Most Frequent Letters – 3rd Assigment

Sobral, Pedro – 98491 – sobral@ua.pt

***Resumo* – Este relatório procura mostrar através de diferentes aborgagens algorítmicas como contar o número de vezes que uma determinada letra aparece numa determinada obra literária. Todos os algoritmos foram escritos em Python (3.10), e serão feitas analises sobre a complexidade computacional, os erros relativos e absolutos das contagens, entre outros tipos de comparação entre os algoritmos.**

***Abstract* – This report seeks to through different algorithmic approaches, how to count the number of times a given letter appears in a given literary work. All algorithms were written in Python (3.10), and analyzes will be made on computational complexity, relative and absolute errors of counts, among other types of comparison between algorithms.**

I. INTRODUÇÃO

Através deste trabalho é pedido que se façam comparações em diferentes nível e modos às 3 abordagens distintas da frequência das letras em determinadas obras literárias. Além deste relatório, foram criados os ficheiros de código em Python, e os ficheiros que guardam os resultados da implementação dos algoritmos. Em relação aos algoritmos desenvolvidos, como já referido há 3 tipos de abordagens, uma que conta o numero exato de letras, uma abordagem que faz uma contagem aproximada, e uma outra que faz a contagem de acordo com o LossyCount.

De modo a correr o algoritmo, basta correr o seguinte comando dentro da pasta /src, ou ler o README do projeto. O ficheiro "process_data.py" serve para adaptar as obras literárias para o formato que o enunciado do projeto pede, e o ficheiro "main.py" para correr os algoritmos e obter os resultados.

```
$ python3 process_data.py
$ python3 main.py
```

II. PREPROCESSAMENTO DOS DADOS

Todas as obras literárias foram consumidas do Projeto Guttenber REFERENCIA, e foram analisadas as seguintes obras:

- Alice no País das Maravilhas em Francês, Alemão, e em Inglês.
- Uma obra do filosofo Friedrich Nietzsche em Inglês.

Em todas as obras referidas anteriormente o cabeçalho dos ficheiros foi retirado, foram removidas as stopwords, os acentos, e as letras foram todas convertidas para maiúscula.

Assim feito, os dados (obras literárias) estão prontas para serem analisados pelos algoritmos de contagem de letras.

III. ALGORITMO DE CONTAGEM EXATA

A abordagem para a criação deste algoritmo foi a seguinte: Criar um dicionário onde o par chave/valor seja letra/numero de ocorrências, desta forma é criada uma iteração sobre as obras literárias de modo a passar por todas as letras e assim poder fazer a contagem certa das letras.

IV. ALGORITMO DE CONTAGEM APROXIMADA

Este tipo de algoritmo permite contar um número gigante de letras usando poucos recursos a nível de memória. Para esta implementação há um incremento de 1/sqrt(2)^k para os eventos. [1]

V. ALGORITMO LOSSY-COUNT

O algoritmo Lossy-count é direcionado para identificar elementos num data stream cuja frequência excede o threshold inserido pelo utilizador. A frequência neste caso de letras nem sempre é inteiramente precisa, no entanto tem um limite de erro, o threshold tal como referido anterirormente. [2]

V. RESULTADOS

A nível de resultados, os mesmos serão divididos primeiramente numa avaliação sobre cada obra literária, e posteriormente será feita uma analise comparativa em relação as várias obras, e que conclusões se podem tirar por exemplo da mesma obra, mas escrita em línguas diferentes.

A. Alice no País das maravilhas em Francês.

Tabela 1 - Frequência das Letras

Letter	Exact Counter	Approximate Counter	Lossy Counter
E	21351	26	21351
A	10822	26	10822
T	10011	25	10011
I	9823	25	9823
S	9245	24	9245
R	8414	23	8414
N	8292	23	8292
U	7687	23	7687
L	7546	23	7546
O	6779	23	6779
D	4544	22	4544
C	4413	22	4413
P	3663	22	3663
M	3022	21	3022
V	2247	21	2247
Q	1466	20	1466
F	1426	19	1426
G	1324	18	1324
B	1242	18	1242
H	1154	18	1154
J	890	18	890
X	431	16	431
Y	409	15	409
Z	389	15	389
W	197	11	197
K	138	11	138
--TIME--	0.021	0.085	0.029

