Count the Number of Occurrences of Letters in Text Files using Probabilistic Methods

João Carvalho, 89059

***Resumo* - Neste relatório é abordado e analisado o processo de desenvolvimento e resultados de *counters* de letras num determinado texto usando algoritmos probabilísticos.**

O objetivo deste problema é fazer uma estimativa do número de vezes que cada letra aparece num certo texto usando um *counting method* com probabilidades fixa e decrescente.

Ao longo do relatório são analisados e comparados os resultados obtidos entre os dois métodos e tendo por base a contagem exata de cada letra.

I. INTRODUCTION

A implementação de counters probabilísticos é bastante usada em ambientes de grande quantidade de dados, uma vez que usa menos espaço de memória para armazenamento do *counter*.

Neste trabalho foi tratado o desenvolvimento de um *counter* com probabilidade fixa(***FC***) - 1 / 2^k , especificamente com k de valor 2 (¼), e um *counter* com probabilidade decrescente(***DC***) - 1 / 2^k - com k representando o valor do *counter*.

Este métodos foram testados em várias repetições para a obtenção de resultados fidedignos cujos valores são, posteriormente, analisados e comparados em diferentes aspetos recorrendo ao counter exato***(EX)***.

II. FIXED COUNTER

Primeiramente será abordado o *counter* com probabilidade fixa. Neste tipo de *counter*, um evento é contabilizado tendo em conta uma certa probabilidade, neste caso p = ¼. Com esta definição é possível perceber que, no resultado final, a contagem do número de vezes que uma letra aparece num texto será aproximadamente ¼ do valor exato dessa letra. Em outras palavras, para estimar o número real da contagem do *FC*, apenas se multiplica o valor do counter por 4.

Abaixo, na Fig.1, encontra-se um gráfico com a comparação do *FC* com o *EX* para texto escrito em finlandês(filandes.txt) com um total de 30878 letras(*len*). O *FC,* neste caso, realizou 100 repetições, isto é, efetuou a contagem das letras 100 vezes, e os valores no gráfico representam o mean value.

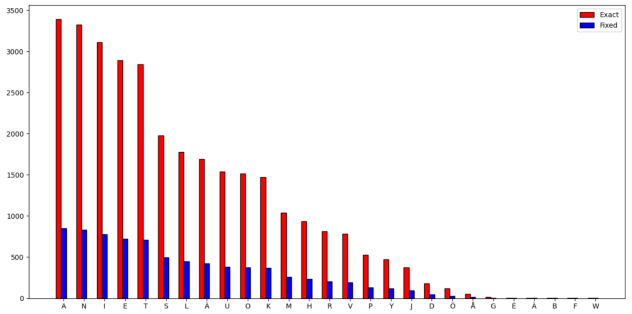


Fig. 1: comparação do FC e EC

O número de repetições é bastante relevante pois, como se trata de um processo probabilístico, um número baixo de repetições pode levar a resultados pouco precisos.

Na Fig.2 está retratado um gráfico que compara a exatidão média em percentagem do *FC* com quatro diferentes números de repetições: 2(vermelho), 10(verde), 100(azul), 1000(preto).

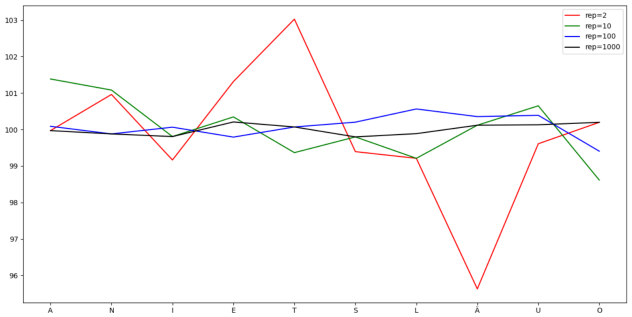


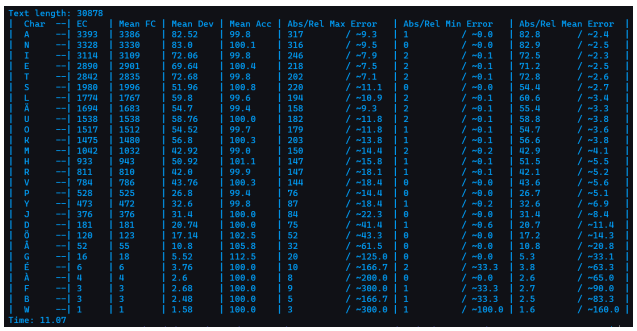
Fig. 2: comparação de exatidão do *FC* com diferentes números de repetições

A exatidão foi calculada com a divisão do valor médio esperado pelo valor exato. Pode-se observar que quanto maior o número de repetições maior a exatidão(valores mais próximos de 100%).

