Problem 1: Hashing Distribution

To understand the distribution of collisions, let's start by supposing that there are N spaces m elements can be hashed into. Let the conditional expectation E_m represent the expected value of m elements being hashed into unique locations. Let 1 or 0 represent if a location is occupied or not. Then,

$$E_0 = 0$$

$$E_1 = 1 \qquad \left(E_m = \sum x_i P(x_i) \right)$$

$$E_2 = E_1 + 1 * {}^{N-E_1}/N$$

$$E_3 = E_2 + 1 * {}^{N-E_2}/N$$

$$\vdots$$

$$E_m = E_{m-1} + 1 * {}^{N-E_{m-1}}/N$$

We can simplify the above by letting $C = {}^{N-1}/N$.

Then,

$$\left.\begin{array}{l} E_0 = 0 \\ E_1 = 1 \\ E_2 = CE_1 + 1 \\ E_3 = CE_2 + 1 \\ \vdots \\ E_m = CE_{m-1} + 1 \end{array}\right\} \quad E_m = \sum_{i=0}^{m-1} C^i .$$

$$* \sum_{i=0}^{N} C^i = \frac{1 - C^{N+1}}{1-C} .$$

Therefore, $E_m = \dfrac{1-C^m}{1-C} = N(1-C^m)$.

We can make another approximation by using $e^{-1/N} \approx 1 - 1/N$ for C. Then, we get

$$E_m \approx N(1-e^{-m/N}).$$

If we have $N = 512$ and we hash $m = 512$ elements, we should see:

$$N(1-C^m) = 512(1-(511/512)^{512}) \approx 324.$$

This means that we had no collisions for 324 lists or spaces. Consequently, we should have around $512 - 324 = 188$ collisions. This matches with our simulation results.