Absolute and Relative Error for Approximate Counter		
Letter	Absolute Error	Relative Error
E	21325	0.9987822584422276
N	10797	0.9976898909628534
A	9987	0.9976026370991908
T	9799	0.9975567545556348
R	9221	0.9974040021633316
I	8391	0.9972664606608034
U	8269	0.9972262421611191
O	7664	0.9970079354754781
D	7523	0.9969520275642725
S	6756	0.996607169198997
L	4522	0.9951584507042254
P	4392	0.9952413324269205
M	3643	0.9945399945399945
V	3002	0.9933818663136995
C	2228	0.9915442812639075
Q	1447	0.9870395634379263
B	1408	0.9873772791023843
F	1307	0.9871601208459214
G	1225	0.9863123993558777
H	1138	0.9861351819757366
Z	875	0.9831460674157303
X	416	0.9651972157772621
J	394	0.9633251833740831
Y	376	0.9665809768637532
W	185	0.9390862944162437
K	128	0.927536231884058
---MEAN---	4862.231	0.985
--HIGHEST--	21325	0.999
--LOWEST--	100	0.928

Tabela 2 - Erros Absolutos e Relativos

Analisando a Tabela 1, podemos averiguar que o a frequência das letras é igual no Contador Exato e no Contador "Lossy Counter", relativamente a este útlimo algoritmo o threshold usado foi: 0.00000775. Já quanto ao Algoritmo de contagem aproximada, não se pode dizer o mesmo, a frequência das letras na mesma obra literária é bastante mais pequena comparado com os outros algoritmos.

Ainda nesta tabela, podemos fazer a comparação entre as três aproximações, a versão mais rápida é a de contagem exata, com 0.02 segundos, sendo também aquela que nos dá os resultados sempre corretos, o algoritmo que se segue é o algoritmo Lossy-count com 0.029 segundos, e por último o algoritmo de contagem aproximada com um tempo de execução de 0.085 segundos, sendo portanto este o algoritmo mais demorado, sendo aproximadamente 3 vezes mais lento em relação aos outros 2, e sendo excessivamente pior a nível de resultados comparado com os outros algoritmos.

As 3,5,10, letras mais frequentes no algoritmo Lossy-Count são respetivamente:

- E, A, T
- E, A, T, I, S
- E, A, T, I, S, R, N, U, L, O

Que comparando com o algoritmo de contagem exata, verifica-se que o Top 3, 5, e 10, são iguais aos do algoritmo Lossy-Count.

Vendo os erros quer relativos quer absolutos do algoritmo [3] de contagem aproximada comparando com os resultados do algoritmo de contagem exata, tal como já visto na tabela 1, a maioria do erro relatico de cada letra é sempre à volta de 99%. Como se pode ver na parte final é possível ver alguns cálculos relativamente aos erros, de entre os quais a média, que a nível absoluto, o algoritmo não conta cerca de 4862 em cada letra, representado um erro relativo de 98.5%.

B. Alice no País das Maravilhas em Alemão

Counters			
Letter	Exact Counter	Approximate Counter	Lossy Counter
E	21097	26	10557
N	11821	25	5915
I	11029	25	5518
A	8550	24	4278
R	8397	24	4203
S	8232	24	4119
T	7708	23	3858
H	7171	23	3587
D	6249	22	3126
U	5554	22	2779
C	5281	22	2643
L	4808	21	2406
G	4011	20	2007
O	3505	20	1757
M	3145	20	1574
W	2570	19	1286
F	2134	19	1067
B	2092	19	1048
K	1849	18	926
Z	1483	17	741
P	1164	17	583
V	771	17	386
J	280	13	140
Y	184	13	92
X	36	8	18
Q	34	7	17
--TIME--	0.020	0.084	0.027