Neste ponto, será analisado mais profundamente o algoritmo *FC* tendo por base algumas métricas: *Estimated count*(*Mean FC*), *Mean absolute deviation*, *Mean accuracy*, *Max error*, *Min error*, *Mean error*. Estes três últimos com valores absolutos e relativos.

Na Fig.3 está representada uma tabela com estes valores resultantes de um teste com o exemplo de texto testado anteriormente(filandes.txt) e com 100 repetições. As letras

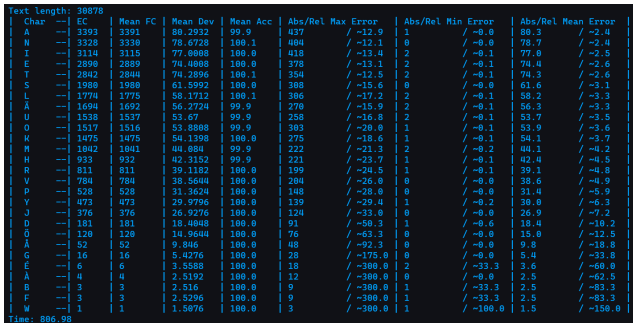
estão organizadas por ordem decrescente do seu número(contagem).



```
Text length: 30878
| Char  --| EC   | Mean FC | Mean Dev | Mean Acc | Abs/Rel Max Error  | Abs/Rel Min Error  | Abs/Rel Mean Error |
| A     --| 3393 | 3386    | 82.52    | 99.8     | 317        / ~9.3  | 1          / ~0.0  | 82.8       / ~2.4  |
| N     --| 3328 | 3330    | 83.0     | 100.1    | 316        / ~9.5  | 0          / ~0.0  | 82.9       / ~2.5  |
| I     --| 3114 | 3109    | 72.06    | 99.8     | 246        / ~7.9  | 2          / ~0.1  | 72.5       / ~2.3  |
| E     --| 2890 | 2901    | 69.64    | 100.4    | 218        / ~7.5  | 2          / ~0.1  | 71.2       / ~2.5  |
| T     --| 2842 | 2835    | 72.68    | 99.8     | 202        / ~7.1  | 2          / ~0.1  | 72.8       / ~2.6  |
| S     --| 1980 | 1996    | 51.96    | 100.8    | 220        / ~11.1 | 0          / ~0.0  | 54.4       / ~2.7  |
| L     --| 1774 | 1767    | 59.8     | 99.6     | 194        / ~10.9 | 2          / ~0.1  | 60.6       / ~3.4  |
| Á     --| 1694 | 1683    | 54.7     | 99.4     | 158        / ~9.3  | 2          / ~0.1  | 55.4       / ~3.3  |
| U     --| 1538 | 1538    | 58.76    | 100.0    | 182        / ~11.8 | 2          / ~0.1  | 58.8       / ~3.8  |
| O     --| 1517 | 1512    | 54.52    | 99.7     | 179        / ~11.8 | 1          / ~0.1  | 54.7       / ~3.6  |
| K     --| 1475 | 1480    | 56.8     | 100.3    | 203        / ~13.8 | 1          / ~0.1  | 56.6       / ~3.8  |
| M     --| 1042 | 1032    | 42.92    | 99.0     | 150        / ~14.4 | 2          / ~0.2  | 42.9       / ~4.1  |
| H     --| 933  | 943     | 50.92    | 101.1    | 147        / ~15.8 | 1          / ~0.1  | 51.5       / ~5.5  |
| R     --| 811  | 810     | 42.0     | 99.9     | 147        / ~18.1 | 1          / ~0.1  | 42.1       / ~5.2  |
| V     --| 784  | 786     | 43.76    | 100.3    | 144        / ~18.4 | 0          / ~0.0  | 43.6       / ~5.6  |
| P     --| 528  | 525     | 26.8     | 99.4     | 76         / ~14.4 | 0          / ~0.0  | 26.7       / ~5.1  |
| Y     --| 473  | 472     | 32.6     | 99.8     | 87         / ~18.4 | 1          / ~0.2  | 32.6       / ~6.9  |
| J     --| 376  | 376     | 31.4     | 100.0    | 84         / ~22.3 | 0          / ~0.0  | 31.4       / ~8.4  |
| D     --| 181  | 181     | 20.74    | 100.0    | 75         / ~41.4 | 1          / ~0.6  | 20.7       / ~11.4 |
| Ó     --| 120  | 122     | 17.14    | 102.5    | 52         / ~43.3 | 0          / ~0.0  | 17.2       / ~14.2 |
| Â     --| 52   | 55      | 10.8     | 105.8    | 32         / ~61.5 | 0          / ~0.0  | 10.8       / ~20.8 |
| G     --| 16   | 18      | 5.52     | 112.5    | 20         / ~125.0| 0          / ~0.0  | 5.3        / ~33.1 |
| É     --| 6    | 6       | 3.76     | 100.0    | 10         / ~166.7| 2          / ~33.3 | 3.8        / ~63.3 |
| À     --| 4    | 4       | 2.6      | 100.0    | 8          / ~200.0| 1          / ~0.0  | 2.6        / ~65.0 |
| Ê     --| 3    | 3       | 2.68     | 100.0    | 9          / ~300.0| 0          / ~33.3 | 2.7        / ~90.0 |
| B     --| 3    | 3       | 2.48     | 100.0    | 5          / ~166.7| 1          / ~33.3 | 2.5        / ~83.3 |
| W     --| 1    | 1       | 1.58     | 100.0    | 3          / ~300.0| 1          / ~100.0| 1.6        / ~160.0|
Time: 11.07
```

