Approximate Couting

Mário Francisco Costa Silva, nMec: 93430

Abstract **–The approximate counting algorithms allow counting a very large number of events using a small amount of memory, in other others, it is possible to use a small counter to keep approximate counts of large numbers. In this article, it is explored the fixed and the decreasing probability counters.**

Keywords **–Approximate, Counting, Probability, Algorithm, Decreased, Fixed**

I. Introduction

The algorithms of approximate counters are very important since they allow estimating exact values, with an expected error that can be controlled [1].

They can be useful in examining large data streams for patterns, such as, in applications of data compression, sight and sound recognition, and other artificial intelligence applications [2].

II. Problem Description

The goal is to count the number of occurrences of letters in text files and also to identify the most common ones. This will be done in three different ways: with an exact counter, a fixed probability counter (1 / 8) and with a decreasing probability counter (1 / $2\hat{k}$).

A fixed probability counter will use the same probability for every event that will decide to either count that event or not.

A decreased probability counter will use a probability different for each letter, this probability will get smaller every time an event is counted for the letter in analysis.

A series of tests will be performed to analyse the computational efficiency and limitations of the developed counters.

III. Approach and Implementation

A. Parameters Management

The program allows the user to pass some arguments that have different actions. The user can set a few parameters that allow for an easier use of the program. It can be passed the name of the file to perform the letter counting, the number of most occurrent letters to display and also the amount of repetitions/trials to perform the counting.

```
Usage: python3 main.py
        -f <File Name for Counting Letters: str>
        -k <Top k Most Occurrent Letters: int>
        -r <Repetitions for Testing: int>
```

Fig. 1
Program arguments

B. Exact Counter

This is a counter with a very simple implementation. All it does is read the given file in chunks, count the letters and store in a dictionary with the exact number of occurrences of each letter. It also removes all non-alphabetical characters and transforms every letter to upper case.

Reading the file chunk by chunk could be crucial since it might be necessary to handle very large files, therefore, it may not have memory to read and store the whole file.

C. Fixed Probability Counter

This counter has a similar implementation to the previous one, except it only counts the letter occurrence if a random generated number by *python* is smaller or equal to 1 / 8.

To estimate the events after the counting, it can be simply done by multiplying the occurrences of the letter by the inverse of the probability, in this case, 8.

D. Decreasing Probability Counter

For the decreasing probability counter, each letter will have different probabilities of being counted, as the more occurrences it has, the less probability it will have. The probability used was 1 / 2 $\hat{k}$, where k is the number of occurrences of a letter.

In order to speed this process, it was kept another dictionary with the probability of each letter already calculated, this increased the performance by a big margin, because after a while the probability of counting a letter becomes so small it is rarely even counted, and to count each time the same probability takes way more time than having it only counted once every time it's value changes, that is when it is effectively counted. Since this problem is mainly focused in decreasing memory usage, this optimization might not be viable, but was used to speed up the testing process.

After the counting, the following formula was used to estimate the events of each letter:

```
int(a ** k - 1)
```

In this case, the a represents the base used, that was 2, and k the value of the counter of a given letter.

E. Results

For the testing part, it was used the text file of the Bible, for these firsts tests, it was always used the English version.

As it was mentioned previously, the program allows setting the number of trials to execute and also choosing the number of the most frequent letters to display.

F. Results for 1 Trial

For a first analysis, it was performed a test with a single repetition and chosen to display the top 10 most frequent letters.

```
Exact Counter

Results for 1 repetition:
Total Elapsed Time: 1.043 s
Total Events Counted: 3218643

Average Values for a Repetition:
Measure                    Value
-----------------  ----------------
Counting Time (s)          1.043
Alphabet Size             26
Events                     3.21864e+06
Mean                  123794
Minimum                 1234
Maximum               400817

Top 10 Most Frequent Letters:
Letter      Exact Events
--------  --------------
E                400817
T                284578
O                279886
A                267258
H                241710
N                216071
I                208186
S                185532
R                174573
D                141033
```

Fig. 2
Exact Counter Results for 1 Repetition

As it can be seen, it took slightly over a second to perform the test, and counted more than 3 million events. It also displays the alphabet size, which was the full English alphabet, 26, the mean, minimum, and maximum number of events for all letters, and then the top 10 most frequent letter.

