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 recaiu 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 computacional 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. Ou seja,

Isto permite fazer contagens de grandes números usando pequenos contadores.

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 [REF. STOR]. Ou seja:

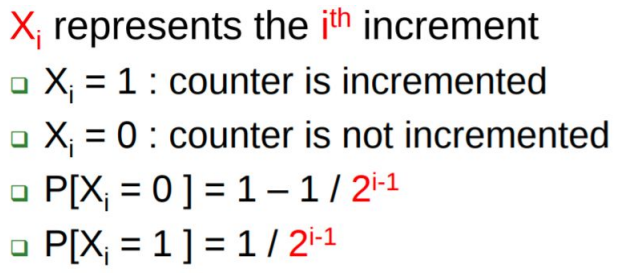

Fig. 1 - Forma como o contador logarítmico de base 2 é incrementado [REF. STOR]

III. ANÁLISE DOS RESULTADOS

Por forma a comparar os desempenhos dos três contadores, foram criados três exemplos, sendo que cada exemplo é referente a um idioma:

- 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 sejam letras.

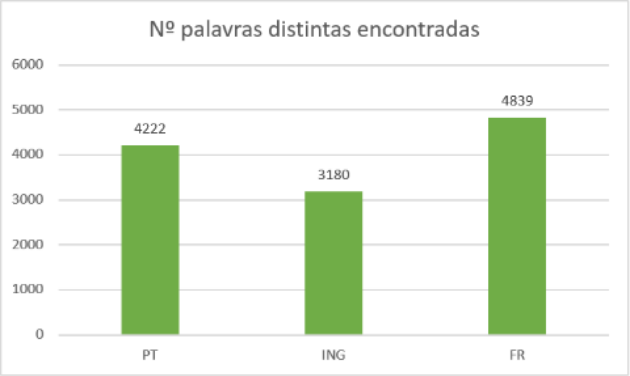

Fig. 2 - Número de palavras distintas encontradas, para cada idioma

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

A. Tempo de execução para cada contador

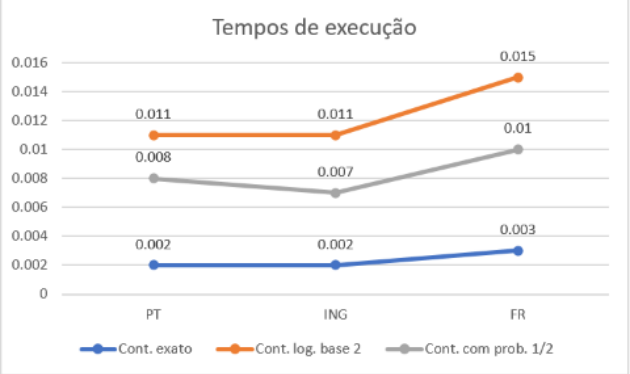

Fig. 3 - Tempo de execução, para cada contador e em cada idioma

Como seria de esperar, os tempos de execução (Figura 3) são maiores para o idioma que apresenta um maior número de palavras, independentemente do contador em questão.

Ao compararem-se os três diferentes contadores (Figura 3) , é possível concluir que o contador logarítmico de base 2 é o que demora mais tempo a concluir a sua contagem - isto deve-se ao facto de ser sempre calculada uma nova probabilidade à medida que uma nova palavra ocorre (Figura 1).

Desta forma, o contador exato será sempre o mais rápido (pois apenas incrementa o seu contador, não calculando qualquer probabilidade). Comparando os dois contadores probabilísticos, devido ao mencionado no parágrafo anterior, pode concluir-se que o contador de probabilidade fixa será sempre mais rápido que o contador logarítmico, visto que tem sempre a probabilidade definida como ½.

