Algoritmos Probabilísticos: Contagem Aproximada de Ocorrências – Palavras em Ficheiros de Texto

Ana Sofia Medeiros de Castro Moniz Fernandes | 88739

***Resumo* - O presente relatório tem como propósito a apresentação de uma solução para contar o número de ocorrências de palavras em ficheiros de texto (através de três diferentes contadores), no âmbito do segundo trabalho prático da Unidade Curricular Algoritmos Avançados, do Mestrado de Engenharia Informática. Ao longo do documento serão apresentados e explicados os detalhes da solução (encontrada com três diferentes contadores), bem como alguns testes realizados e as suas interpretações.**

***Abstract* - The present report aims to present a solution to count the number of words present in a text file, in the context of the first practical work of the course Algoritmos Avançados of the Informatics Engineer Master's Degree. Over the document, it will be presented and explained all the constituent details of the solution (found by using three different counters), as well as some performed tests and its interpretations.**

I. INTRODUÇÃO

No segundo trabalho prático da Unidade Curricular de Algoritmos Avançados, foi proposta aos alunos a escolha de um tema, de entre três.

A escolha aqui apresentada recai sobre o tema "Contagem Aproximada de Ocorrências - Palavras em Ficheiros de Texto". Assim sendo, foram desenvolvidos e testados três contadores, indicados pelo docente - um contador exato, um contador aproximado com probabilidade ½ e um contador logarítmico de base 2. Além disso, foi realizada uma análise da eficiência e das limitações dos contadores desenvolvidos.

II. TIPOS DE CONTADORES

Como mencionado anteriormente, foram implementados três tipos de contadores - um contador exato, um contador aproximado com probabilidade ½ e um contador logarítmico de base 2. Assim deve-se, numa primeira fase, começar por se entender as diferenças entre cada um dos três.

A. Contador Exato

Tal como o próprio nome indica, um contador exato é aquele capaz de encontrar a solução exata para o problema. Por exemplo, caso queiramos contar o número de palavras distintas presentes na lista ["algoritmos","avancados","avancados"], a solução do contador exato será sempre {"algoritmos":1, "avancados":2}.

B. Contador aproximado com probabilidade ½

Neste tipo de contador, ao ser encontrada uma palavra (a ocorrência de uma palavra é, então, um evento), o contador em causa é incrementado em uma unidade, com uma probabilidade de ½.

C. Contador logarítmico de base 2

Neste contador, também conhecido como contador aproximado com probabilidade decrescente, à medida que o valor do contador aumenta, vai sendo incrementado com uma probabilidade mais pequena [1]. Ou seja:

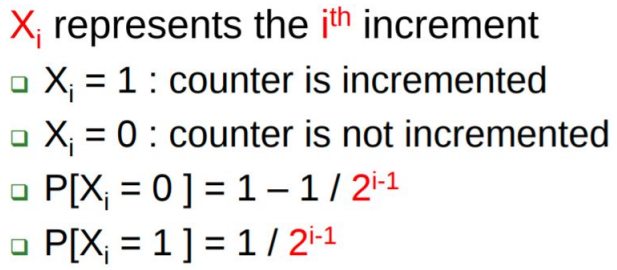

Fig. 1 - Forma como o contador logarítmico de base 2 é incrementado [1]

III. ANÁLISE DOS RESULTADOS

Por forma a comparar os desempenhos dos três contadores, foram criados quatro "scripts" - um que permite repetir a experiências um certo número de vezes (contando as palavras da obra original) e três exemplos, sendo que cada exemplo é referente a um idioma:

- RepeatingExperience.py: repete, 10 vezes para cada contador, a contagem das palavras da obra original - o livro "Hamlet", na língua inglesa;
- ExampleENG.py: os contadores vão realizar as suas contagens para o livro "Hamlet", na versão escrita em inglês;
- ExampleFR.py: os contadores vão realizar as suas contagens para o livro "Hamlet", na versão escrita em francês;
- ExamplePT.py: os contadores vão realizar as suas contagens para o livro "Hamlet", na versão escrita em português.