These results are stored to be used for comparison with the next counters.

For the next two approximate counters, other statistics are calculated.

As it can be seen from the figures 3 and 4, the execution time of the fixed probability counter was a bit smaller than the decreasing one, this can be explained due to the fact that the decreasing probability counter has to calculated new probabilities for each letter. However, both of them take less time than the exact counter because of less memory accesses are done to increment the counters. As it can be seen, the fixed probability counted less a million events, and the decreasing one, only counted 402 events. The difference is huge when compared to the 3 million of the exact counter.

The alphabet size was the same in all counters, which means, every counter managed to count every letter at least once. The total and mean number of events has a smaller error in the fixed probability one, this is a bit obvious since it counted a lot more events.

```
Fixed Probability Counter with 1 / 8

Results for 1 repetition:
Total Elapsed Time: 0.649 s
Total Events Counted: 403682.0

Average Values for a Repetition:
Measure                    Value  Absolute Error    Relative Error (%)
-----------------  ----------------  ----------------  --------------------
Counting Time (s)          0.649     -                 -
Alphabet Size             26         0.0               0.0
Events                     3.22946e+06  10813.0           0.34
Mean                  124210           415.89            0.34
Minimum                 1144            90.0              7.29
Maximum               402584           1767.0            0.44

Top 10 Most Frequent Letters:
Letter       Min     Max    Mean    Mean Absolute Error    Mean Relative Error (%)
--------  ------  ------  ------  ---------------------  -------------------------
E         402584  402584  402584                   1767                       0.44
T         286784  286784  286784                   2206                       0.78
O         279688  279688  279688                    198                       0.07
A         266584  266584  266584                    674                       0.25
N         216912  216912  216912                    841                       0.39
I         207368  207368  207368                    818                       0.39
S         186656  186656  186656                   1124                       0.61
R         176280  176280  176280                   1707                       0.98
D         143440  143440  143440                   2407                       1.71
Accuracy: 100.00 %
Precision: 100.00 %
Average Precision (relative order): 100.00 %
```

Fig. 3
Fixed Probability Counter Results for 1 Repetition

```
Decreasing Probability Counter with 1 / 2^k

Results for 1 repetition:
Total Elapsed Time: 0.793 s
Total Events Counted: 402.0

Average Values for a Repetition:
Measure                    Value  Absolute Error    Relative Error (%)
-----------------  ----------------  ----------------  --------------------
Counting Time (s)          0.793     -                 -
Alphabet Size             26         0.0               0.0
Events                     3.30493e+06  86291.0           2.68
Mean                  127113           3318.89           2.68
Minimum                  511            723.0             58.59
Maximum               524287           123470.0          30.8

Top 10 Most Frequent Letters:
Letter       Min     Max    Mean    Mean Absolute Error    Mean Relative Error (%)
--------  ------  ------  ------  ---------------------  -------------------------
T         524287  524287  524287                 239709                      84.23
E         524287  524287  524287                 123470                      30.8
H         262143  262143  262143                  20433                       8.45
R         262143  262143  262143                  87570                      50.16
S         262143  262143  262143                  76611                      41.29
O         262143  262143  262143                  17743                       6.34
L         262143  262143  262143                 146243                     126.18
A         131071  131071  131071                 136187                      50.96
I         131071  131071  131071                  77115                      37.04
N         131071  131071  131071                  85000                      39.34
Accuracy: 92.31 %
Precision: 90.00 %
Average Precision (relative order): 0.00 %
```

Fig. 4
Decreasing Probability Counter Results for 1 Repetition

From the top 10 most frequent letters, it is also shown the error values compared to the exact counter. Overall, all errors are much smaller on the fixed probability one than on the decreasing one, which is to be expected since the total events counted is much smaller. It also shows the accuracy, precision, and the average precision considering relative order of the top letters displayed. The fixed probability counter had the same order as the exact counter, but the decreasing one, although it had high accuracy and precision, it didn't have any letter in the correct order.