Tabela 1 - Frequência das Letras

Absolute and Relative Error for Approximate Counter		
Letter	Absolute Error	Relative Error
E	21072	0.9988149973929943
N	11796	0.9978851197022248
R	11005	0.9978239187596337
A	8526	0.9971929824561403
I	8373	0.9971418363701322
S	8209	0.9972060252672498
L	7685	0.9970160871821484
H	7148	0.9967926370101798
U	6226	0.9963194111057769
C	5531	0.9958588404753331
D	5259	0.9958341223253172
T	4787	0.9956322795341098
G	3990	0.9947643979057592
B	3485	0.9942938659058488
Z	3125	0.9936406995230525
W	2550	0.9922178988326849
O	2114	0.9906279287722587
M	2072	0.9904397705544933
K	1830	0.989724175229854
F	1464	0.9871881321645314
P	1147	0.9853951890034365
V	755	0.9792477302204928
J	268	0.9571428571428572
Y	173	0.9402173913043478
X	28	0.7777777777777778
Q	26	0.7647058823529411
---MEAN---	4947.846	0.973
--HIGHEST--	21072	0.999
--LOWEST--	26	0.765

Tabela 2 - Erros Absolutos e Relativos

Analisando a Tabela 3, podemos averiguar que o a frequência das letras é diferente no Contador Exato e no Contador "Lossy Counter", relativamente a este útlimo algoritmo o threshold usado foi: 0.00000775.

Ainda nesta tabela, podemos fazer a comparação entre as três aproximações, a versão mais rápida é a de contagem exata, com 0.02 segundos, sendo também aquela que nos dá os resultados sempre corretos, o algoritmo que se segue é o algoritmo Lossy-count com 0.027 segundos, e por último o algoritmo de contagem aproximada com um tempo de execução de 0.084 segundos, sendo portanto este o algoritmo mais demorado, sendo aproximadamente 3 vezes mais lento em relação aos outros 2, e sendo excessivamente pior a nível de resultados comparado com os outros algoritmos.

As 3,5,10, letras mais frequentes no algoritmo Lossy-Count são respetivamente:

- E, N, I
- E, N, I, A, R
- E, N, I, A, R, S, T, H, D, U

Que comparando com o algoritmo de contagem exata, verifica-se que o Top 3, 5, e 10, são iguais aos do algoritmo Lossy-Count.

Vendo os erros quer relativos quer absolutos do algoritmo [3] de contagem aproximada comparando com os resultados do algoritmo de contagem exata, tal como já visto na tabela 3, a maioria do erro relatico de cada letra é sempre à volta de 99%. Como se pode ver na parte final é possível ver alguns cálculos relativamente aos erros, de entre os quais a média, que a nível absoluto, o algoritmo não conta cerca de 4947 em cada letra, representado um erro relativo de 97.3%.

C. Alice no País das Maravilhas em Inglês

Counters			
Letter	Exact Counter	Approximate Counter	Lossy Counter
E	9057	24	9057
T	5639	24	5639
I	5387	23	5387
A	5272	23	5272
O	4763	22	4763
N	4749	22	4749
S	4481	22	4481
R	4123	22	4123
L	4095	22	4095
D	3644	21	3644
H	2827	21	2827
U	2510	20	2510
G	2409	20	2409
C	2234	19	2234
M	1693	18	1693
W	1495	18	1495
P	1433	17	1433
Y	1427	17	1427
K	1158	17	1158
F	940	17	940
B	938	16	938
V	576	16	576
Q	209	13	209
X	148	12	148
J	98	11	98
Z	78	10	78
--TIME--	0.011	0.045	0.015

Tabela 3 - Frequência das Letras

Absolute and Relative Error for Approximate Counter		
Letter	Absolute Error	Relative Error
E	9031	0.9971292922601303
S	5616	0.995921262635219
T	5365	0.9959160943010952
N	5250	0.9958270106221547
D	4741	0.9953810623556582
H	4728	0.9955780164245104
A	4460	0.9953135460834636
R	4102	0.9949066213921901
O	4074	0.9948717948717949
G	3623	0.9942371020856202
K	2806	0.9925716307039264
C	2490	0.9920318725099602
I	2389	0.991697799916978
L	2214	0.991047448522829
U	1673	0.9881866509155346
W	1476	0.9872909698996656
M	1415	0.9874389392882066
P	1410	0.9880868955851436
Y	1141	0.9853195164075993
B	924	0.9829787234042553
F	922	0.9829424307036247
V	561	0.9739583333333334
Q	197	0.9425837320574163
X	136	0.918918918918919
J	88	0.8979591836734694
Z	70	0.8974358974358975
---MEAN---	2727.000	0.979
--HIGHEST--	9031	0.997
--LOWEST--	70	0.897