B. Palavras com maior ocorrência

Por forma a ter-se noção da precisão de cada contador, foram contabilizadas as 20 palavras mais contadas por cada um. Os resultados dos contadores probabilísticos são postos lado a lado com os resultados do contador exato, de maneira a poder retirar conclusões acerca da precisão dos mesmos. Além disso, as experiências foram repetidas três vezes, mas as diferenças entre as mesmas não foram notórias, pelo que neste relatório serão apenas analisados os resultados de uma das experiências (a mais recente).

```
--- Top 20 words - exact counter:
hamlet -> 406
para -> 236
mais -> 175
como -> 164
minha -> 139
horacio -> 121
rainha -> 101
polonio -> 100
está -> 79
laerte -> 73
porque -> 68
rosencrantz -> 66
quem -> 63
esta -> 62
ophelia -> 62
seus -> 60
quando -> 58
este -> 57
suas -> 56
primeiro -> 52
```

Fig. 4 - 20 palavras mais contadas com o contador exato, na língua portuguesa

```
--- Top 20 words - exact counter:
that -> 351
with -> 275
this -> 254
your -> 250
what -> 190
have -> 176
will -> 151
shall -> 109
they -> 98
good -> 97
thou -> 97
from -> 94
most -> 82
would -> 75
more -> 75
like -> 74
enter -> 69
very -> 64
hath -> 63
such -> 61
```

Fig. 5 - 20 palavras mais contadas com o contador exato, na língua inglesa

```
--- Top 20 words - exact counter:
vous -> 441
dans -> 354
pour -> 327
nous -> 249
plus -> 220
comme -> 204
votre -> 197
mais -> 187
avec -> 167
cette -> 159
tout -> 141
fait -> 116
bien -> 113
sont -> 102
cela -> 97
même -> 94
sans -> 87
notre -> 85
elle -> 82
aussi -> 80
```

Fig. 6 - 20 palavras mais contadas com o contador exato, na língua francesa

```
--- Top 20 words - counter with log base 2:
para -> 10
hamlet -> 9
rainha -> 9
horacio -> 8
está -> 8
como -> 8
porque -> 8
então -> 7
marcello -> 7
mesmo -> 7
minha -> 7
mais -> 7
quando -> 7
pois -> 7
todos -> 7
seus -> 7
quanto -> 7
seja -> 7
laerte -> 7
polonio -> 7
```

Fig. 7 - 20 palavras mais contadas com o contador logarítmico de base 2, na língua portuguesa

```
--- Top 20 words - counter with log base 2:
what -> 10
this -> 9
will -> 9
with -> 9
that -> 9
than -> 9
hath -> 8
from -> 8
where -> 8
they -> 8
shall -> 8
enter -> 7
most -> 7
your -> 7
have -> 7
good -> 7
them -> 7
think -> 7
give -> 7
there -> 7
```

Fig. 8 - 20 palavras mais contadas com o contador logarítmico de base 2, na língua inglesa

```
--- Top 20 words - counter with log base 2:
vous -> 10
nous -> 10
dans -> 10
mais -> 9
premier -> 8
cette -> 8
encore -> 8
avec -> 8
bien -> 8
avait -> 8
fait -> 8
pour -> 8
plus -> 8
comme -> 8
soit -> 8
trop -> 8
aussi -> 8
votre -> 7
faire -> 7
elle -> 7
```

Fig. 9 - 20 palavras mais contadas com o contador logarítmico de base 2, na língua francesa

```
--- Top 20 words - counter with probability 1/2:
hamlet -> 213
para -> 119
mais -> 88
como -> 84
minha -> 76
horacio -> 60
polonio -> 53
laerte -> 40
porque -> 36
está -> 35
rainha -> 35
rosencrantz -> 35
seus -> 33
este -> 31
suas -> 31
esta -> 30
ophelia -> 30
primeiro -> 28
vossa -> 28
quem -> 27
```

Fig. 10 - 20 palavras mais contadas com o contador de probabilidade fixa ½, na língua portuguesa

```
--- Top 20 words - counter with probability 1/2:
that -> 178
with -> 138
this -> 120
your -> 118
what -> 89
have -> 84
will -> 84
they -> 61
shall -> 60
from -> 55
good -> 43
thou -> 43
give -> 37
would -> 37
most -> 36
more -> 35
when -> 33
such -> 31
must -> 31
their -> 31
```

Fig. 11 - 20 palavras mais contadas com o contador de probabilidade fixa ½, na língua inglesa

```
--- Top 20 words - counter with probability 1/2:
vous -> 228
dans -> 188
pour -> 148
nous -> 135
plus -> 115
comme -> 103
mais -> 93
votre -> 91
cette -> 88
avec -> 82
tout -> 65
bien -> 56
fait -> 55
sont -> 50
même -> 50
aussi -> 48
notre -> 45
cela -> 44
elle -> 41
sans -> 41
```