G. Results for 100 Trials

In order to better visualize the differences, these algorithms will be executed multiple times to better understand the results, in this case, 100 trials.

For the exact counter, it can be seen the total elapsed time, the number of counted events, and the average time for 1 repetition, all the other results are obviously the same as with 1 repetition, so they are not displayed here, but can be consulted on the figure 2.

```
Exact Counter

Results for 100 repetitions:
Total Elapsed Time: 93.004 s
Total Events Counted: 3218643

Average Values for a Repetition:
Measure                        Value
-----------------  ----------------
Counting Time (s)       0.93
```

Fig. 5
Exact Counter Results for 100 Repetitions

Visualizing the figures 6 and 7, the same conclusions regarding time elapsed, events counted, and alphabet size can be made from these results as the previous one with only 1 repetition.

```
Fixed Probability Counter with 1 / 8

Results for 100 repetitions:
Total Elapsed Time: 54.592 s
Total Events Counted: 402230.96

Average Values for a Repetition:
Measure                 Value  Absolute Error    Relative Error (%)
----------------  ------------  ----------------  --------------------
Counting Time (s)       0.546  -                 -
Alphabet Size              26  0.0               0.0
Events                3.21785e+06  795.32            0.02
Mean                  123763       30.59             0.02
Minimum                 1241.44    7.44              0.6
Maximum               400748       68.84             0.02

Top 10 Most Frequent Letters:
Letter       Min     Max    Mean    Mean Absolute Error    Mean Relative Error (%)          Variance    Standard Deviation
--------  ------  ------  ------  ---------------------  -------------------------  --------------  --------------------
E         396368  403936  400748                  68.84                       0.02       2.83761e+06               1684.52
T         281808  287576  284653                  75.2                        0.03       1.63044e+06               1276.88
O         275656  280096  279576                 310.32                       0.11       2.38577e+06               1544.59
A         263928  269816  267233                  24.96                       0.01       1.64979e+06               1284.44
H         236112  244392  241454                 255.92                       0.11       1.79761e+06               1340.75
N         212904  219280  215953                 118.04                       0.05       1.53058e+06               1237.17
I         203984  210976  208120                  66.32                       0.03       1.47738e+06               1215.47
S         183280  187576  185579                  47.2                        0.03  956088                          977.798
R         171136  177280  174455                 117.8                        0.07       1.36985e+06               1170.41
D         138344  143440  140882                 151.08                       0.11       1.11304e+06               1055.01
Accuracy: 100.00 %
Precision: 100.00 %
Average Precision (relative order): 100.00 %
```

Fig. 6
Fixed Probability Counter Results for 100 Repetitions

```
Decreasing Probability Counter with 1 / 2^k

Results for 100 repetitions:
Total Elapsed Time: 84.678 s
Total Events Counted: 407.6

Average Values for a Repetition:
Measure                 Value  Absolute Error    Relative Error (%)
----------------  ------------  ----------------  --------------------
Counting Time (s)       0.847  -                 -
Alphabet Size              26  0.0               0.0
Events                3.19325e+06  25394.12          0.79
Mean                  122817       976.7             0.79
Minimum                 861.72     372.28            30.17
Maximum               574094       173277.36         43.23

Top 10 Most Frequent Letters:
Letter       Min      Max     Mean    Mean Absolute Error    Mean Relative Error (%)     Variance    Standard Deviation
--------  ------  -------  ------  ---------------------  -------------------------  -------------  --------------------
E         131071  1048575  367001                33816.4                       8.44  4.43241e+10                210533
A          65535  1048575  294911                27653                        10.35  4.66863e+10                216070
O          65535  1048575  285081                 5194.6                       1.86  5.27744e+10                229727
T          65535  1048575  280493                 4084.92                      1.44  3.28204e+10                181164
H          65535   524287  227409                14301.1                       5.92  1.73988e+10                131874
N          65535  1048575  218889                 2818.24                      1.3   2.26104e+10                150368
R          32767  1048575  199884                25310.8                      14.5   2.50933e+10                158409
I          65535   524287  195952                12234.4                       5.88  1.7566e+10                 132537
S          32767   524287  163184                22348.4                      12.05  9.6203e+09                  98083.1
D          32767   524287  145817                 4783.6                       3.39  1.13468e+10                106521
Accuracy: 100.00 %
Precision: 100.00 %
Average Precision (relative order): 34.33 %
```

