Probabilistic Counters

ABLORDEPPEY Prosper

***Abstract* –In this project, the goal is to count the number of occurrences of letters in text files and, for instance, identify the most common ones. Three types of counters were analyzed. The *Exact Counter*; which provides the exact count or frequency of each letter present in the text, the *Fixed Probability Counter*, which approximates the frequency or number of counts of each letter in the text using a fixed probability value of $\frac{1}{2}$. The last counting method considered, being the *Decreasing Probability Counter* approximates the count of each letter in the text with each future encounter of the letter having a decreased probability of being counter, with probability $\left(\frac{1}{\sqrt{2}^k}\right)$ where k is the number of occurrence of the letter of interest.**

***Keywords* –**

1. **Let k_n denote the counter value of a given counter for the $n'th$ occurrence of a letter in the text stream.**
2. **Let $\mathbb{E}[k_n]$ denote the Expected Counter Value for the $n'th$ occurrence.**
3. **From a Counter Value k_n, we denote an estimate of the actual/exact number of occurrence for n as $\hat{n}$.**

I. Basic Derivations

Let X_i be a random variable modeling the $i'th$ increment with binary states $S = \{0, 1\}$ where $X_i = 0$ is the event of not increasing the counter value and $X_i = 1$ be the event of increasing the counter value.

Again, let p and q be the probability of increasing and not increasing the counter value respectively. Thus

$$\begin{aligned}\mathbb{P}(X_i = 1) &= p\\ \mathbb{P}(X_i = 0) &= 1 - p = q\end{aligned}$$

Moments

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i \in S} x_i \cdot \mathbb{P}(x = i) \qquad (1)\\ &= 0 \cdot (1-p) + 1 \cdot p = p\\ \therefore \mathbb{E}[X] &= p\\ \mathbb{E}[X^2] &= \sum_{i \in S} x_i^2 \cdot \mathbb{P}(x = i)\\ &= 0^2 \cdot (1-p) + 1^2 \cdot p = p\\ \therefore \mathbb{E}[X^2] &= p\end{aligned}$$

For the $n'th$ occurrence, the counter value is expressed as

$$\begin{aligned}k_n &= \sum_{i=1}^{n} X_i\\ \mathbb{E}[k_n] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\\ &= \sum_{i=1}^{n} \mathbb{E}[X]\\ &= n \cdot \mathbb{E}[X]\\ &= n \cdot p\end{aligned}$$

$$\begin{aligned}\sigma^2[k_n] &= \sum_{i=1}^{n} \sigma^2\left[X\right]\\ &= \sum_{i=1}^{n} \left[\mathbb{E}[X^2] - \mathbb{E}[X]^2\right]\\ &= \sum_{i=1}^{n} \left[p - p^2\right]\\ &= n(p - p^2)\\ &= n \cdot p \cdot = (1-p) = n \cdot p \cdot q\end{aligned}$$

It has been shown from the above that, for the $n'th$ occurrence of a given letter, the expected counter value, variance and standard deviation are expressed respectively as;

$$\mathbb{E}[k_n] = n \cdot p \qquad (1)$$

$$\sigma^2[k_n] = n \cdot p \cdot q \qquad (2)$$

$$\sigma[k_n] = \sqrt{n \cdot p \cdot q} \qquad (3)$$

II. Outline Of Implementation

III. EXACT COUNTER

This counter could be considered as a fixed probability counter, where the probability of increasing the counter value for a given new encounter of a letter is $p = 1$. Anytime a new letter is observed from the text stream, the counter value get increased. Thus, $q = 0$. Hence, the expected value, variance and standard deviation of the counter values is given as

$$\begin{aligned}\mathbb{E}[k_n] &= n \cdot p\\ &= n \cdot 1 = n\\ \sigma^2[k_n] &= n \cdot p \cdot q\\ &= n \cdot 1 \cdot 0 = 0\\ \sigma[k_n] &= \sqrt{n \cdot p \cdot q}\\ &= \sqrt{n \cdot 1 \cdot 0} = 0\end{aligned}$$

The Table (III) below presents the expected counter value for a given occurrence n of a particular letter in a given text stream. The estimated occurrence from the (k_n) counter values given as $(\widehat{n})$ is an identity relation.
More formally,