Fig. 12 - 20 palavras mais contadas com o contador de probabilidade fixa ½, na língua francesa

Daqui, é possível inferir que:

• 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 que ocorrem são bastante idênticas às que aparecem no contador exato;

• A palavra com mais ocorrências é a mesma, em cada idioma, tanto no contador exato (Figuras 4, 5 e 6) como no contador de probabilidade fixa ½ (Figuras 10,11,12);

• 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. 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 têm 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" (Figura 4), enquanto que essa palavra nem consta no top 20 das palavras em inglês (Figura 5).

C. Erros relativos

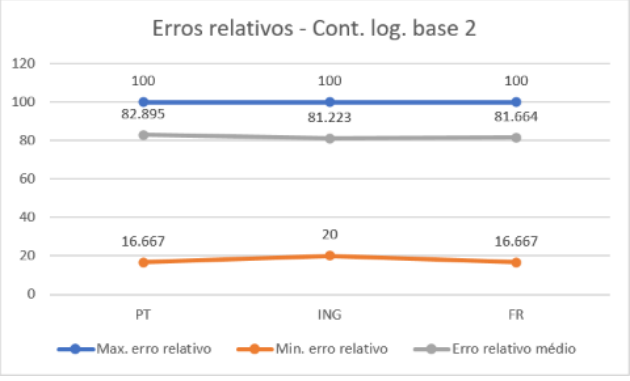

Fig. 13 - Erro relativo máximo, mínimo e médio em percentagem do contador logarítmico de base 2, para cada idioma/exemplo

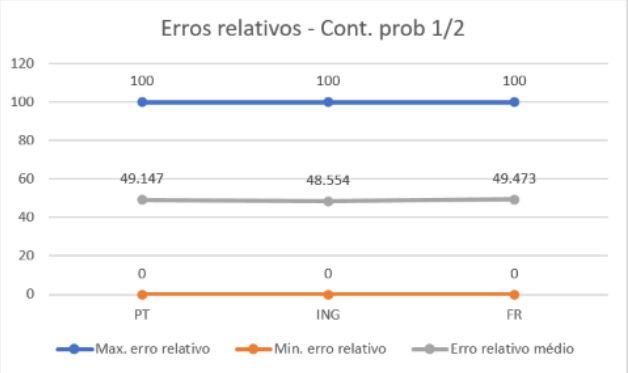

Fig. 14 - Erro relativo máximo, mínimo e médio em percentagem do contador de probabilidade fixa ½, para cada idioma/exemplo

A análise dos erros relativos vai permitir uma melhor percepção sobre a precisão de cada contador. É de notar que, para o cálculo destes erros, foi utilizado o valor obtido com o contador exato como valor real. Desta forma, nota-se que:

• O contador logarítmico de base 2, em qualquer que seja o idioma, não é capaz de ter uma única palavra com a contagem igual à real - nunca apresenta um erro relativo mínimo de 0 (Figura 13);

• O contador logarítmico de base 2, em qualquer que seja o idioma, tem pelo menos uma contagem a diferir em 100% da contagem real (o erro máximo é de 100%) (Figura 13);

• O contador de probabilidade fixa ½ apresenta, para cada idioma, um erro relativo mínimo de 0 (Figura 14), o que indica que tem pelo menos uma contagem igual à contagem real ;

• O contador de probabilidade fixa ½ apresenta, para cada idioma, um erro relativo máximo de 100% (Figura 14), o que indica que tem pelo menos uma contagem a diferir em 100% da contagem real;

• O erro relativo médio do contador de probabilidade fixa ½ (Figura 14) é bastante menor que o erro relativo médio do contador logarítmico de base 2 (Figura 13), o que indica que o primeiro apresenta uma maior precisão.

VI. 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 rápido é o contador de probabilidade fixa ½ - o contador logarítmico acaba por ser mais lento devido ao facto de calcular sempre uma nova probabilidade, pois esta não é um valor fixo;

• 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