Fig. 7
Decreasing Probability Counter Results for 100 Repetitions

However, averaging all repetitions and getting the of number of estimated events, the mean, minimum and maximum, much smaller errors are obtained. The minimum and maximum values have much higher relative errors on the decreasing probability counter comparing to the fixed one. This can be easily explained, letters with little occurrences will be harder to estimate using a decreased probability because the probability grows smaller than the fixed one really fast, so it might end up skipping a lot of values. A similar explanation for the maximum can also be made, if by chance the letter that occurs many times is counted even with a minimal probability, it will estimate a much larger number of events than what it really happens.

Considering the top most frequent letters, it also displays the minimum, maximum, the mean, the mean absolute and relative errors, the variance and the standard deviation of the registered values. These values help visualize the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the all the estimated counters for that letter, while a high standard deviation indicates that the values are spread out over a wider range [3].

It can be observed that the minimum and maximum values do not deviate too much from the mean on the fixed probability counter, but in the decreasing one, the deviations are much higher, for example on the letter E, has the minimum value of 130 071 and the maximum of over a million, but the mean is 367 001, which shows that it can have a big variance in some repetitions, having a standard deviation of 210 533. On the fixed probability counter, they are around 1 000, which is significantly smaller. The accuracy and precision of the top letters was also higher than with just 1 repetition for the decreasing probability counter, and also the average precision considering the relative order increased from 0% to 34.33%.

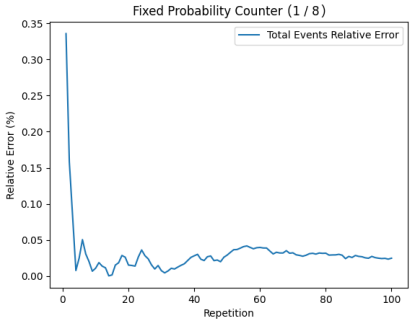

Fig. 8
Fixed Probability Counter Relative Error of the Average Total Estimated Events for 100 Repetitions

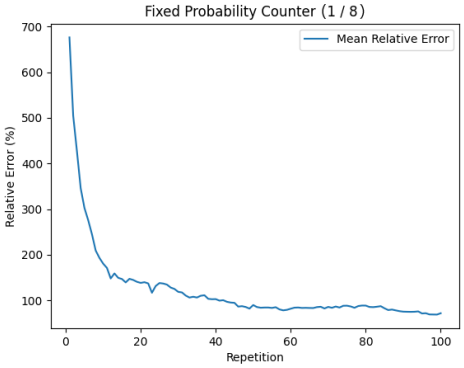

Fig. 9
Fixed Probability Counter Mean Relative Error of all the Estimated Events for 100 Repetitions

Regarding the fixed probability counters, the graphics above, 8 and 9, show, respectively, the estimated total events relative error, and the average of the approximations relative errors, this is, the mean error of the estimated occurrences of each letter comparing to the exact counter value. Both of these two graphics, tend to stabilize after approximately 60 repetitions, this demonstrates that if the problem is only memory related, it can be obtained approximations with minimal errors by averaging the results of 60 repetitions, however, the elapsed time would be much higher.

Comparing the previous graphs with the ones for the decreased probability counter, 10 and 11, it can be observed they also tend to stabilize after 60 repetitions, however, they stabilize on much higher error values and have much higher fluctuations, which means the graph doesn't look as linear as the previous.