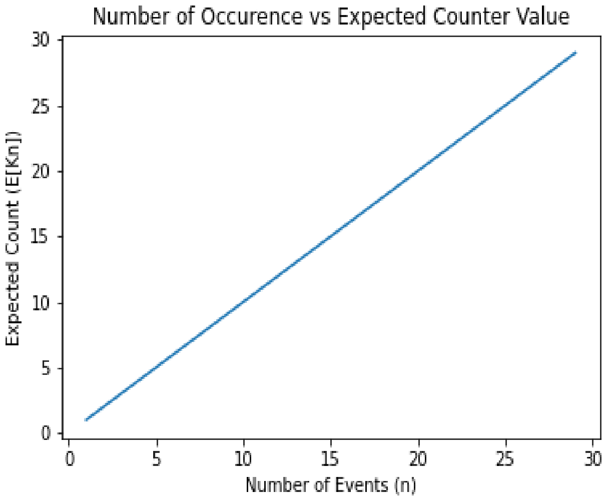
$$\widehat{n} = \mathbb{E}[k_n]$$

Although such a counter will report the accurate/exact number of occurrence of each letter in the text stream, for large text streams, this becomes a problem as it occupies a lot of memory, hence, expensive. This motivates the notion of analyzing approximate counters with fixed probability and decreasing probabilities.

TABLE I
EXACT COUNTER ESTIMATES

Occurrence (n)	Expected Value	$\mathbb{E}[k_n]$	$\hat{n}$
1	1	1	1
3	$1+1+1$	3	3
13	$\mathbb{E}[k_{11}]+1+1$	13	13
27	$\mathbb{E}[k_{25}]+1+1$	27	27
51	$\mathbb{E}[k_{49}]+1+1$	51	51

Fig. 1 - Exact Expected Counter Estimates



IV. FIXED PROBABILITY COUNTER

With this counter, the probability of increasing and decreasing the counter are equal. Thus $p = q = \frac{1}{2}$. Inferring from Equations (1), (2) and (3);

$$\begin{aligned}\mathbb{E}[k_n] &= n \cdot p \\ &= n \cdot \frac{1}{2} = \frac{n}{2} \\ \sigma^2[k_n] &= n \cdot p \cdot q \\ &= n \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{n}{4} \\ \sigma[k_n] &= \sqrt{n \cdot p \cdot q} \\ &= \sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{\sqrt{n}}{2}\end{aligned}$$

An expression for $\widehat{n}$ from the counter value k_n for this counter is expressed as

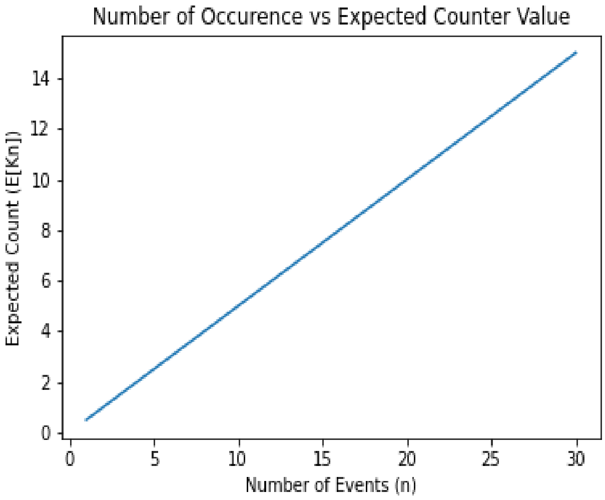
$$\widehat{n} = 2 \cdot \mathbb{E}[k_n]$$

This estimate relation evaluates to the actual number of occurrence (n). The memory allocation for this counter is lesser than that of the exact counter by half $(\frac{1}{2})$.

TABLE II
FIXED PROBABILITY COUNTER ESTIMATES $(p = \frac{1}{2})$

Occurrence (n)	Expected Value	$\mathbb{E}[k_n]$	$\widehat{n}$
2	$\frac{1}{2}+\frac{1}{2}$	1	1
4	$\mathbb{E}[k_2]+\frac{1}{2}+\frac{1}{2}$	2	4
6	$\mathbb{E}[k_4]+\frac{1}{2}+\frac{1}{2}$	3	6
28	$\mathbb{E}[k_{26}]+\frac{1}{2}+\frac{1}{2}$	14	28
50	$\mathbb{E}[k_{48}]+\frac{1}{2}+\frac{1}{2}$	25	50