Tabela 4 - Tabela 4 - Erros Absolutos e Relativos

Analisando a Tabela 5, podemos averiguar que o a frequência das letras é igual no Contador Exato e no Contador "Lossy Counter", relativamente a este útlimo algoritmo o threshold usado foi: 0.00000775. Já quanto ao Algoritmo de contagem aproximada, não se pode dizer o mesmo, a frequência das letras na mesma obra literária é bastante mais pequena comparado com os outros algoritmos.

Ainda nesta tabela, podemos fazer a comparação entre as três aproximações, a versão mais rápida é a de contagem exata, com 0.011 segundos, sendo também aquela que nos dá os resultados sempre corretos, o algoritmo que se segue é o algoritmo Lossy-count com 0.015 segundos, e por último o algoritmo de contagem aproximada com um tempo de execução de 0.045 segundos, sendo portanto este o algoritmo mais demorado, sendo aproximadamente 3 vezes mais lento em relação aos outros 2, e sendo excessivamente pior a nível de resultados comparado com os outros algoritmos.

As 3,5,10, letras mais frequentes no algoritmo Lossy-Count são respetivamente:

- E, T, I
- E, T, I, A, O
- E, T, I, A, O, N, S, R, L, D

Que comparando com o algoritmo de contagem exata, verifica-se que o Top 3, 5, e 10, são iguais aos do algoritmo Lossy-Count.

Vendo os erros quer relativos quer absolutos do algoritmo [3] de contagem aproximada comparando com os resultados do algoritmo de contagem exata, tal como já visto na tabela 5, a maioria do erro relatico de cada letra é sempre à volta de 99%. Como se pode ver na parte final é possível ver alguns cálculos relativamente aos erros, de entre os quais a média, que a nível absoluto, o algoritmo não conta cerca de 2727 em cada letra, representado um erro relativo de 97.9%.

D. Conclusões sobre a obra “Alice no País das Maravilhas” nas diferentes línguas.

Fazendo agora uma comparação entre as 3 diferentes línguas que foram previamente analisadas, podemos logo à partida verificar que a obra em Inglês tem sensivelmente metade das letras das restantes obras nas outras línguas analisadas.

A nível do erro absoluto médio é na versão em inglês, que este é menor, também muito em caso derivado ao número de letras ser sensivelmente metade tal como já referido no segmento anterior. Já relativamente ao erro relatico maior o mesmo é encontrado na obra em Francês.

O erro relativo menor é encontrado na obra em Alemão, e o maior nas obras em Alemão e em Francês sendo o memso de 0.999.

Podemos fazer algumas observações como: A letra mais utilizada nas 3 diferentes linguas é a letra “E”

E. Uma obra do filosofo Friedrich Nietzsche em Inglês.

Counters			
Letter	Exact Counter	Approximate Counter	Lossy Counter
E	27776	26	19325
S	16909	26	11699
I	16623	26	11473
N	16201	25	11317
T	15458	25	10699
A	15025	25	10508
R	14447	25	10085
O	13925	24	9627
L	10989	24	7643
D	7782	24	5439
C	7398	24	5091
U	7042	24	4911
H	6208	23	4313
M	6069	23	4237
P	6035	23	4116
G	5394	22	3799
Y	4539	22	3170
F	3416	22	2392
V	2717	21	1898
W	2661	19	1894
B	2367	19	1670
K	1498	18	1059
X	494	15	336
Q	361	15	245
J	260	14	178
Z	180	14	126
--TIME--	0.032	0.138	0.044