É de notar que são excluídas das contagens palavras que sejam de comprimento inferior a três e que contenham caracteres que não letras.

A. Funcionamento dos contadores - 10 experiências

Para se iniciar a análise do funcionamento de cada contador, a contagem foi repetida 10 vezes (para cada contador), com o ficheiro original ("Hamlet", na língua inglesa):

Repetition	Max value	Min value	Mean value	Exec.time
1	that : 351	paces : 1	3.68459	0.003
2	that : 351	paces : 1	3.68459	0.002
3	that : 351	paces : 1	3.68459	0.002
4	that : 351	paces : 1	3.68459	0.002
5	that : 351	paces : 1	3.68459	0.002
6	that : 351	paces : 1	3.68459	0.002
7	that : 351	paces : 1	3.68459	0.002
8	that : 351	paces : 1	3.68459	0.002
9	that : 351	paces : 1	3.68459	0.002
10	that : 351	paces : 1	3.68459	0.002

Tabela 1 - Repetição da experiência 10 vezes, para a obra original, com o contador exato

Repetition	Max value	Min value	Mean value	Exec.time
1	that : 174	approaches : 0	1.84245	0.007
2	that : 168	paces : 0	1.85597	0.007
3	that : 173	approaches : 0	1.83679	0.007
4	that : 186	paces : 0	1.81887	0.007
5	that : 168	carefully : 0	1.83491	0.009
6	that : 176	paces : 0	1.87075	0.007
7	that : 152	paces : 0	1.85063	0.007
8	that : 174	platform : 0	1.82327	0.007
9	that : 174	platform : 0	1.83113	0.007
10	that : 175	carefully : 0	1.87767	0.007

Tabela 2 - Repetição da experiência 10 vezes, para a obra original, com o contador de probabilidade fixa de ½

Repetition	Max value	Min value	Mean value	Exec.time
1	your : 9	paces : 0	0.89748	0.009
2	this : 11	paces : 0	0.89748	0.01
3	that : 9	paces : 0	0.88994	0.014
4	that : 11	paces : 0	0.89874	0.011
5	will : 9	paces : 0	0.89874	0.018
6	this : 9	paces : 0	0.88994	0.012
7	your : 10	paces : 0	0.90314	0.01
8	your : 10	paces : 0	0.89403	0.009
9	your : 10	paces : 0	0.88553	0.009
10	with : 10	paces : 0	0.88616	0.009

Tabela 3 - Repetição da experiência 10 vezes, para a obra original, com o contador logarítmico de base 2

Através da análise das três tabelas, é possível observar-se que:

• No contador de probabilidade fixa de ½, independentemente da experiência, a palavra com maior contagem é sempre a mesma que a do contador exato, neste caso "that" (Tabelas 1 e 2);

• Apesar do contador de probabilidade fixa de ½ acertar sempre na palavra com maior número de ocorrências, em seis das experiências (1, 3, 5, 8, 9 e 10) não foi capaz de encontrar a palavra correta com menor contagem, que é "paces" (Tabela 2);

• No contador logarítmico de base 2, a palavra com maior contagem varia bastante - apenas conseguiu encontrar a correta em duas das experiências (3 e 4) (Tabela 3);

• Apesar de apresentar uma grande variação no cálculo da contagem da palavra com mais ocorrências, o contador logarítmico foi sempre capaz de encontrar a palavra que tem, efetivamente, um menor número de ocorrências (Tabela 3);

• As contagens feitas pelo contador logarítmico (Tabela 3) são muito mais pequenas que as contagens reais (Tabela 1), e que contagens realizadas pelo contador de probabilidade fixa (Tabela 2);

• O valor médio de palavras contadas ("mean value") pelo contador logarítmico de base 2, em cada experiência, é inferior ao valor médio contabilizado pelo contador de probabilidade fixa de ½ - isto deve-se ao facto de, como já mencionado, o contador logarítmico incrementar com uma probabilidade mais pequena, à medida que o valor do contador aumenta.

