3th Project - CSURÖS Counter and Space-Saving Count

Nuno cunha - 98124

Abstract **–CSURÖS approximate counter is a mathematical algorithm that is used to approximate the number of elements in a set or a stream of data. It is based on the concept of random sampling and is particularly useful when dealing with large datasets where it is not practical to count the number of elements in the set. The CSURÖS approximate counter is designed to provide a fast and accurate estimate of the number of elements in a set, with a low error rate. It has been widely used in various fields including computer science, data analysis, and statistics.**

Space-Saving Count is a mathematical method designed to reduce the amount of storage space needed to represent large numbers in a computer system. This method involves using a smaller number of bits to represent each digit in the number, resulting in a more compact representation that requires less memory to store. The Space-Saving Count method is particularly useful for situations where memory is limited or where it is necessary to store large numbers quickly and efficiently. This method has been used in various applications including data compression, image processing, and scientific simulations.

In this paper, we will compare the accuracy of the results and the performance of the algorithms.

I. Introduction

Accurately counting the number of elements in a large dataset can be a time-consuming and resource-intensive task. This is where the CSURÖS approximate counter comes in. This mathematical algorithm is designed to provide a fast and accurate estimate of the number of elements in a set or a stream of data, without the need to actually count every single element. Based on the concept of random sampling, the CSURÖS approximate counter is able to provide reliable results with a low error rate. In this paper, we will explore the principles and applications of the CSURÖS approximate counter, as well as its potential limitations. We will also discuss the various fields where the CSURÖS approximate counter has been used, including computer science, data analysis, and statistics.

As technology continues to advance, the need for efficient storage of large numbers in computer systems becomes increasingly important. Traditional methods of representing numbers in a computer system often require a large amount of storage space, which can be limiting in certain situations. The Space-Saving Count method offers a solution to this problem by using a smaller number of bits to represent each digit in a number, resulting in a more compact representation that requires less memory to store. This method has proven to be effective in a variety of applications including data compression, image processing, and scientific simulations. In this paper, we will explore the origins and principles of the Space-Saving Count method, as well as its practical applications and potential limitations.

II. Method

The text sample used to collected results was the English version of the literary work "The tragedy of Othello, moor of Venice". Two more versions, one in French and one in German, were also employed to gather data pertaining to the most and least frequent letters of this same literary work in various languages. All letters would be capitalized when processing the text files, and each letter would have a matching counter that would track how many times it appeared in the text. The counters would come in one of the three categories listed below. All the tests presented are average of 10 consecutive try's of each counter.

Exact counter - The exact counter is implemented by processing the provided text file and, each time that particular letter is discovered, adding one unit to the counter corresponding to that letter. This ensures that every letter is accounted for and that the outcomes are entirely accurate. We did this using the collections.Counter() method in the collections library.

CSURÖS counter - This counter is a binary floating-point counter that employs a binary exponent along with a d-bit significand. D + loglogn bits are used in the counter. The parameter M = 2^d, with a non-negative integer d, defines the counting chain. The value of X is increased using the FP-increment method once the counter is setup with X = 0. (X). The first M-1 steps of the chain are deterministic, therefore the counter will give accurate results when counting events with a number of occurrences less than or equal to M-1. The likelihood of the counter increasing decreases by half every M steps in the following probabilistic process.

Space Saving Counter - this counter works by maintaining a count of the number of times that each item occurs in the stream. When a new item is encountered, its count is incremented. If the Space Saving Counter is already at capacity, the least frequently occurring item is removed to make room for the new item.

III. Experimental results

The experimental results can be found in the "Counters.xlsx" file. These results contains the exact number of occurrences of each letter, of the K most frequent letters using the Space-Saving Count and the counts obtained using the CSURÖS Counter with different values of D. In the file there is also some graphics to visualize the information. All the results that are presented are average values of 10 executions.

A. CSURÖS Counter

In the CSURÖS Counter depending on the value of D we got different results and we concluded that by increasing the value of D we got better results, see Fig. 1 and Fig. 2.