Fig. 2 - Fixed Prob Expected Counter Estimates



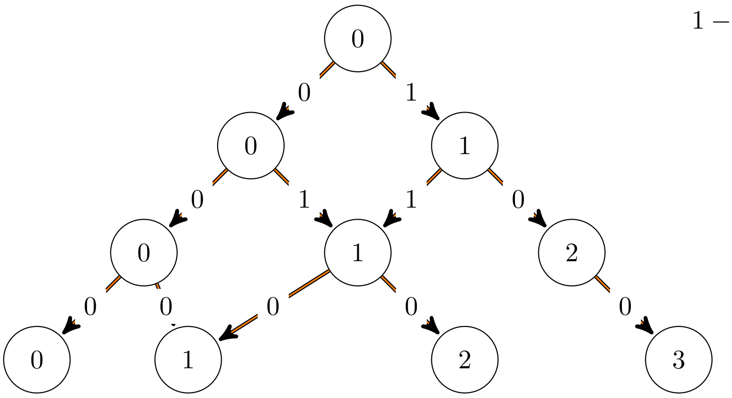


Fig. 3 - Corrected Acyclic Graph

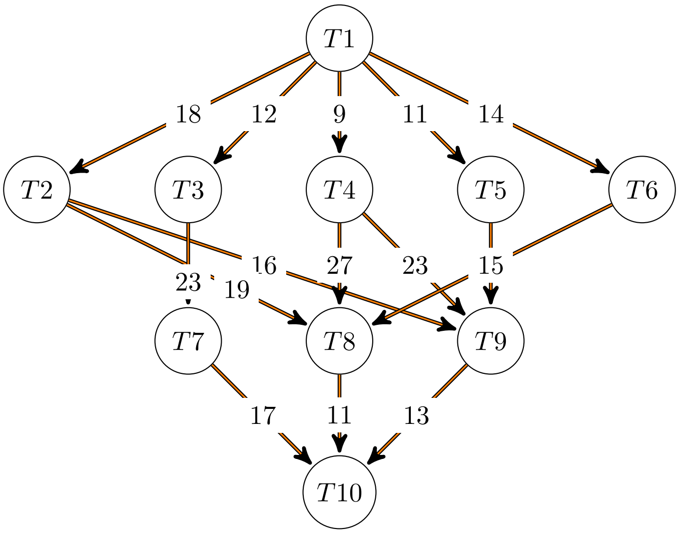
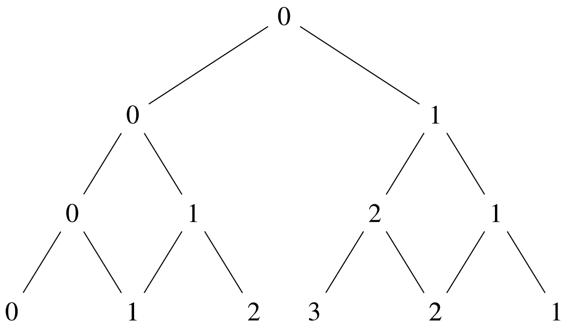


Fig. 4 - Corrected Acyclic Graph

V. DECREASING PROBABILITY COUNTER



The probability of increasing the counter value for a given occurrence in this case is given as $p = \frac{1}{\sqrt{2}^{k_n}}$. Hence, $q = 1 - \frac{1}{\sqrt{2}^{k_n}}$. Inferring from Equations (1), (2) and (3);

$$\begin{aligned}
\mathbb{E}[k_n] &= n \cdot p \\
&= n \cdot \frac{1}{\sqrt{2}^{k_n}} = \frac{n}{\sqrt{2}^{k_n}} \\
\sigma^2[k_n] &= n \cdot p \cdot q \\
&= n \cdot \frac{1}{\sqrt{2}^{k_n}} \cdot \left(1 - \frac{1}{\sqrt{2}^{k_n}}\right) \\
&= \frac{n}{\sqrt{2}^{k_n}} \cdot \left(1 - \frac{1}{\sqrt{2}^{k_n}}\right) \\
\sigma[k_n] &= \sqrt{n \cdot p \cdot q} \\
&= \sqrt{n \cdot \frac{1}{\sqrt{2}^{k_n}} \cdot \left(1 - \frac{1}{\sqrt{2}^{k_n}}\right)} \\
&= \frac{\sqrt{n}}{2}
\end{aligned}$$