Fig. 3: *FC* com 100 repetições

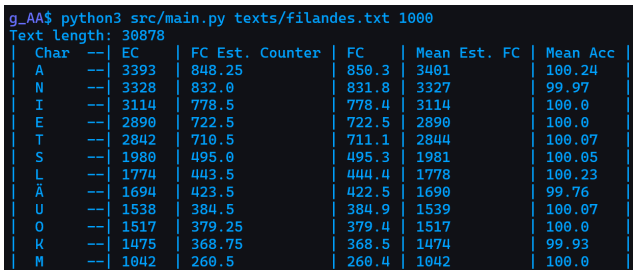
E, para se poder ter uma comparação de valores com um número diferente de repetições, tem-se na Fig.4 uma tabela de um teste que utilizou o mesmo ficheiro de texto porém com 10000 repetições.



```
Text length: 30878
| Char  --| EC   | Mean FC | Mean Dev | Mean Acc | Abs/Rel Max Error  | Abs/Rel Min Error  | Abs/Rel Mean Error |
| A     --| 3393 | 3391    | 80.2932  | 99.9     | 437        / ~12.9 | 1          / ~0.0  | 80.3       / ~2.4  |
| N     --| 3328 | 3336    | 78.6728  | 100.1    | 404        / ~12.1 | 0          / ~0.0  | 78.7       / ~2.4  |
| I     --| 3114 | 3115    | 77.8008  | 100.0    | 418        / ~13.4 | 2          / ~0.1  | 77.8       / ~2.5  |
| E     --| 2890 | 2889    | 74.4008  | 100.0    | 378        / ~13.1 | 2          / ~0.1  | 74.4       / ~2.6  |
| T     --| 2842 | 2844    | 74.2896  | 100.1    | 354        / ~12.5 | 2          / ~0.1  | 74.3       / ~2.6  |
| S     --| 1980 | 1980    | 61.5992  | 100.0    | 308        / ~15.6 | 0          / ~0.0  | 61.6       / ~3.1  |
| L     --| 1774 | 1775    | 58.1712  | 100.1    | 306        / ~17.2 | 2          / ~0.1  | 58.2       / ~3.3  |
| Á     --| 1694 | 1692    | 56.2724  | 99.9     | 270        / ~15.9 | 2          / ~0.1  | 56.3       / ~3.3  |
| U     --| 1538 | 1537    | 53.67    | 99.9     | 258        / ~16.8 | 2          / ~0.1  | 53.7       / ~3.5  |
| O     --| 1517 | 1516    | 53.8808  | 99.9     | 303        / ~20.0 | 1          / ~0.1  | 53.9       / ~3.6  |
| K     --| 1475 | 1475    | 54.1398  | 100.0    | 275        / ~18.6 | 1          / ~0.1  | 54.1       / ~3.7  |
| M     --| 1042 | 1041    | 44.084   | 99.9     | 222        / ~21.3 | 2          / ~0.2  | 44.1       / ~4.2  |
| H     --| 933  | 932     | 42.3152  | 99.9     | 221        / ~23.7 | 1          / ~0.1  | 42.4       / ~4.5  |
| R     --| 811  | 811     | 39.1182  | 100.0    | 199        / ~24.5 | 1          / ~0.1  | 39.1       / ~4.8  |
| V     --| 784  | 784     | 38.5648  | 100.0    | 204        / ~26.0 | 0          / ~0.0  | 38.6       / ~4.9  |
| P     --| 528  | 528     | 31.3624  | 100.0    | 148        / ~28.0 | 0          / ~0.0  | 31.4       / ~5.9  |
| Y     --| 473  | 473     | 29.9796  | 100.0    | 139        / ~29.4 | 1          / ~0.2  | 30.0       / ~6.3  |
| J     --| 376  | 376     | 26.9276  | 100.0    | 124        / ~33.0 | 0          / ~0.0  | 26.9       / ~7.2  |
| D     --| 181  | 181     | 18.4048  | 100.0    | 91         / ~50.3 | 1          / ~0.6  | 18.4       / ~10.2 |
| Ó     --| 120  | 120     | 14.9644  | 100.0    | 76         / ~63.3 | 0          / ~0.0  | 15.0       / ~12.5 |
| Â     --| 52   | 52      | 9.846    | 100.0    | 48         / ~92.3 | 0          / ~0.0  | 9.8        / ~18.8 |
| G     --| 16   | 16      | 5.4276   | 100.0    | 28         / ~175.0| 0          / ~0.0  | 5.4        / ~33.8 |
| É     --| 6    | 6       | 3.5588   | 100.0    | 18         / ~300.0| 2          / ~33.3 | 3.6        / ~60.0 |
| À     --| 4    | 4       | 2.5192   | 100.0    | 12         / ~300.0| 0          / ~0.0  | 2.5        / ~62.5 |
| B     --| 3    | 3       | 2.516    | 100.0    | 9          / ~300.0| 1          / ~33.3 | 2.5        / ~83.3 |
| F     --| 3    | 3       | 2.5296   | 100.0    | 9          / ~300.0| 1          / ~33.3 | 2.5        / ~83.3 |
| W     --| 1    | 1       | 1.5076   | 100.0    | 3          / ~300.0| 1          / ~100.0| 1.5        / ~150.0|
Time: 806.98
```