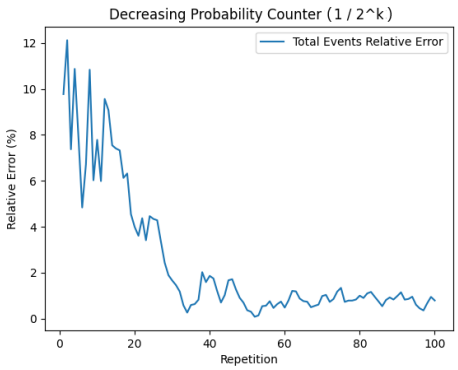

Fig. 10
Decreased Probability Counter Relative Error of the Average Total Estimated Events for 100 Repetitions

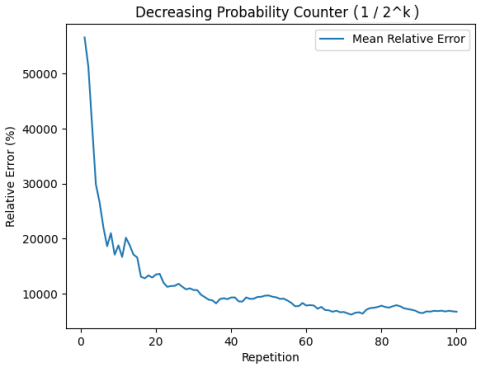

Fig. 11
Decreased Probability Counter Mean Relative Error of all the Estimated Events for 100 Repetitions

H. Different Languages Analysis

For the previous tests, the English version of the Bible was the one used, but it is also interesting to test the same book but in different languages. So it was also used in the German and Portuguese versions. These two languages have different alphabets, as it can be seen from the figures 2 and 12, the alphabet size of the English one has 26 letters, the German has 30 letters and the Portuguese 42, a big reason for that difference is the accents, since English does not have any, and also, German alphabet has other letters, for example, 'ß', and the Portuguese as well, such as 'Ç'.

```
Exact Counter

Results for 1 repetition:
Total Elapsed Time: 0.915 s
Total Events Counted: 3217235

Average Values for a Repetition:
Measure                    Value
-----------------  ---------------
Counting Time (s)          0.915
Alphabet Size             30
Events                     3.21724e+06
Mean                  107241
Minimum                   42
Maximum               534434

Top 10 Most Frequent Letters:
Letter      Exact Events
--------  --------------
E                 534434
N                 338314
I                 245716
R                 232890
D                 219780
S                 202860
A                 190211
H                 177934
T                 171036
U                 136012
```

```
Exact Counter

Results for 1 repetition:
Total Elapsed Time: 0.796 s
Total Events Counted: 2985343

Average Values for a Repetition:
Measure                    Value
-----------------  ---------------
Counting Time (s)          0.796
Alphabet Size             42
Events                     2.98534e+06
Mean                   71079.6
Minimum                    1
Maximum               406759

Top 10 Most Frequent Letters:
Letter      Exact Events
--------  --------------
E                 406759
A                 348622
O                 332322
S                 282939
R                 198395
I                 161611
D                 153589
U                 136263
M                 132984
N                 132549
```

Fig. 12
Exact Counters for German Bible (left) and Portuguese Bible (Right)

However, they still have similar letters, and the most frequent ones tend to be the vowels 'A', 'E', 'I', but 'O' in English and the 'U' in German do not show in the top 10. Also, consonants like 'N', 'S' and 'R' are very frequent in these 3 languages.

IV. Conclusion

Overall, it was tested two different approximate counters, in order to analyse their computational efficiency and evaluate the approximations' errors.

From the results, it can be concluded that it is possible to count very large number of events while having an amount of memory and estimations' errors that can be controlled by adjusting the probabilities used or/and the algorithms used.

References

[1] R. Morris, Counting Large Numbers of Events in Small Registers, Commun. ACM, Vol. 21, N. 10, October 1978

[2] Wikipedia's contributors. (2021a, March 22). Approximate counting algorithm. Wikipedia. https://en.wikipedia.org/wiki/Approximate_counting_algorithm

[3] Wikipedia's contributors. (2021b, November 23). Standard deviation. Wikipedia. https://en.wikipedia.org/wiki/Standard_deviation