Seguidamente, para cada repetição, registaram-se em duas tabelas os erros relativos para cada um dos contadores probabilísticos. É de notar que, para o cálculo destes erros, foi utilizado o valor obtido com o contador exato como valor real.

Repetition	Max rel error	Min rel error	Mean rel error	Standard deviation
1	100	0	50.062	6.829
2	100	0	49.703	6.894
3	100	0	51.372	6.846
4	100	0	50.09	6.819
5	100	0	50.504	6.824
6	100	0	50.045	6.949
7	100	0	49.025	6.703
8	100	0	51.418	6.843
9	100	0	49.903	6.773
10	100	0	48.257	6.907

Tabela 4 - Erros relativos em cada repetição da experiência, com a obra original, com o contador de probabilidade fixa de ½

Repetition	Max rel error	Min rel error	Mean rel error	Standard deviation
1	100	16.667	81.194	1.526
2	100	20	81.128	1.526
3	100	20	81.198	1.502
4	100	20	81.249	1.541
5	100	16.667	81.054	1.514
6	100	20	81.287	1.511
7	100	16.667	81.024	1.531
8	100	20	81.248	1.52
9	100	16.667	81.306	1.502
10	100	20	81.331	1.504

Tabela 5 - Erros relativos em cada repetição da experiência, com a obra original, com o contador logarítmico de base 2

• Pela análise dos erros relativos, pode concluir-se que o contador de probabilidade fixa apresenta sempre um erro relativo mínimo de 0 (Tabela 4), enquanto que o logarítmico não é capaz de o apresentar em nenhuma das experiências (Tabela 5). Já o erro relativo máximo é igual em ambos;

• Relativamente ao erro relativo médio, o contador logarítmico de base 2 é quem apresenta maior valor para todas as experiências, mostrando assim que não é tão preciso quanto o contador exato, no que toca à contagem de ocorrências de cada palavra;

• Quanto ao desvio padrão do (última coluna das tabelas 4 e 5), pode concluir-se que o valor do desvio padrão do contador logarítmico é bastante mais pequeno (Tabela 5) que o do contador de probabilidade fixa (Tabela 4), indicando assim que as contagens realizadas pelo contador logarítmico se encontram próximas do valor médio das mesmas (Tabela 3, coluna "mean value");

• O facto de o contador de probabilidade fixa apresentar um desvio padrão elevado não é necessariamente mau - pode apenas refletir a existência duma grande variação nas

contagens, o que vai de encontro com o analisado previamente neste relatório.

B. Comparação das palavras mais frequentes em cada idioma da obra "Hamlet"

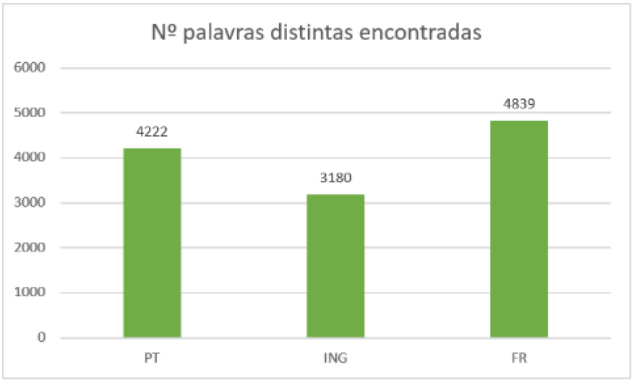

Fig. 2 - Número de palavras distintas encontradas, para cada idioma

É de referir que, em cada idioma, foi contabilizado um número diferente de palavras (Figura 2).

