Most Frequent Letters

Renato Alexandre Lourenço Dias, nMec 98380

***Abstract* – Counting the most frequent letters can be done with approximate counting algorithms which allow counting a very large number of events using a small amount of memory. In this article, it is explored the fixed probability counters and the lossy-Count.**

I. INTRODUCTION

Letter frequency is the number of times letters of the alphabet appear on average in written language. Letter frequency analysis dates back to the Arab mathematician Al-Kindi, who formally developed the method to break ciphers.

The use of letter frequencies and frequency analysis plays a fundamental role in cryptograms and several word puzzle games, including Hangman, Scrabble, Wordle and the television game show Wheel of Fortune [1].

II. PROBLEM DESCRIPTION

The goal is to identify the most frequent letters in text files. This will be done in three different ways: with an exact counter, a fixed probability counter (1 / 32) and with a lossy-count algorithm.

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

In a lossy-count algorithm the frequency count exceed a user-given threshold [2].

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, the amount of repetitions/trials to perform the counting and also the epsilon value that will be used in the lossy counter.

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

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 1/32.

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, 32.

D. Lossy-Count

The counter works by dividing the Data Stream into ‘Buckets’ as for frequent items, but fill as many buckets as possible in main memory one time. The frequency computed by this algorithm is not always accurate, but has an error threshold that can be specified by the user. The run time space required by the algorithm is inversely

proportional to the specified error threshold, hence larger the error, the smaller the footprint.

The algorithm can be divided in three steps:

- Divide the incoming data stream into buckets of width w = 1 / ϵ, where ϵ is mentioned by user as the error bound (along with minimum support threshold = σ);
- Increment the frequency count of each item according to the new bucket values. After each bucket, decrement all counters by 1;
- Repeat – Update counters and after each bucket, decrement all counters by 1.

E. Results

For the testing part, it was used the text file of the book Association Football, and How To Play It, for the first tests, and also the text file of the Bible in Portuguese and German for different language test.

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: 0.048 s
Total Events Counted: 113981

Average Values for a Repetition:
Measure                Value
-----------------  ----------
Counting Time (s)       0.048
Alphabet Size              27
Events                 113981
Mean                  4221.52
Minimum                     1
Maximum                 14126

Top 10 Most Frequent Letters:
Letter      Exact Events
--------  --------------
E                  14126
T                  10922
A                   9496
O                   8838
I                   7745
N                   7685
S                   6909
R                   6648
H                   6313
L                   5227
```

Fig. 2
Exact Counter Results for 1 Repetition

As it can be seen, it took less than a second to perform the test, and counted more than 100 thousand events. It also displays the alphabet size, which was the full English alphabet (26) plus the "&", 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 approximate counter, other statistics are calculated.

As it can be seen from the figure 3, the execution time of the fixed probability counter was a bit smaller than the exact counter, this can be explained due to the fact that the probability choosed for the fixed probability counter was too low. If the probability choosen was higher, probably exact counter's execution time would be lower than fixed probability counter's execution time because they can retrieve counts in constant time, whereas fixed probability counters have a variable time complexity depending on the probability of success choosen. As it can be seen, the fixed probability counted less than four thousand events. The difference is huge when compared to the 100 thousand of the exact counter.

The alphabet size was not the same in all counters, which means the fixed probability counter does not count the "&".

```
Fixed Probability Counter with 1/32

Results for 1 repetition:
Total Elapsed Time: 0.029 s
Total Events Counted: 3443.0

Average Values for a Repetition:
Measure                Value  Absolute Error    Relative Error (%)
-----------------  ----------  ----------------  --------------------
Counting Time (s)       0.029  -                 -
Alphabet Size              26  1.0               3.7
Events                 110176  3805.0            3.34
Mean                  4237.54  16.02             0.38
Minimum                    32  31.0              3100.0
Maximum                 13728  398.0             2.82

Top 10 Most Frequent Letters:
Letter      Min    Max    Mean    Mean Absolute Error    Mean Relative Error (%)
--------  -----  -----  ------  ---------------------  -------------------------
E         13728  13728   13728                    398                       2.82
T         10656  10656   10656                    266                       2.44
A          9056   9056    9056                    440                       4.63
O          8608   8608    8608                    230                       2.6
I          7712   7712    7712                     33                       0.43
S          7264   7264    7264                    355                       5.14
R          6880   6880    6880                    232                       3.49
N          6624   6624    6624                   1061                      13.81
H          6016   6016    6016                    297                       4.7
L          5280   5280    5280                     53                       1.01
Accuracy: 100.00 %
Precision: 100.00 %
Average Precision (relative order): 63.67 %
```

Fig. 3
Fixed 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 and the accuracy, precision, and the average precision considering relative order of the top letters displayed. As we can see, the fixed probability counter had the same order as the exact counter.

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 as we can see in figure 2.