Fig. 4: Análise do FC com 10000 repetições

A seguir, na Fig. 5, pode-se ver a comparação de valores dos quatro tipos de valor do *counter*: valor exato, estimativa do valor do *FC*, valor do *FC* e a estimativa do valor exato. Teste realizado com 1000 repetições sobre o mesmo ficheiro sendo apresentadas apenas as letras mais comuns(para o *FC*).



```
g_AA$ python3 src/main.py texts/filandes.txt 1000
Text length: 30878
|  Char  --| EC   | FC Est. Counter | FC    | Mean Est. FC | Mean Acc |
|  A     --| 3393 | 848.25          | 850.3 | 3401         | 100.24   |
|  N     --| 3328 | 832.0           | 831.8 | 3327         | 99.97    |
|  I     --| 3114 | 778.5           | 778.4 | 3114         | 100.0    |
|  E     --| 2890 | 722.5           | 722.5 | 2890         | 100.0    |
|  T     --| 2842 | 710.5           | 711.1 | 2844         | 100.07   |
|  S     --| 1980 | 495.0           | 495.3 | 1981         | 100.05   |
|  L     --| 1774 | 443.5           | 444.4 | 1778         | 100.23   |
|  Á     --| 1694 | 423.5           | 422.5 | 1690         | 99.76    |
|  U     --| 1538 | 384.5           | 384.9 | 1539         | 100.07   |
|  O     --| 1517 | 379.25          | 379.4 | 1517         | 100.0    |
|  K     --| 1475 | 368.75          | 368.5 | 1474         | 99.93    |
|  M     --| 1042 | 260.5           | 260.4 | 1042         | 100.0    |
```

Fig. 5: Comparação de valores dos *counters*

A estimativa do valor exato foi calculada, como já referido, pela multiplicação do valor do *counter* por 4(fazendo a média posteriormente), já a estimativa do valor do *FC* é o inverso, divide-se o valor exato por 4.

A compreensão destas tabelas será feita no bloco de análise de resultados a seguir no relatório.

III. DECREASING COUNTER

Agora será abordado o *counter* com probabilidade decrescente.

Neste tipo de counter, um evento é contabilizado tendo em conta uma probabilidade que vai decrescendo à medida que aumenta a contagem, neste caso p = 1 / 2^k (k representa o valor do *counter*). Nesta situação também se pode definir como *log counter* pois o valor do *counter* aumenta de forma logarítmica em relação ao número de eventos. Assim, é possível perceber que, no resultado final, a contagem do número de vezes que uma letra aparece num texto será aproximadamente o valor inteiro de log2 (n+1) com n sendo o valor exato dessa letra. Para estimar o número de surgimentos da letra no texto com base no counter pode-se recorrer à expressão 2^k -1.