For example, for D=2 we got very bad results.

char	count	%	E relativo(%)	E absoluto
E	47.1	5.00	99.65620438	-13652.9
O	44.8	4.76	99.58800809	-10829.2
T	45.7	4.86	99.55239961	-10164.3
A	44.1	4.69	99.50560538	-8875.9
I	43.4	4.61	99.48632974	-8405.6
S	43.7	4.64	99.42935492	-7614.3
H	42.9	4.56	99.41529235	-7294.1
N	43.4	4.61	99.39730593	-7157.6
R	42.2	4.48	99.31670984	-6133.8
L	41.2	4.38	99.22527266	-5276.8
D	40.5	4.30	99.1426757	-4683.5
U	39.3	4.18	98.90254119	-3541.7
M	38.2	4.06	98.92847125	-3526.8
Y	37.3	3.96	98.63069016	-2686.7
W	36.6	3.89	98.56526852	-2514.4
C	36.8	3.91	98.47619048	-2378.2
F	36.5	3.88	98.3941927	-2236.5
G	36.7	3.90	98.33786232	-2171.3
B	33.5	3.56	97.97949337	-1624.5
P	34.0	3.61	97.79792746	-1510
V	33.6	3.57	97.40740741	-1262.4
K	30.8	3.27	96.85393258	-948.2
X	18.8	2.00	86.17647059	-117.2
J	18.4	1.96	78.85057471	-68.6
Q	15.8	1.68	78.05555556	-56.2
Z	11.8	1.25	57.85714286	-16.2
6	1.0	0.11	0	0
0	1.0	0.11	0	0
5	1.0	0.11	0	0
1	1.0	0.11	0	0
		media	82.49762933	-3824.90

TABLE I: Counts and errors for each letter. CSURÖS Counter, d=2 (English)

By looking at the table we can conclude that the error is very big for d=2, and that the counts deviate too much from the exact count. This happens because the M value that depends on the D value is small, and therefore will do much more probabilistic updates.

But for D=12 the results were much better

char	count	%	E relativo(%)	E absoluto
E	8528.3	9.22	37.75	-5171.7
O	7459.0	8.07	31.40	-3415
T	7169.9	7.75	29.78	-3040.1
A	6480.7	7.01	27.35	-2439.3
I	6269.0	6.78	25.80	-2180
S	5863.4	6.34	23.43	-1794.6
H	5701.1	6.16	22.30	-1635.9
N	5636.4	6.01	21.73	-1564.6
R	5136.7	5.55	16.83	-1039.3
L	4702.9	5.09	11.57	-615.1
D	4414.8	4.77	6.55	-309.2
U	3581.0	3.87	0	0
M	3565.0	3.85	0	0
Y	2724.0	2.95	0	0
W	2551.0	2.76	0	0
C	2415.0	2.61	0	0
F	2273.0	2.46	0	0
G	2208.0	2.39	0	0
B	1658.0	1.79	0	0
P	1544.0	1.67	0	0
V	1296.0	1.40	0	0
K	979.0	1.06	0	0
X	136.0	0.15	0	0
J	87.0	0.09	0	0
Q	72.0	0.08	0	0
Z	28.0	0.03	0	0
6	1.0	0.001	0	0
0	1.0	0.001	0	0
5	1.0	0.001	0	0
1	1.0	0.001	0	0
		media	8.482565709	-773.49

TABLE II: Counts and errors for each letter. CSURÖS Counter, d=12 (English)

By looking at the table we can conclude that the error is small for d=12, and that the counts deviate a bit from the exact count. We can also see that the more the count is the more error is associated. This happens because the M value that depends on the D value is bigger than in the previous example, and therefore will do much less probabilistic updates.

We found that the D=14 and more already gives us the best result possible. For more tests see "Counters.xlsx" file, CSURÖS Counter page.