```
Exact Counter

Results for 100 repetitions:
Total Elapsed Time: 1.839 s
Total Events Counted: 113981

Average Values for a Repetition:
Measure                 Value
----------------  ----------
Counting Time (s)       0.018
Alphabet Size          27
Events                113981
Mean                    4221.52
Minimum                    1
Maximum                14126

Top 10 Most Frequent Letters:
Letter       Exact Events
--------  --------------
E                   14126
T                   10922
A                    9496
O                    8838
I                    7745
N                    7685
S                    6909
R                    6648
H                    6313
L                    5227
```

Fig. 4
Exact Counter Results for 100 Repetitions

Visualizing the figure 5, 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 1/32

Results for 100 repetitions:
Total Elapsed Time: 1.127 s
Total Events Counted: 3562.0

Average Values for a Repetition:
Measure                 Value  Absolute Error    Relative Error (%)
----------------  ----------  ----------------  --------------------
Counting Time (s)       0.011  -                 -
Alphabet Size          25.57   1.43              5.3
Events                113984   3.0               0.0
Mean                    4459.72  238.2             5.64
Minimum                   64.96  63.96             6396.0
Maximum                14152.3   26.32             0.19

Top 10 Most Frequent Letters:
Letter      Min    Max      Mean    Mean Absolute Error    Mean Relative Error (%)    Variance    Standard Deviation
--------  -----  -----  --------  ---------------------  -------------------------  ----------  --------------------
E         12480  15680  14152.3                  26.32                       0.19      521085               721.862
T          9344  13088  10780.5                 141.52                       1.3       410438               640.654
A          7680  11168   9520.32                 24.32                       0.26      358185               598.485
O          7744  10048   8872.64                 34.64                       0.39      234863               483.801
I          6656   9088   7760.96                 15.96                       0.21      248452               498.45
N          6272   9184   7673.6                  11.4                        0.15      218337               467.266
S          5440   8032   6894.4                  14.6                        0.21      227090               476.54
R          5696   7776   6680.96                 32.96                       0.5       191008               437.045
H          5248   7520   6316.48                  3.48                       0.06      190011               435.903
L          4320   6336   5262.72                 35.72                       0.68      182752               427.495
Accuracy: 100.00 %
Precision: 100.00 %
Average Precision (relative order): 100.00 %
```

Fig. 5
Fixed Probability Counter Results for 100 Repetitions

However, averaging all repetitions and getting the number of estimated events, the mean, minimum and maximum, much smaller errors are obtained.

The minimum value has a much higher relative error and maximum value has a lower relative error for 100 trials comparing with the values for 1 trial to the fixed one.

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.

It can be observed that the minimum and maximum values do not deviate too much from the mean on the fixed probability counter (standard deviation values are between 400 and 750).

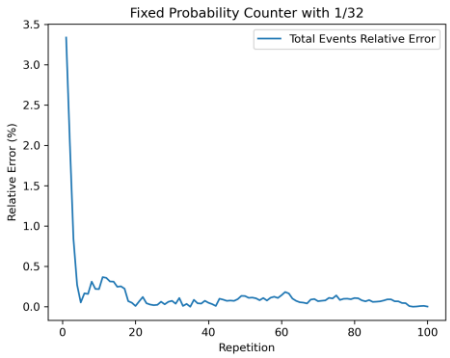

Fig. 6
Fixed Probability Counter Relative Error of the Average Total Estimated Events for 100 Repetitions

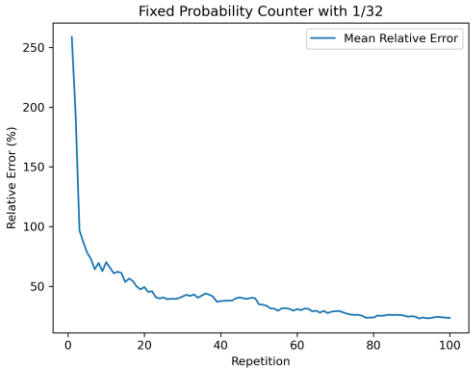

Fig. 7
Fixed Probability Counter Mean Relative Error of all the Estimated Events for 100 Repetitions

Regarding the fixed probability counters, the graphics above, 6 and 7, 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.