Abaixo, na Fig.6, encontra-se um gráfico com a comparação do *DC*(100 repetições) com o *EX* para texto escrito em finlandês(filandes.txt) com um total de 30878 letras(*len*). Contudo, contrariamente ao exemplo do *FC*(Fig.1), os valores do *DC* são já os valores estimados pois os valores do *counter* são demasiado pequenos para serem visíveis.

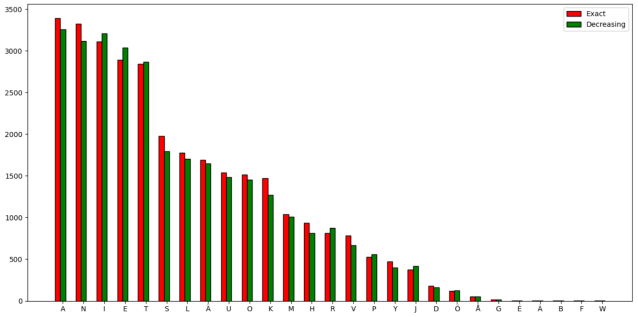


Fig. 6: comparação do *DC* e *EC*

Na Fig.7, abaixo, está o gráfico da exatidão para o *DC* com quatro diferentes números de repetições: 2(vermelho), 10(verde), 100(azul), 1000(preto). Estão expostas apenas as letras mais comuns.

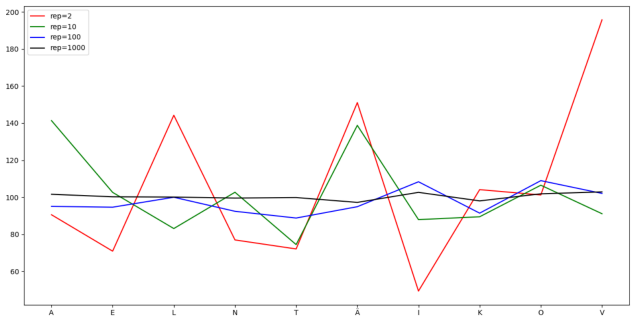
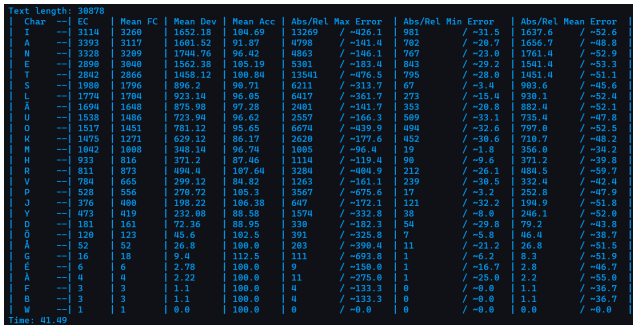


Fig. 7: comparação de exatidão do *FC* com diferentes números de repetições

Fazendo já uma rápida análise de ambos os gráficos de exatidão(Fig.2, Fig.6), pode-se já verificar que o *DC*

precisa de realizar mais repetições para adquirir uma exatidão razoável, conclusão que, já seria esperada.

Agora, irá analisar-se o *DC* recorrendo à tabela com as métricas já referidas anteriormente. Na Fig.8 pode-se observá-la num teste com 100 repetições do mesmo ficheiro usado anteriormente(filandês.txt).



