；4)3.4.中四个问题的回答。其中，对哈希函数的描述可以只用“纯数学公式”，对冲突处理的描述需要结合代码细节。不鼓励大家追求过多的字数和篇幅，将内容简洁地讲述清楚即可。实验报告必须是**pdf格式**。
2. 提供完整的源代码文件和所有测试数据。**代码必须含有注释**。
3. 鼓励大家查阅资料、相互讨论，但**严禁抄袭**。你应当在实验报告的末尾写明你所参考的网页、博文、某位同学的思路等。一旦发现抄袭，本次作业按0分处理。