The expected counter values for the various occurrences n, could be expressed as

$$\begin{aligned}
\mathbb{E}[k_n] &= \sum_{k=1}^{n} \frac{1}{\left(\sqrt{2}\right)^{k-1}} = \sum_{k=0}^{n-1} \frac{1}{\left(\sqrt{2}\right)^{k}} \\
&= \frac{2^{(1-\frac{n}{2})} - 2}{\sqrt{2} - 2} \\
\therefore \mathbb{E}[k_n] &= \frac{2^{(1-\frac{n}{2})} - 2}{\sqrt{2} - 2}
\end{aligned} \tag{3}$$

An expression for n estimate $\widehat{n}$ from the counter value is given by

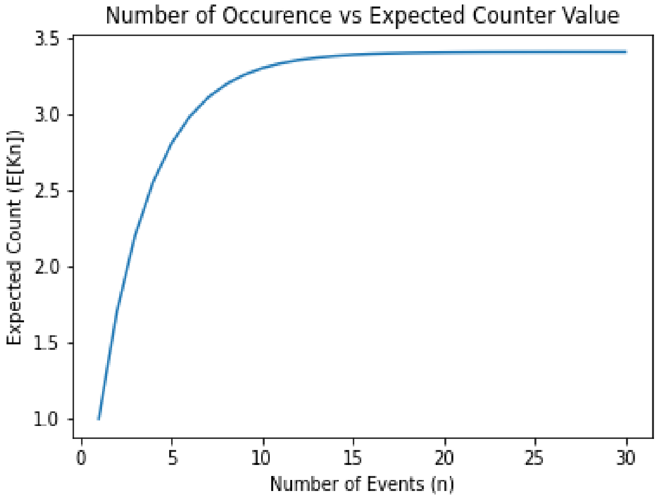
$$\widehat{n} =$$

As it can be clearly observed from Fig. (V), the expected counter values as $n \to \infty$ becomes constant for $n > 10$. Hence, we can determine the upper-bound for the expected counter values by evaluating the limit of the sum function of Equation (3) as $n \to \infty$.

TABLE III
DECREASING PROBABILITY COUNTER ESTIMATES $\left(prob = \frac{1}{(\sqrt{2})^k}\right)$

# of Events	$\mathbb{E}[S]$	Expected Counter V
1	1	1
3	$1 + \frac{1}{\sqrt{2}} + \frac{1}{(\sqrt{2})^2}$	2.2071
13	$\mathbb{E}[S]_{11} + \frac{1}{(\sqrt{2})^{11}} + \frac{1}{(\sqrt{2})^{12}}$	3.3765
27	$\mathbb{E}[S]_{25} + \frac{1}{(\sqrt{2})^{25}} + \frac{1}{(\sqrt{2})^{26}}$	3.4139
51	$\mathbb{E}[S]_{49} + \frac{1}{(\sqrt{2})^{49}} + \frac{1}{(\sqrt{2})^{50}}$	3.4142

Fig. 5 - Decreasing Prob Expected Counter Estimates



Let $\phi(k_n) \geq \mathbb{E}[k_n], \forall n$ be an upper-bound for $\mathbb{E}[k_n]$

$$\begin{aligned}
\phi(k_n) &= \lim_{n\to\infty} \frac{2^{(1-\frac{n}{2})} - 2}{\sqrt{2} - 2} \\
&= \lim_{n\to\infty} \frac{2^{(1-\frac{n}{2})} - 2}{\sqrt{2} - 2} \cdot \frac{\sqrt{2} + 2}{\sqrt{2} + 2} \\
&= \lim_{n\to\infty} 2^{\frac{1}{2}} + 2 - 2^{\frac{1-n}{2}} - 2^{\frac{2-n}{2}} \\
\therefore \phi(k_n) =& \sqrt{2} + 2 \\
\approx& 3.14142
\end{aligned}$$

Hence, after about 10 events, the expected counter values for all future occurrences tend to approach $\sqrt{2} + 2$.

VI. Auxiliary Functions

References

[1] Paul E Black. greedy algorithm, dictionary of algorithms and data structures. *US Nat. Inst. Std. & Tech Report*, 88:95, 2012.

[2] Nykamp DQ. Adjacency matrix definition.

[3] Anthony Kim. Min cut and karger's algorithm : Min cut and karger's algorithm, 2016.