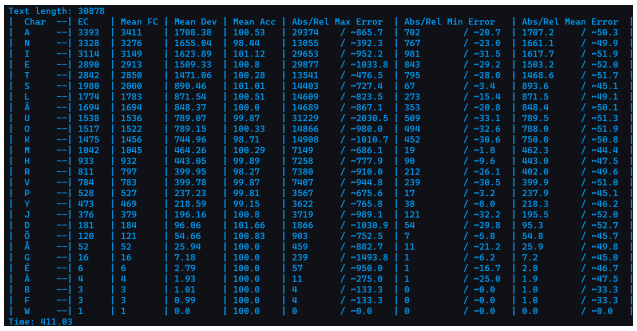
Text length: 30878

Char	EC	Mean FC	Mean Dev	Mean Acc	Abs/Rel Max Error	Abs/Rel Min Error	Abs/Rel Mean Error
I	3114	3260	1652.18	104.69	13269 / ~426.1	981 / ~31.5	1637.6 / ~52.6
A	3393	3117	1601.52	91.87	4798 / ~141.4	702 / ~20.7	1656.7 / ~48.8
N	3328	3209	1744.76	96.42	4863 / ~146.1	767 / ~23.0	1761.4 / ~52.9
E	2890	3040	1562.38	105.19	5381 / ~183.4	843 / ~29.2	1541.4 / ~53.3
T	2842	2866	1458.12	100.84	13541 / ~476.5	795 / ~28.0	1451.4 / ~51.1
S	1980	1796	896.2	90.71	6211 / ~313.7	67 / ~3.4	903.6 / ~45.6
L	1774	1704	923.14	96.85	6417 / ~361.7	273 / ~15.4	930.1 / ~52.4
Ã	1694	1648	875.98	97.28	2401 / ~141.7	353 / ~20.8	882.4 / ~52.1
O	1538	1486	723.94	96.62	2557 / ~166.3	509 / ~33.1	735.4 / ~47.8
U	1517	1451	781.12	95.65	6674 / ~439.9	494 / ~32.6	797.0 / ~52.5
K	1475	1271	629.12	86.17	2620 / ~177.6	452 / ~30.6	710.7 / ~48.2
M	1042	1008	348.14	96.74	1005 / ~96.4	19 / ~1.8	356.0 / ~34.2
H	933	816	371.2	87.46	1114 / ~119.4	90 / ~9.6	371.2 / ~39.8
R	811	873	494.4	107.66	3281 / ~404.9	212 / ~26.1	484.5 / ~59.7
V	784	665	299.12	84.82	1263 / ~161.1	239 / ~30.5	332.4 / ~42.4
P	528	556	270.72	105.3	3567 / ~675.6	17 / ~3.2	252.8 / ~47.9
J	376	400	198.22	106.38	647 / ~172.1	121 / ~32.2	194.9 / ~51.8
Y	473	419	232.08	88.58	1574 / ~332.8	38 / ~8.0	246.1 / ~52.0
D	181	161	88.95	100.0	330 / ~182.3	54 / ~29.8	79.2 / ~43.8
Ô	120	123	45.6	102.5	391 / ~325.8	7 / ~5.8	46.4 / ~38.7
Â	52	52	26.8	100.0	203 / ~390.4	11 / ~21.2	26.8 / ~51.5
G	16	18	9.4	112.5	111 / ~693.8	1 / ~6.2	8.3 / ~51.9
É	6	6	2.78	100.0	9 / ~150.0	1 / ~16.7	2.8 / ~46.7
À	4	4	2.22	100.0	11 / ~275.0	1 / ~25.0	2.2 / ~55.0
F	3	3	1.1	100.0	4 / ~133.3	0 / ~0.0	1.1 / ~36.7
B	3	3	1.1	100.0	4 / ~133.3	0 / ~0.0	1.1 / ~36.7
W	1	1	0.0	100.0	0 / ~0.0	0 / ~0.0	0.0 / ~0.0

Time: 41.49

Fig. 8: *DC* com 100 repetições

E na Fig. 9 encontra-se um teste com 10000 repetições.