Tabela 5 - Frequência das Letras

Absolute and Relative Error for Approximate Counter		
Letter	Absolute Error	Relative Error
E	27749	0.9990279377880185
S	16882	0.998403217221598
I	16596	0.9983757444504602
N	16175	0.9983951607925436
T	15432	0.9983180230301462
R	15000	0.9983361064891847
A	14423	0.9983387554509586
O	13901	0.9982764811490126
D	10965	0.9978159978159978
L	7758	0.9969159599074788
C	7375	0.9968910516355772
G	7019	0.9967338824197671
W	6186	0.9964561855670103
U	6048	0.9965397923875432
H	6014	0.9965202982601491
M	5373	0.9961067853170189
Y	4518	0.9953734302709848
P	3396	0.9941451990632318
B	2697	0.9926389400073611
V	2641	0.9924840285606915
F	2348	0.9919729615547106
K	1480	0.9879839786381842
J	478	0.9676113360323887
Q	347	0.961218836565097
X	246	0.9461538461538461
Z	167	0.9277777777777778
---MEAN---	8123.615	0.989
--HIGHEST--	27749	0.999
--LOWEST--	100	0.928

Tabela 6 - Erros Absolutos e Relativos

Analisando a Tabela 7, podemos averiguar que o a frequência das letras não é igual no Contador Exato e no Contador “Lossy Counter”, relativamente a este útlimo algoritmo o threshold usado foi: 0.00000775, no entanto apresenta valores signigicativamente melhores que o algoritmo de contagem aproximada.

Ainda nesta tabela, podemos fazer a comparação entre as três aproximações, a versão mais rápida é a de contagem exata, com 0.032 segundos, sendo também aquela que nos dá os resultados sempre corretos, o algoritmo que se segue é o algoritmo Lossy-count com 0.044 segundos, e por último o algoritmo de contagem aproximada com um tempo de execução de 0.138 segundos, sendo portanto este o algoritmo mais demorado, sendo aproximadamente 3 vezes mais lento em relação aos outros 2, e sendo excessivamente pior a nível de resultados comparado com os outros algoritmos.

As 3,5,10, letras mais frequentes no algoritmo Lossy-Count são respetivamente:

- E, S, I
- E, T, I, N, T
- E, T, I, N, T, A, R, O, L, D

Que comparando com o algoritmo de contagem exata, verifica-se que o Top 3, 5, e 10, são iguais aos do algoritmo Lossy-Count.

Vendo os erros quer relativos quer absolutos do algoritmo [3] de contagem aproximada comparando com os resultados do algoritmo de contagem exata, tal como já visto na tabela 7, a maioria do erro relatico de cada letra é sempre à volta de 99%. Como se pode ver na parte final é possível ver alguns cálculos relativamente aos erros, de entre os quais a média, que a nível absoluto, o algoritmo não conta cerca de 8123 em cada letra, representado um erro relativo de 98.9%.

VI. CONCLUSÃO

Tal como proposto no início do trabalho foram implementados os 3 algoritmos de contagem mencionados e foi feita uma avaliação elaborada dos algoritmos a vários níveis.

O algoritmo Lossy-Count é tendencialmente bom, e será tão bom quanto menor o valor do threshold utilizado, no entanto o valor para uma comparação mais justa o valor foi fixo para todos os testes feitos de modo a ser possível assim comparar de igual forma os resultados obtidos.

Pelos resultados podemos confirmar que o algoritmo de pesquisa por aproximação é de facto mau quer a nível de retornar um bom resultado, quer a nível de tempo de execução comparando com os outros algoritomos, e o algoritmo Lossy-Count é na sua maioria bom, tanto melhor quanto menor for o valor do threshold tal como referido, mas mesmo quando não apresenta resultados iguais aos da contagem exata, apresenta resultados substancialmente superiores aos do algoritmo de contagem por aproximação.

VII. REFERÊNCIAS

[1] - https://www.geeksforgeeks.org/approximation-algorithms/

[2] - https://en.wikipedia.org/wiki/Lossy_Count_Algorithm

[3] - https://www.vedantu.com/maths/absolute-and-relative-error