Por forma a poderem obter-se mais conclusões, foram contabilizadas as 20 palavras mais contadas por cada contador, em cada idioma. Os resultados dos contadores probabilísticos são postos lado a lado com os resultados do contador exato, de maneira a poderem retirar-se conclusões acerca da precisão dos mesmos. Analisem-se os resultados das seguintes tabelas:

• Português:

Comparasion between exact counter and counter with probability 1/2

Word	Exact counting	Prob 1/2 counting	Pos. in prob 1/2 counter	Counting diff	Top 20 of prob 1/2 counter?
hamlet	406	190	0	216	True
para	236	122	1	114	True
mais	175	97	2	78	True
como	164	66	4	98	True
minha	139	66	3	73	True
horacio	121	65	5	56	True
rainha	101	46	7	55	True
polonio	100	53	6	47	True
está	79	33	14	46	True
laerte	73	32	16	41	True
porque	68	37	8	31	True
rosencrantz	66	35	11	31	True
quem	63	29	17	34	True
esta	62	33	15	29	True
ophelia	62	26	21	36	False
seus	60	36	10	24	True
quando	58	34	12	24	True
este	57	36	9	21	True
suas	56	34	13	22	True
primeiro	52	27	19	25	True

Tabela 6 - Comparação das 20 palavras mais contadas em português com o contador exato e o contador de probabilidade fixa de ½

Comparasion between exact counter and counter with log base 2

Word	Exact counting	Prob log 2 counting	Pos. in prob log 2 counting	Counting diff	Top 20 of prob log 2?
hamlet	406	8	9	398	True
para	236	9	0	227	True
mais	175	7	17	168	True
como	164	7	19	157	True
minha	139	8	5	131	True
horacio	121	8	2	113	True
rainha	101	7	25	94	False
polonio	100	8	11	92	True
está	79	8	3	71	True
laerte	73	8	10	65	True
porque	68	8	7	60	True
rosencrantz	66	6	53	60	False
quem	63	7	13	56	True
esta	62	8	4	54	True
ophelia	62	7	30	55	False
seus	60	7	22	53	False
quando	58	6	41	52	False
este	57	8	8	49	True
suas	56	7	24	49	False
primeiro	52	7	12	45	True

Tabela 7 - Comparação das 20 palavras mais contadas em português com o contador exato e o contador logarítmico de base 2

• Inglês:

Comparasion between exact counter and counter with probability 1/2

Word	Exact counting	Prob 1/2 counting	Pos. in prob 1/2 counter	Counting diff	Top 20 of prob 1/2 counter?
that	351	177	0	174	True
with	275	143	1	132	True
this	254	102	3	152	True
your	250	126	2	124	True
what	190	97	4	93	True
have	176	96	5	80	True
will	151	79	6	72	True
shall	109	52	7	57	True
they	98	44	11	54	True
good	97	51	8	46	True
thou	97	49	9	48	True
from	94	39	13	55	True
most	82	45	10	37	True
would	75	38	14	37	True
more	75	40	12	35	True
like	74	30	24	44	False
enter	69	36	15	33	True
very	64	36	16	28	True
hath	63	33	17	30	True
such	61	29	25	32	False

Tabela 8 - Comparação das 20 palavras mais contadas em inglês com o contador exato e o contador de probabilidade fixa de ½

Comparasion between exact counter and counter with log base 2

Word	Exact counting	Prob log 2 counting	Pos. in prob log 2 counting	Counting diff	Top 20 of prob log 2?
that	351	10	3	341	True
with	275	10	2	265	True
this	254	10	1	244	True
your	250	10	0	240	True
what	190	8	8	182	True
have	176	8	5	168	True
will	151	8	7	143	True
shall	109	8	13	101	True
they	98	8	12	90	True
good	97	7	18	90	True
thou	97	8	9	89	True
from	94	7	22	87	False
most	82	7	16	75	True
would	75	8	10	67	True
more	75	8	11	67	True
like	74	7	25	67	False
enter	69	7	14	62	True
very	64	7	28	57	False
hath	63	6	37	57	False
such	61	5	82	56	False