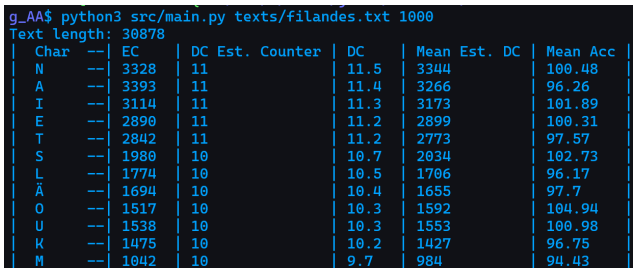
Text length: 30878

Char	EC	Mean FC	Mean Dev	Mean Acc	Abs/Rel Max Error	Abs/Rel Min Error	Abs/Rel Mean Error
A	3393	3411	1708.38	100.53	29374 / ~865.7	702 / ~20.7	1707.2 / ~50.3
N	3328	3276	1655.04	98.44	13055 / ~392.3	767 / ~23.0	1661.1 / ~49.9
I	3114	3149	1623.89	101.12	29653 / ~952.2	981 / ~31.5	1617.7 / ~51.9
E	2890	2913	1509.33	100.8	29877 / ~1033.8	843 / ~29.2	1503.2 / ~52.0
T	2842	2850	1471.06	100.28	13541 / ~476.5	795 / ~28.0	1468.6 / ~51.7
S	1980	2000	890.46	101.01	14403 / ~727.4	67 / ~3.4	893.6 / ~45.1
L	1774	1783	871.54	100.51	14689 / ~823.5	273 / ~15.4	871.5 / ~49.1
Ã	1694	1694	848.37	100.0	14689 / ~867.1	353 / ~20.8	848.4 / ~50.1
U	1538	1536	789.07	99.87	31229 / ~2030.5	509 / ~33.1	789.5 / ~51.3
O	1517	1522	789.15	100.33	14866 / ~980.0	494 / ~32.6	788.0 / ~51.9
K	1475	1456	744.96	98.71	14908 / ~1010.7	452 / ~30.6	750.0 / ~50.8
M	1042	1045	464.26	100.29	7149 / ~686.1	19 / ~1.8	462.3 / ~44.4
H	933	932	443.05	99.89	7258 / ~777.9	90 / ~9.6	443.0 / ~47.5
R	811	797	399.95	98.27	7380 / ~910.0	212 / ~26.1	402.0 / ~49.6
V	784	783	399.78	99.87	7407 / ~944.8	239 / ~30.5	399.9 / ~51.0
P	528	527	237.23	99.81	3567 / ~675.6	17 / ~3.2	237.9 / ~45.1
Y	473	469	218.59	99.15	3622 / ~765.8	38 / ~8.0	218.3 / ~46.2
J	376	379	196.16	100.8	3719 / ~989.1	121 / ~32.2	195.5 / ~52.0
D	181	184	96.06	101.66	1866 / ~1030.9	54 / ~29.8	95.3 / ~52.7
Ô	120	121	54.66	100.83	903 / ~752.5	7 / ~5.8	54.8 / ~45.7
Â	52	52	25.94	100.0	459 / ~882.7	11 / ~21.2	25.9 / ~49.8
G	16	16	7.18	100.0	239 / ~1493.8	1 / ~6.2	7.2 / ~45.0
É	6	6	2.79	100.0	57 / ~950.0	1 / ~16.7	2.8 / ~46.7
À	4	4	1.93	100.0	11 / ~275.0	1 / ~25.0	1.9 / ~47.5
B	3	3	1.01	100.0	4 / ~133.3	0 / ~0.0	1.0 / ~33.3
F	3	3	0.99	100.0	4 / ~133.3	0 / ~0.0	1.0 / ~33.3
W	1	1	0.0	100.0	0 / ~0.0	0 / ~0.0	0.0 / ~0.0

Time: 411.03

Fig. 9: *DC* com 10000 repetições

A seguir, na Fig. 10, pode-se ver a comparação de valores dos quatro tipos de valor do *counter*. Teste realizado com 1000 repetições sobre o mesmo ficheiro sendo apresentadas apenas as letras mais comuns(para o *DC*).



```
g_AA$ python3 src/main.py texts/filandes.txt 1000
Text length: 30878
```

Char	EC	DC Est. Counter	DC	Mean Est. DC	Mean Acc
N	3328	11	11.5	3344	100.48
A	3393	11	11.4	3266	96.26
I	3114	11	11.3	3173	101.89
E	2890	11	11.2	2899	100.31
T	2842	11	11.2	2773	97.57
S	1980	10	10.7	2034	102.73
L	1774	10	10.5	1706	96.17
Ã	1694	10	10.4	1655	97.7
O	1517	10	10.3	1592	104.94
U	1538	10	10.3	1553	100.98
K	1475	10	10.2	1427	96.75
M	1042	10	9.7	984	94.43

Fig.10: Comparação de valores dos *counters*

A estimativa do valor exato foi calculada, como já referido, pela expressão 2^k -1, já a estimativa do valor do *DC* é o valor inteiro dado pela expressão log2 (n+1).

IV. RESULTS ANALYSIS

Numa visão geral das tabelas de métricas pode-se tirar algumas conclusões. Tendo em conta os valores absolutos, *mean deviation*, *abs max error*, *abs mean error*, é possível verificar um **aumento** destes à medida que **aumenta** o *counter* da letra(letras mais comuns), por outro lado, os valores relativos **diminuem** à medida que o counter da letra aumenta mas apenas no *FC, no caso de DC* os valores relativos parecem estar uniformemente distribuídos pelas letras. Estes valores já eram esperados uma vez que as letras com o maior counter passaram, obviamente, mais vezes pelo processo de contagem probabilística, tendo uma maior quantidade de informação. Pode-se ainda observar que, como já referido, o teste com 10000 repetições apresenta uma ***mean accuracy***, ou seja, uma **exatidão mais próxima** dos 100% num geral de todas as letras, aliás, no caso do *FC*, todas as letras obtiveram uma *mean accuracy* de 100% com uma margem de 0.1%.

Falando de **exatidão** pode-se extrair das tabelas que o *FC* tende a ter valores mais exatos quando comparado ao *DC* em testes com o mesmo número de repetições, resultado previsível pois o *FC* usa mais memória(*counters maiores*).

