

Homework 1

Problem 1

- 若跳表中，第k层的key生长到第k+1层的概率为1/2，请计算跳表所需层数的期望及其空间复杂度，给出计算过程。（提示：几何分布取得成功所需次数的期望为1/p，p为停止生长的概率。）

- **Solution**

设有n个元素

存储在第*i*层列表 S_i ，至少一个会出现在 S_i 中，相应的概率为 $P(|S_i| > 0) < n \times 2^{-i} = \frac{n}{2^i}$

因此， S_i 为空的概率为 $P(|S_i| = 0) > 1 - \frac{n}{2^i}$

以第 $i = 3 \times \log n$ ，该层为空的概率为 $P(h < i) = 1 - \frac{1}{n^2}$

一般地， $i = a \times \log n$ 层列表为空的概率 $1 - \frac{1}{n^{a-1}}$ ， $a > 3$ 后概率迅速接近100%。

跳转表的高度h极大可能不会超过 $3 \times \log n$ ，所以h的期望值为 $E(h) = O(\log n)$

则跳表中数据项总数为

$$\sum_{i=0}^k \frac{n}{2^i} = n \sum_{i=0}^k \frac{1}{2^i} < 2n$$

因此，期望空间复杂度为 $O(n)$

Problem 2

- 请实现以下两种对Bloom Filter的使用方案：
 - (i). 同时使用两个相同大小的Bloom Filter进行筛选，两者各自使用两种哈希函数。（第一个Bloom Filter使用哈希1和哈希2，第二个Bloom Filter使用哈希3和哈希4）
 - (ii). 只使用一个Bloom Filter，同时使用(i)中的四个哈希函数，同时哈希表的长度是(i)的两个Bloom Filter的和。

请随机生成大量数据插入两种方案的Bloom Filter中，并比较两者的假阳性率变化曲线（注意在方案(i)中，只有两个Bloom Filter同时报假阳性时才算入）。

请在回答中给出你的**实现方案（包括使用的哈希函数，哈希表的长度，生成数据的方式等）和比较的结果。**

- **Solution**

使用的哈希函数：

hash1

```

int hash1(int num, int size) {
    unsigned int b = 590127;
    unsigned int a = 32934;
    unsigned int hash = 0;
    unsigned int i = 0;

    for (i = 0; i < size; num /= 10, i++) {
        hash = hash * a + num;
        a = a * b;
    }

    return hash % size;
}

```

hash2

9cc8ec58c9a4d22880b551a69fbb9b5

hash3

5dc0f83922942d2b6a95ece401e2aa7

hash4

df60625688c545da50ca51f939faad3

哈希表的长度:

1000

生成数据的方式:

随机生成数据

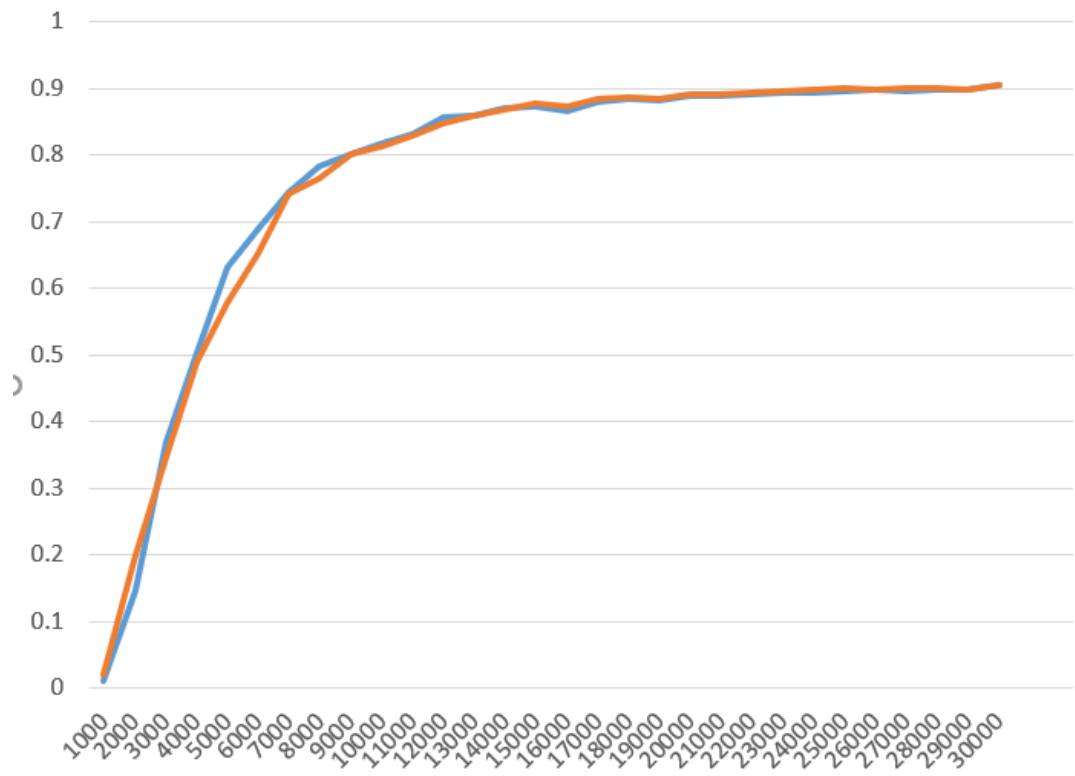
```

//工具函数
default_random_engine e((unsigned)time(NULL));
int randint(int a, int b) { return (e() % b + a); }

```

比较结果:

图表标题



注：系列1 表示(i)方案， 系列2表示(ii)方案

横轴表示生成的测试数据

纵轴表示假阳性率