If we analyse your count by language (Table 3 and 4) we can see that d=14 is not enough to get the same results as the exact count. This happens because these languages have more characters and the number of occurrences of each letter is also bigger. We can also observe that the letters that appear more frequent are mostly the same in the 3 languages used (English,French and German). See the exact results at the end of this report.

char	count	percentage
E	19212.9	13.677494769379281
S	11502.0	8.188172781693575
O	10218.0	7.274104458645883
A	9974.0	7.100403001618129
I	9835.0	7.001450122409694
N	9458.0	6.733067133477467
U	8987.0	6.397766370116516
R	8977.0	6.390647457943246
T	8890.0	6.328712922035809
L	7666.0	5.457358072027729
M	4892.0	3.4825718351630126
D	4646.0	3.307446595700604
C	4362.0	3.1052694899797753
Ã	4359.0	3.1031338163277944
P	3480.0	2.477381436297482
V	2965.0	2.110757459374148
Q	1875.0	1.3347960324878676
J	1690.0	1.2030961572823982
G	1632.0	1.16180646667744
H	1595.0	1.135466491636346
F	1429.0	1.0172925495600869
B	1270.0	0.9041018460051157
Z	628.0	0.44706768448126977
X	422.0	0.30041809371193606
Y	313.0	0.222821951023308
§	55.0	0.039154016952977454
1	25.0	0.01779728043317157
K	20.0	0.014237824346537253
2	19.0	0.013525933129210393
W	18.0	0.012814041911883529
Â	14.0	0.009966477042576078
3	7.0	0.004983238521288039
5	6.0	0.0042713473039611765
4	6.0	0.0042713473039611765
8	4.0	0.0028475648693074506
6	4.0	0.0028475648693074506
0	4.0	0.0028475648693074506
7	4.0	0.0028475648693074506
9	4.0	0.0028475648693074506
Š	3.0	0.0021356736519805883

TABLE III: Counts for each letter in French. CSURÖS Counter, d=14

char	count	percentage
E	20771.3	13.756347093896922
N	14708.0	9.740765048746873
I	13212.0	8.749999172154181
S	9848.0	6.5221005031315755
R	9645.0	6.387658342069866
H	9597.0	6.355869062606999
T	8501.0	5.630013848204865
A	8280.0	5.483650707344582
D	7023.0	4.651168951410749
L	6104.0	4.042536705027938
U	5616.0	3.7193456971554557
C	5471.0	3.6233155821113776
O	5314.0	3.5193381472015832
G	4819.0	3.1915112027407657
M	4641.0	3.073625958065967
B	3042.0	2.014645585959205
W	2954.0	1.956365240277282
Ã	2601.0	1.7225815808941132
F	2054.0	1.360316250348523
Z	1522.0	1.007985069635079
K	1387.0	0.9185777211457651
V	1177.0	0.7794996234957213
Ÿ	678.0	0.4490235724129984
J	543.0	0.35961622392368453
P	527.0	0.34901979743606215
Y	513.0	0.33974792425939254
¶	396.0	0.26226155556865394
Q	22.0	0.014570086420480773
X	12.0	0.007947319865716786
1	9.0	0.005960489899287589
3	3.0	0.0019868299664291965
2	2.0	0.0013245533109527976
Œ	2.0	0.0013245533109527976

TABLE IV: Counts for each letter in German. CSURÖS Counter, d=14

B. Space-Saving Count

In the Space-Saving Counter depending on the value of K we got different results and we concluded that by increasing the value of k we got better results, because k is the number of letters that will be counter, the more letters we have more the counter is close to the exact answer.

For example, for K=3 we got very bad results.

char	count	percent	E relativo	E absoluto
I	38563.0	33.333	356.4208782	30114
V	38563.0	33.333	2875.540123	37267
E	38562.0	33.332	181.4744526	24862

TABLE V: Space-Saving Count for k=3 (English)

By looking at the table we can conclude that the error is very big for k=3, and that the counts deviate to much from the exact count. This happens because the

this algorithm has to distribute all the counts of all the letters just by k letters, and if the k is low like 3 the relative and absolute errors will be very high.

char	count	percent	E relativo	E absoluto
O	23139.0	20.001	112.7919809	12265
E	23138.0	20.000	68.89051095	9438
I	23137.0	19.999	173.8430584	14688
M	23137.0	19.999	549.0042076	19572
C	23137.0	19.999	858.0538302	20722

TABLE VI: Space-Saving Count for k=5 (English)