Outra observação importante é relacionada com a diferença de valores de *errors* quando comparados os dois tipos. Consegue-se observar que, o *mean deviation* e o abs *mean error* são **muito maiores no *DC***, isto significa que, apesar de ser razoavelmente exata, este tipo de *counter* é muito menos **preciso** que o *FC*.

Uma nota adicional é que, quando se ordena as letras por ordem decrescente do seu valor do *counter*, a lista pode ser diferente quando comparados os *DC* e *FC* com o *EC* sobretudo com números de repetições baixos(e sobretudo o *DC*), contudo com números de repetições mais elevados essa diferença praticamente não existe.

V. EXTRA TESTS

Neste bloco serão apresentados alguns testes destes *counters* com ficheiros de linguagens diferentes.

Assim podemos verificar outros valores dos *counters*, além de curiosidades como saber as letras mais comuns de cada linguagem. Todos os testes foram feitos com 100 repetições. É curioso reparar que as letras mais comuns de algumas das línguas mais faladas da europa são muito coicidentes.

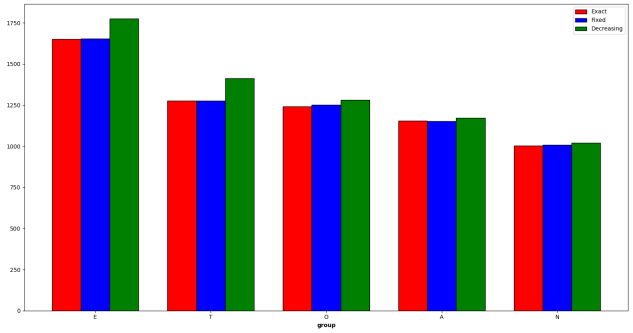


Fig.11: Texto em inglês com *len*=14296

E - T - O - A - N

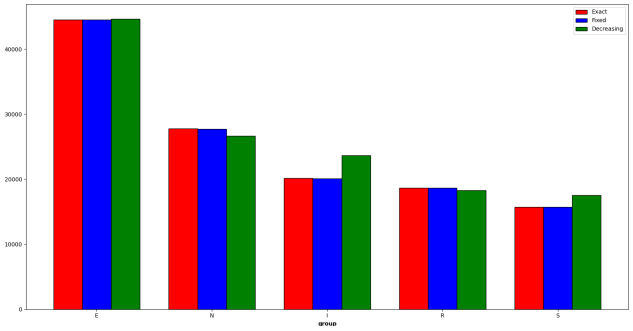


Fig.12: Texto em alemão com *len*=257920

E - N - I - R - S

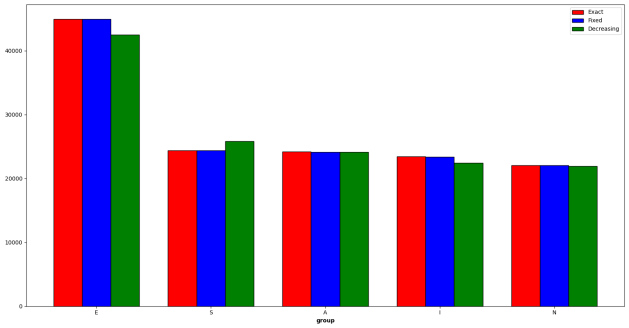


Fig.13: Texto em francês com *len*=311776

E - S - A - I - N

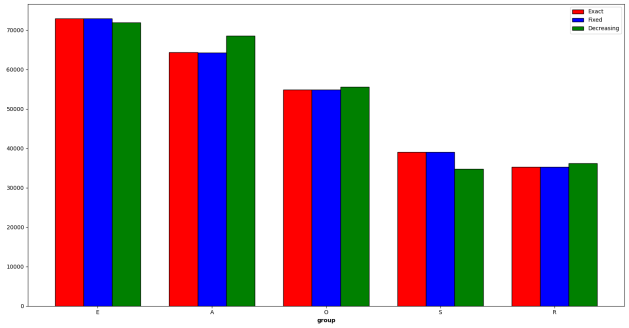


Fig.13: Texto em português com *len*=533657

E - A - O - S - R

VI. CONCLUSION

Por fim pode-se concluir que os resultados obtidos foram ao encontro das expectativas prévias. Os algoritmos provaram-se bastante fidedignos com um número razoavelmente alto de repetições.

Assim conclui-se que o processo de contagem probabilística é uma boa técnica no que diz respeito ao menor uso de memória em *Big Data*.

REFERENCES

[1]https://stackoverflow.com/questions/14576083/can-someone-explain-how-probabilistic-counting-works

[2]https://elearning.ua.pt/pluginfile.php/2931377/mod_resource/content/0/AA_09_Probabilistic_Counters.pdf