Tabela 9 - Comparação das 20 palavras mais contadas em inglês com o contador exato e o contador logarítmico de base 2

• Francês:

Comparasion between exact counter and counter with probability 1/2

Word	Exact counting	Prob 1/2 counting	Pos. in prob 1/2 counter	Counting diff	Top 20 of prob 1/2 counter?
vous	441	219	0	222	True
dans	354	180	1	174	True
pour	327	169	2	158	True
nous	249	115	4	134	True
plus	220	116	3	104	True
comme	204	105	5	99	True
votre	197	104	6	93	True
mais	187	102	7	85	True
avec	167	67	9	100	True
cette	159	69	8	90	True
tout	141	56	12	85	True
fait	116	59	11	57	True
bien	113	51	14	62	True
sont	102	62	10	40	True
cela	97	55	13	42	True
même	94	46	16	48	True
sans	87	41	18	46	True
notre	85	49	15	36	True
elle	82	37	23	45	False
aussi	80	40	20	40	False

Tabela 10 - Comparação das 20 palavras mais contadas em francês com o contador exato e o contador de probabilidade fixa de ½

Comparasion between exact counter and counter with log base 2

Word	Exact counting	Prob log 2 counting	Pos. in prob log 2 counting	Counting diff	Top 20 of prob log 2?
vous	441	11	0	430	True
dans	354	10	3	344	True
pour	327	10	1	317	True
nous	249	9	5	240	True
plus	220	10	2	210	True
comme	204	9	7	195	True
votre	197	8	8	189	True
mais	187	8	16	179	True
avec	167	7	22	160	False
cette	159	9	4	150	True
tout	141	7	25	134	False
fait	116	9	6	107	True
bien	113	8	12	105	True
sont	102	8	11	94	True
cela	97	7	28	90	False
même	94	6	51	88	False
sans	87	8	14	79	True
notre	85	6	49	79	False
elle	82	6	47	76	False
aussi	80	6	67	74	False

Tabela 11 - Comparação das 20 palavras mais contadas em francês com o contador exato e o contador logarítmico de base 2

• Quando comparados ao contador exato, é o contador de probabilidade fixa ½ o mais preciso - conta um número de ocorrências mais próximo do número real e as palavras presentes no seu "top 20" diferem pouco, independentemente do idioma (Tabelas 6, 8 e 10);

• O contador logarítmico de base 2, apesar de contar muito menos ocorrências do que aquelas que realmente acontecem, acaba por conseguir ter um top 20 de palavras bastante idêntico ao do exato (ainda assim, não tão idêntico quanto o contador de probabilidade fixa). Isto deve-se ao facto de, como já mencionado anteriormente, este contador incrementar com uma probabilidade mais pequena à medida que o número de ocorrências aumenta;

• As palavras com mais ocorrências diferem bastante (não apresentam o mesmo significado) entre cada idioma e para cada contador - por exemplo, para a língua portuguesa a palavra mais contada (com o contador exato) é "hamlet" (Tabela 6), enquanto que essa palavra não consta, sequer, no top 20 das palavras em inglês (Tabela 8).

IV. CONCLUSÃO

Tendo em conta o estudado neste relatório, é possível concluir que:

• Os contadores probabilísticos permitem ter um pequeno contador, por forma a manter contagens aproximadas de números bastante grandes;

• De entre os dois contadores probabilísticos analisados, o mais preciso é o contador de probabilidade fixa ½;

• Apesar de apresentar grandes discrepâncias de contagem, o contador logarítmico de base 2 é aquele que vai permitir usarem-se contadores mais pequenos quando se está a lidar com grandes volumes de informação, mantendo a ordem das palavras mais contadas relativamente idêntica à ordem real.

REFERÊNCIAS

[1] J. Madeira, "Probabilistic Counters", Aveiro, 2020
[2] R. Morris, "Counting Large Numbers of Events in Small Registers", Bell Laboratories, Murray Hill, N.J., 1978