In this case, k=5 we can see a big improvement over the previous experiment, we can see that the error is much smaller than for k=3. So if we keep increasing k we will eventually get an exact count

char	count	percent	E relativo	E absoluto
E	13700.0	11.842	0	0
O	11352.0	9.812	4.395806511	478
T	11340.0	9.802	11.0675808	1130
A	11329.0	9.792	27.00672646	2409
L	11328.0	9.791	113.0124107	6010
F	11328.0	9.791	398.3721953	9055
R	11328.0	9.791	83.41968912	5152
P	11328.0	9.791	633.6787565	9784
N	11328.0	9.791	57.31148452	4127
C	11327.0	9.790	369.0269151	8912

TABLE VII: Space-Saving Count for k=10 (English)

For k=10 we can see more of the same behavior, but in this case we can see that the higher the exact (real) count the less error is associated with it.

If we analyse your count by language (Table 8 and 9) we can see that appear more frequent are mostly the same in the 3 languages used (English,French and German). See the exact results at the end of this report.

char	count	percentage
E	22006.0	15.360453428635246
T	13475.0	9.405712530712531
A	13474.0	9.405014518650882
H	13473.0	9.404316506589234
K	13473.0	9.404316506589234
L	13473.0	9.404316506589234
I	13473.0	9.404316506589234
O	13473.0	9.404316506589234
R	13472.0	9.403618494527585
P	13472.0	9.403618494527585

TABLE VIII: Counts for k most frequent letters in French. Space Saving Counter, k=10

char	count	percentage
E	25145.0	16.1841563256269
N	14711.0	9.468487719478915
O	14441.0	9.29470676072293
R	14440.0	9.294063127542351
A	14439.0	9.293419494361773
S	14439.0	9.293419494361773
T	14439.0	9.293419494361773
K	14438.0	9.292775861181196
L	14438.0	9.292775861181196
P	14438.0	9.292775861181196

TABLE IX: Counts for k most frequent letters in German. Space Saving Counter, k=10

IV. Conclusion

In conclusion, the CSURÖS Counter is a versatile and efficient tool for counting and tracking items. Its innovative design allows for easy and accurate counting, while its compact size makes it perfect for use in any setting. Whether you need to count inventory in a warehouse or keep track of small items in a workshop, the CSURÖS Counter is an excellent choice. In the results that we got we concluded that CSURÖS Counter gives very good results if the value of D is big enough.

In conclusion, the Space Saving Count is a valuable tool for anyone looking to maximize their space and organization. It allows for accurate counting and tracking of items, helping to reduce clutter and ensure that everything has its own designated place. With its compact size and easy-to-use design, the Space Saving Count is a must-have for anyone looking to declutter and streamline their living or work space. By analysing the results we concluded that Space Saving Count gives very good results if the value of K is big enough, and that it presents with little error the k most frequent letters.

We can also conclude that in the 3 languages used the most frequent letters were mostly the same

References

[1] https://arxiv.org/pdf/0904.3062.pdf

[2] http://romania.amazon.com/techon/presentations/DataStreamsAlgorithms$_F lorin Manolache.pdf$

