作业 2 Bloom Filter

1 理论分析

令哈希数组的大小为 m，已插入元素个数为 n，设置 k 个哈希函数。

在插入某个元素时，对于数组中某位，该位为 false 的概率为 $1-1/m$ ，k 次哈希后为 false 的概率为 $(1-1/m)^k$ ，插入 n 个元素后为 false 的概率为 $(1-1/m)^{nk}$ ，即该位为 true 的概率为 $1-(1-1/m)^{nk}$ 。

则有误报率 f 为

$$[1-(1-\frac{1}{m})^{kn}]^k \approx (1-e^{-\frac{kn}{m}})^k$$

易知 n 不变时，误报率随 m 的增大而减小。

此外，上式可化为 $f=e^g$ ，其中

$$p=e^{-\frac{kn}{m}},\ g=-\frac{m}{n}ln(p)ln(1-p)$$

由对称性 p 为 1/2，解得误报率最小时，有

$$k=ln2\cdot(\frac{m}{n})$$

2 实验数据与折线图

令 n=100，m/n 分别为 2、3、4、5，k 分别为 1、2、3、4、5，插入元素为 0 ~ 99, 测试元素为 100 ~ 199，得到的误报率如表 1。

m/n \ k	1	2	3	4	5
2	0.390	0.420			
3	0.260	0.250	0.230		
4	0.180	0.140	0.160	0.140	
5	0.130	0.090	0.090	0.080	0.110

表 1: 实验误报率

k=1 时，误报率-m/n 曲线如图 1。m/n=5 时，误报率-k 曲线如图 2。

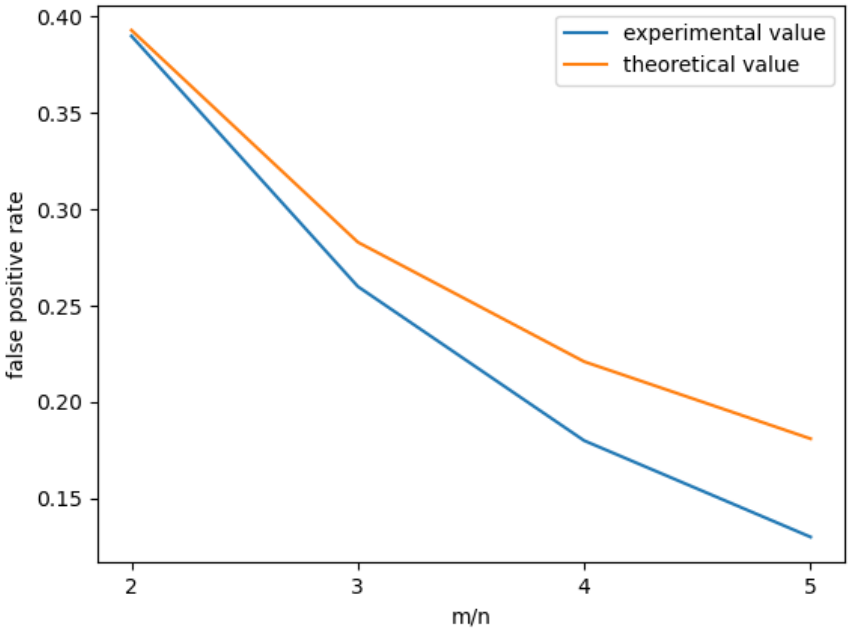

图 1: 误报率-m/n 曲线

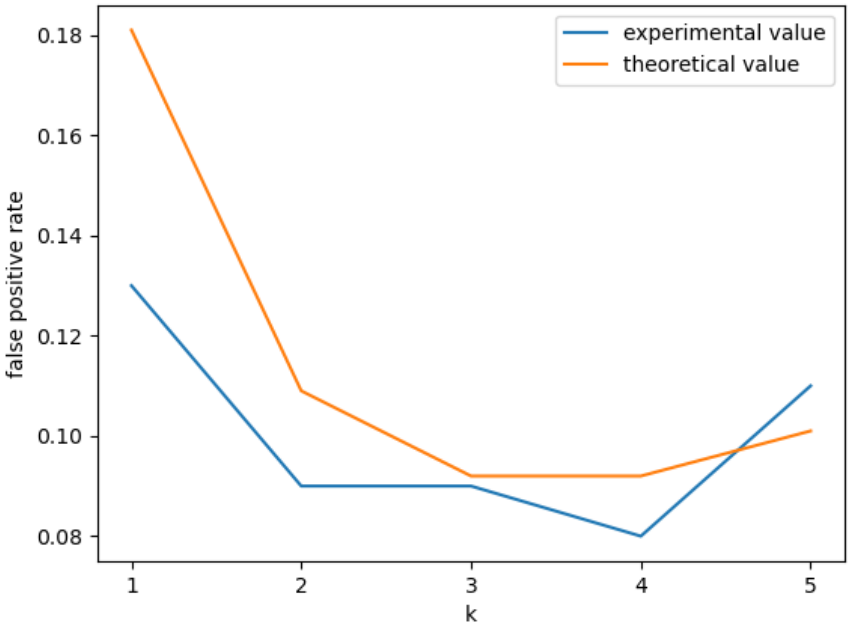

图 2: 误报率-k 曲线

3 实验数据分析

图 1和图 2中，实验值与理论值近似相同。

在 k 固定时，实验得到的误报率随 m 的增大而减小，与理论分析一致。此外，最小化误报率的 k 值与理论 k 值相近，误差可能与 k 的四舍五入等有关。

在 m/n 分别为 2、3、4、5，k 分别为 1、2、3、4、5 时，理论误报率如表 2。

m/n \ k	1	2	3	4	5
2	0.393	0.400			
3	0.283	0.237	0.253		
4	0.221	0.155	0.147	0.160	
5	0.181	0.109	0.092	0.092	0.101

表 2: 理论误报率

将表 1与表 2对照，两张表的数据范围即趋势大致相同。两张表的少量偏差可能与哈希函数的性质等有关。

4 绘图代码

```
import matplotlib.pyplot as plt

def paint_graph(x, x_exp, x_the, title):
    plt.plot(x, x_exp, label='experimental value')
    plt.plot(x, x_the, label='theoretical value')
    plt.legend()
    plt.xlabel(title)
    plt.ylabel('false positive rate')
    plt.xticks(x)
    plt.savefig('./graph/{}.png'.format(title[:1]))
    plt.close()

mn = [2, 3, 4, 5]
mn_exp = [0.39, 0.26, 0.18, 0.13]
mn_the = [0.393, 0.283, 0.221, 0.181]

k = [1, 2, 3, 4, 5]
k_exp = [0.13, 0.09, 0.09, 0.08, 0.11]
k_the = [0.181, 0.109, 0.092, 0.092, 0.101]

paint_graph(mn, mn_exp, mn_the, "m/n")
paint_graph(k, k_exp, k_the, "k")
```