

# Homework 3: Bloom Filter

在本次作业中，你需要自行实现一个简单的 Bloom Filter，并按照实验步骤探究 Bloom Filter 各项参数与误报率（False Positive）的关系。关于 Bloom Filter 的实现原理、误报率的定义以及最优参数的推导可以参考课程 slides 的内容。

## Part1: 实验步骤

Bloom Filter 的误报率主要与三个变量的值有关，分别是：

- $m$ ：哈希数组的大小
- $n$ ：集合中已经插入的元素个数
- $k$ ：哈希函数的个数

理论上，当取值满足  $k = \ln 2 \cdot (\frac{m}{n})$  时，可以取得最小的误报率（详细推导过程请参考 slides）。

在实验中你需要：

1. 构建大小为  $m$  的哈希数组；
2. 选取  $k$  个取值范围在  $0 \sim m-1$  的哈希函数。你可以自己设计，也可以使用 C++ 提供的标准库 `std::hash`。一般来说，后续的哈希函数  $H_i(x)$  可由第一个哈希函数  $H_1(x)$  简单变化生成，如  $H_i(x) = H_1(x + i)$ ；
3. 待插入的元素个数  $n=100$ ，范围是  $0 \sim 99$ ，你也可自行选择输入集合；
4. 用于测试误报率的测试集合为  $100 \sim 199$ ，你也可自行选择测试集合；
5. 控制  $\frac{m}{n}$  与  $k$  的值分别进行多组实验，记录每组实验的误报率，可以以表格的形式展示你的实验结果，如下图（理论值）所示：

$m/n$	$k$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
2	1.39	0.393	0.400						
3	2.08	0.283	0.237	0.253					
4	2.77	0.221	0.155	0.147	0.160				
5	3.46	0.181	0.109	0.092	0.092	0.101			
6	4.16	0.154	0.0804	0.0609	0.0561	0.0578	0.0638		
7	4.85	0.133	0.0618	0.0423	0.0359	0.0347	0.0364		
8	5.55	0.118	0.0489	0.0306	0.024	0.0217	0.0216	0.0229	

**Tips 1:** 你只需要完成并记录  $2 \leq \frac{m}{n} \leq 5$  且  $1 \leq k \leq 5$  的部分即可，即图中红色部分。

6. 记录完成数据后，请观察规律并分析你的数据，同时观察  $k$  值是否在理论值下误报率最小，如果有差别请简单分析可能的原因。

**Tips 2:** 你也可以自己设计实验步骤和方法，请在实验报告中附上相应的说明。

## Part 2: 提交要求

你提交的内容应该包括：

- 你实现的 Bloom Filter 源码；
- 你的**实验结果数据**以及对实验结果的**简单分析**；
- （可选）你自行选择的输入/测试集合；
- （可选）你自己的实验设计方法和思路；

## Part 3: 注意事项

---

- 请将相关的代码和实验报告打包上传 Canvas，命名使用“学号+姓名+hw3”，如“521123456789+张三+hw3.zip”。
- **请勿抄袭**！课后作业采用倒扣分制，如果有遗漏或者得分不足会在最终成绩酌情减分，同时课后作业的内容会体现在期末试卷中，对同学们也是一种练习。
- 本次作业的截止时间是 **2023年3月12日23:59**，迟交将会酌情扣分。
- 有任何作业相关的问题可以询问 江玗璠、熊天磊 助教。