[3] Teacher slides

V. Apendix

English				French				German		
char	count	percentage		char	count	percentage		char	count	percentage
E	13700	11.84219625		E	22006	15.36045343		E	25145	16.18415633
O	10874	9.399419127		S	11502	8.028534733		N	14708	9.46655682
T	10210	8.825461586		O	10218	7.132287246		I	13212	8.503681582
A	8920	7.710393472		A	9974	6.961972303		S	9848	6.338499562
I	8449	7.303263951		I	9835	6.864948626		R	9645	6.207842027
S	7658	6.619528387		N	9458	6.601798079		H	9597	6.176947634
H	7337	6.342057949		U	8987	6.273034398		T	8501	5.471525668
N	7201	6.22450038		R	8977	6.266054277		A	8280	5.329282735
R	6176	5.338496646		T	8890	6.205327228		D	7023	4.520235827
L	5318	4.596846691		L	7666	5.350960465		L	6104	3.928736934
D	4724	4.083396722		M	4892	3.414675006		U	5616	3.614643942
U	3581	3.095394509		D	4646	3.242964038		C	5471	3.521317131
M	3565	3.081564207		C	4362	3.044728613		O	5314	3.420266722
Y	2724	2.354608948		Ă	4359	3.042634577		G	4819	3.101668297
W	2551	2.205068806		P	3480	2.429081975		M	4641	2.987101591
C	2415	2.087511237		V	2965	2.069605763		B	3042	1.957932135
F	2273	1.964767305		Q	1875	1.308772616		W	2954	1.901292415
G	2208	1.908581703		J	1690	1.179640384		Ă	2601	1.674089903
B	1658	1.433165065		G	1632	1.139155685		F	2054	1.322022553
P	1544	1.334624162		H	1595	1.113329238		Z	1522	0.979609701
V	1296	1.120254478		F	1429	0.997459236		K	1387	0.892719221
K	979	0.846241615		B	1270	0.886475318		V	1177	0.757556254
X	136	0.117557569		Z	628	0.438351575		Ÿ	678	0.436383296
J	87	0.075202268		X	422	0.29456109		J	543	0.349492817
Q	72	0.06223636		Y	313	0.218477775		P	527	0.339194686
Z	28	0.024203029		§	55	0.038390663		Y	513	0.330183822
1	1	0.000864394		1	25	0.017450302		¶	396	0.25487874
6	1	0.000864394		K	20	0.013960241		Q	22	0.01415993
0	1	0.000864394		2	19	0.013262229		X	12	0.007723598
5	1	0.000864394		W	18	0.012564217		1	9	0.005792699
				Â	14	0.009772169		3	3	0.0019309
				3	7	0.004886084		Œ	2	0.001287266
				4	6	0.004188072		2	2	0.001287266
				5	6	0.004188072				
				6	4	0.002792048				
				7	4	0.002792048				
				8	4	0.002792048				
				9	4	0.002792048				
				0	4	0.002792048				
				Š	3	0.002094036				

