hw29

本次作業需要證明實作 Bloom Filter 以及證明其 false positive 機率，Bloom Filter 需要用到多個雜湊函數來作為濾波器，在這邊我採用 MurMurHash2 作為濾波器，以不同的 seed 做為不同的 hash 來實作。

False positive

因為 Bloom Filter 的特性因此只會出現 FP 但不會有 TN 的錯誤出現，而 FP 證明如下。

$$P(u) = (1-\frac{1}{n})^u(1-(1-\frac{1}{m})^{uh})^h$$

$$\because\ (1-\frac{1}{x})^q \approx e^{-\frac{q}{x}}\ \therefore\ P(u) \approx e^{-\frac{u}{n}}(1-e^{-\frac{uh}{m}})^h$$

$$y = (1-e^{-\frac{uh}{m}})^h,\ s = \frac{u}{m}$$

$$\Rightarrow \ln y = h\ln(1-e^{sh})$$

$$\Rightarrow \frac{d}{dh}\ln y = \frac{d}{dh}(h\ln(1-e^{sh}))$$

$$\Rightarrow \frac{dy}{dh} \times \frac{1}{y} = ln(1-e^{sh}) + \frac{hse^{-sh}}{1-e^{-sh}}$$

$$\Rightarrow \frac{dy}{dh} = (1-e^{-sh})^h \times [\ln(1-e^{sh}) + \frac{hse^{-sh}}{1-e^{-sh}}]$$

$$\Rightarrow \frac{dP(u)}{dh} = e^{-\frac{u}{n}}(1-e^{-\frac{uh}{m}})^h \times [\ln(1-e^{-\frac{uh}{m}}) + \frac{uh}{m} \times \frac{e^{-\frac{uh}{m}}}{1-e^{-\frac{uh}{m}}}] = 0$$

$$\therefore [\ln(1-e^{-\frac{uh}{m}}) + \frac{uh}{m} \times \frac{e^{-\frac{uh}{m}}}{1-e^{-\frac{uh}{m}}}] = 0$$

$$x = \frac{uh}{m} \Rightarrow \ln(1-e^{-x}) + x \times \frac{e^{-x}}{1-e^{-x}} = 0$$

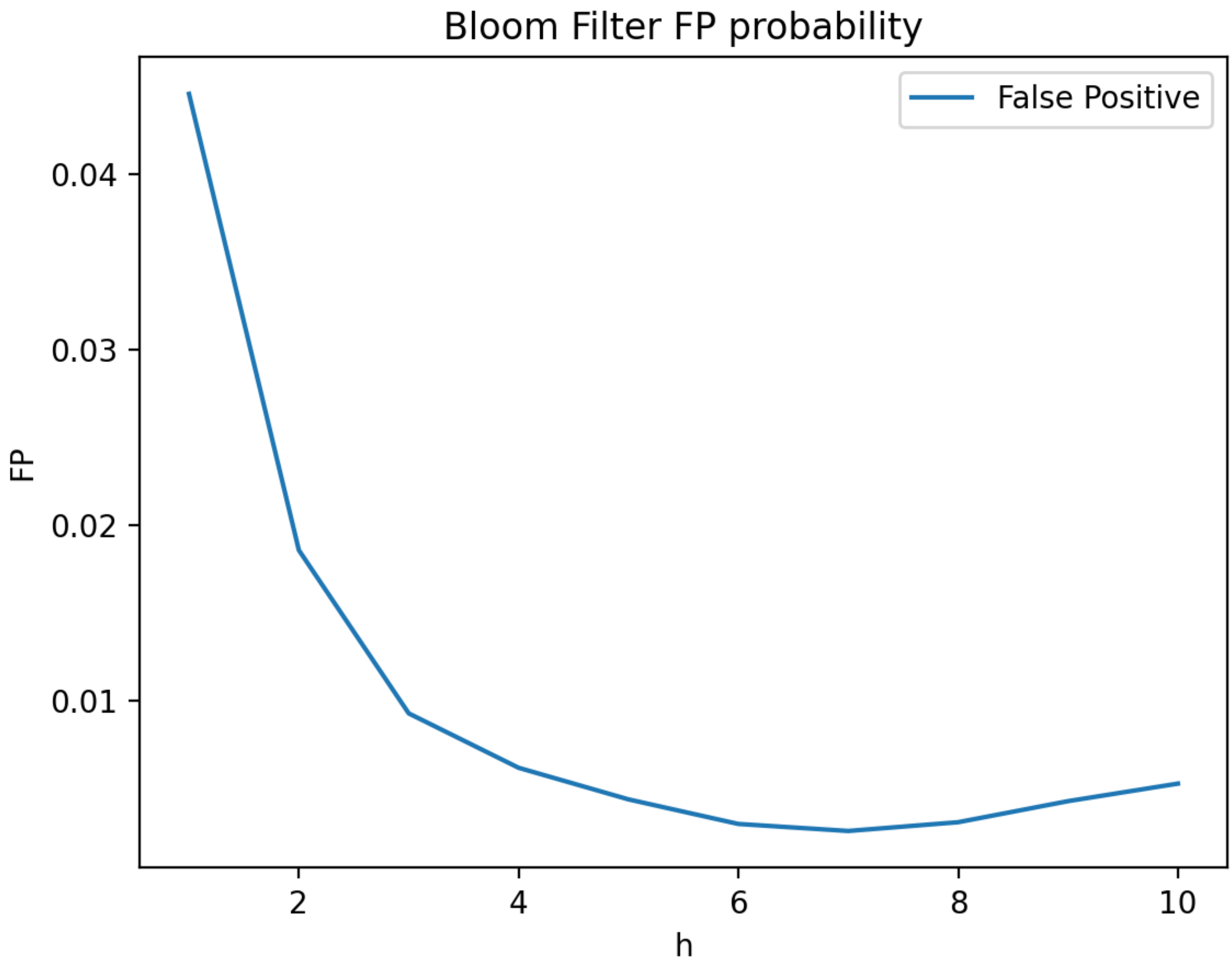
$$z = e^{-x} \Rightarrow \ln(1-z) = \frac{(\ln z) \times z}{1-z}$$

$$(1-z)\ln(1-z) = z(\ln z) \Rightarrow (1-z)^{1-z} = z^z \Rightarrow z = 1-z$$

$$if\ z = \frac{1}{2} \Rightarrow x = \ln 2 \Rightarrow h = \frac{m}{u}\ln 2$$

Test

最後為實際驗證，在這邊我們帶入參數 $n = 10000, m = 100000, u = 10000$，可以得到以下結論，在 $h = 7$ 時會有最小錯誤。



Conclusion

最後的測試結果帶入公式 $h = \frac{100000}{10000} \times \ln 2 \approx 6.9$，我們的實作的確符合理論推導。