H. Lossy Count Results

In figure 8, we can see the results for lossy count algorithm.

```
Lossy Counter

E                   -- Exact: 14126 | Lossy: 14126 -- Acc: 100.0
T                   -- Exact: 10922 | Lossy: 10922 -- Acc: 100.0
A                   -- Exact: 9496  | Lossy: 9496  -- Acc: 100.0
O                   -- Exact: 8838  | Lossy: 8837  -- Acc: 99.99
I                   -- Exact: 7745  | Lossy: 7745  -- Acc: 100.0
N                   -- Exact: 7685  | Lossy: 7684  -- Acc: 99.99
S                   -- Exact: 6909  | Lossy: 6909  -- Acc: 100.0
R                   -- Exact: 6648  | Lossy: 6648  -- Acc: 100.0
H                   -- Exact: 6313  | Lossy: 6312  -- Acc: 99.98
L                   -- Exact: 5227  | Lossy: 5227  -- Acc: 100.0
D                   -- Exact: 3954  | Lossy: 3954  -- Acc: 100.0
C                   -- Exact: 3206  | Lossy: 3205  -- Acc: 99.97
F                   -- Exact: 3027  | Lossy: 3027  -- Acc: 100.0
U                   -- Exact: 2987  | Lossy: 2986  -- Acc: 99.97
P                   -- Exact: 2590  | Lossy: 2589  -- Acc: 99.96
M                   -- Exact: 2516  | Lossy: 2516  -- Acc: 100.0
Y                   -- Exact: 2505  | Lossy: 2505  -- Acc: 100.0
G                   -- Exact: 2430  | Lossy: 2430  -- Acc: 100.0
B                   -- Exact: 2159  | Lossy: 2158  -- Acc: 99.95
W                   -- Exact: 2158  | Lossy: 2158  -- Acc: 100.0
K                   -- Exact: 1028  | Lossy: 1028  -- Acc: 100.0
V                   -- Exact: 980   | Lossy: 978   -- Acc: 99.8
J                   -- Exact: 196   | Lossy: 83    -- Acc: 42.35
```

Fig. 8
Lossy Counter results for 1 repetition

We obtain this values by dividing the number of events calculated with the exact counter algorithm by the number of events calculated with the lossy count algorithm for each letter, which gave us, approximately, an accuracy of 100% for all the letters except for “J”.

I. Different Languages Analysis

It can be also interesting to test the same book but in different languages. So it was used the text files of the Bible in the German and Portuguese versions. These two languages have different alphabets, as it can be seen from the figure 9, the alphabet size of the the German has 30 letters and the Portuguese 42, a big reason for that difference is the accents, and also, German alphabet has other letters, for example, ’ß’, and the Portuguese as well, such as ’Ç’. Comparing with the english version, the alphabet size would be, probably, lower because English does not use accents.

```
Exact Counter

Results for 1 repetition:
Total Elapsed Time: 0.566 s
Total Events Counted: 3217235

Average Values for a Repetition:
Measure                        Value
-----------------  ----------------
Counting Time (s)              0.566
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: 0.606 s
Total Events Counted: 2985343

Average Values for a Repetition:
Measure                        Value
-----------------  ----------------
Counting Time (s)              0.606
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. 9
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’, ’O’ and ’U’. Also, consonants like ’N’, ’S’ and ’R’ are very frequent in these 2 languages.

IV. CONCLUSION

Overall, it was tested three different 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] Wikipedia's contributors. (2022, December 25). Letter Frequency. Wikipedia. https://en.wikipedia.org/wiki/Letter_frequency

[2] Wikipedia's contributors. (2021, April 7). Lossy Count Algorithm Wikipedia. https://en.wikipedia.org/wiki/Lossy_Count_Algorithm