TABLE X: Exact counts for different languages

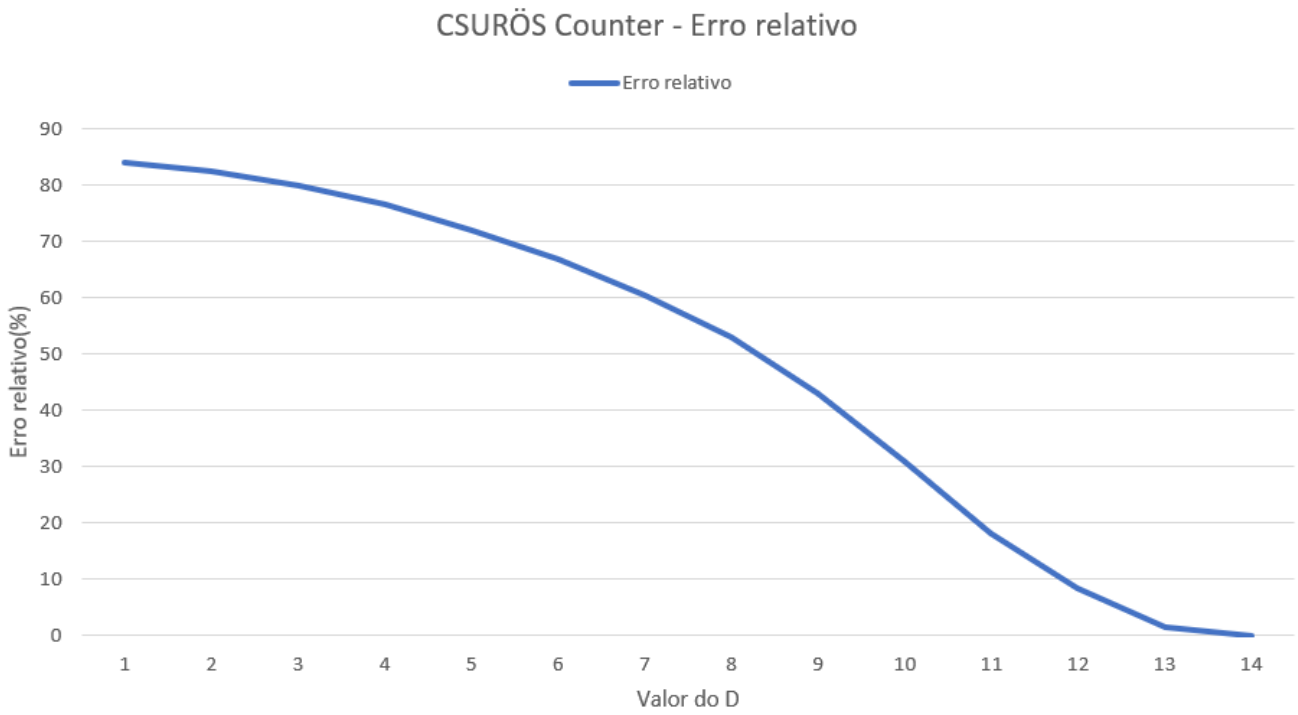

Fig. 1: Evolution of the relative error with the increase of d

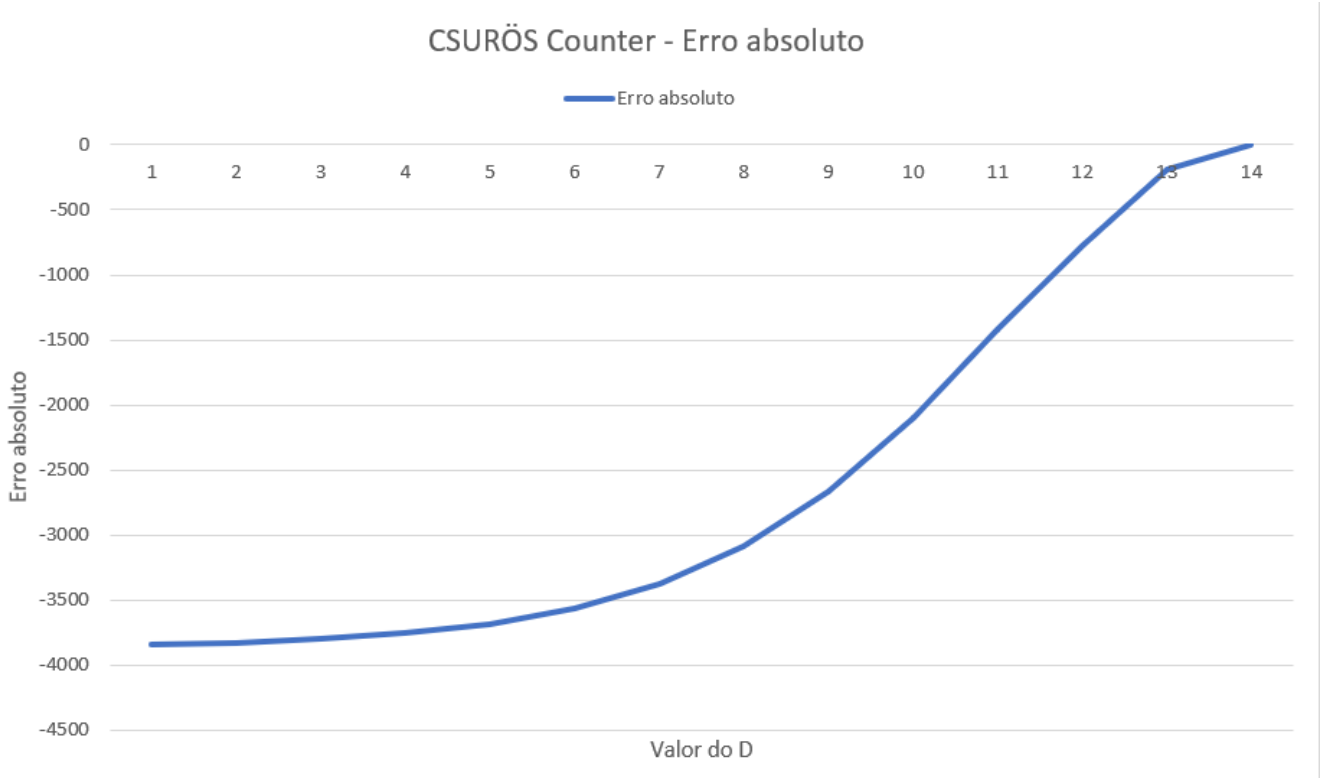

Fig. 2: Evolution of the absolute error with the increase of d