Letras mais frequentes - Contadores
Projeto 3

Vasco Vieira Costa - 97746

***Abstract* –A study into the counters of Csuros by Miklós Csuros and of Space-Saving by Ahmed Metwally, Divyankant Agrawald and Amr El Abbadi in a set of characters taken from the compilation of some of the work by Luís Vaz de Camões and Eça de Queirós. Making a brief study on the statistics related to the probabilistic counter and comparing the obtained counter values with the exact ones. Analyzing the most frequent letters from this technique in combination with the Space-Saving algorithm.**

From here it was possible to draw conclusions related to these counters and verify their utility in the most diverse of situations.

***Resumo* –Estudo dos contadores de Csuros de Miklós Csuros e de *Space-Saving* de Ahmed Metwally, Divyankant Agrawald e Amr El Abbadi para um conjunto de caracteres vindo da compilação de algumas das obras de Luís Vaz de Camões e de Eça de Queirós. Efetuando um breve estudo da estatística relativa ao contador probabilístico e comparando os valores obtidos com os exatos. Analisando a ordem das letras mais frequentes que aqui se obtém assim como do algoritmo de *Space-Saving*.**

Daqui foi possível retirar conclusões relativamente a estes contadores e verificar que estes apresentam utilidade nas mais diversas situações.

***Keywords* –Counters, Csuros, Space-Saving, Most Frequent Letters, Probabilistic Counters**

***Palavras chave* –Contadores, Csuros, Space-Saving, Letras Mais Frequentes, Contadores Probabilísticos**

I. Introdução

Com este trabalho pretendem-se estudar diferentes formas de efetuar a identificação das letras mais frequentes em ficheiros de texto, avaliando o desempenho das estimativas. Tomando três abordagens, uma exata, uma aproximada/probabilística e um algoritmo para verificar quais as que são mais frequentes.

O uso de contadores aproximados foi um tópico primeiramente explorado por Robert Morris. Este pretendia contar até valores bastante elevados sem a necessidade de ocupar grande memória. Obtendo valores aproximados para as contagens com um erro máximo quantificável [1].

O crescimento do uso de memória *flash* devido ás suas inúmeras vantagens face a outros dispositivos de armazenamento tradicionais originou também a necessidade do desenvolvimento de técnicas de contagem. Mais em concreto do número de vezes de escrita em cada um dos blocos para se poder aumentar a longevidade do dispositivo [2]. Também no estudo de redes, a contagem de triângulos é essencial e facilmente se torna computacionalmente difícil. Onde sistemas paralelos e de contagens aproximadas se tornam uma necessidade [3]. Para além do estudo de redes existem também muitas outras áreas em que estes contadores se tornam imprescindíveis devido à grande quantidade de dados a trabalhar.

Implementação

Assim, de forma a analisar os diferentes contadores foi proposto que se efetua-se a contagem das letras mais frequentes em diferentes conjuntos de textos. Para isto utilizaram-se os ficheiros disponíveis em *Project Gutenberg*. Compilando os textos de Luís Vaz de Camões, *Os Lusíadas* [4], *A Morte de D. Ignez de Castro* [5], *Obras Completas de Luis de Camões, Tomo II* [6] e *Obras Completas de Luis de Camões, Tomo III* [7]. Adicionalmente também foi explorada esta combinação com a inclusão de algumas das obras de Eça de Queirós, *A Relíquia* [8] e *Os Maias: episodios da vida romantica* [9].

No código incluído, `clean_text.py`, é feita a leitura dos ficheiros originais, disponíveis em *Project Gutenberg*, e é feita a sua transformação. Removendo caracteres de pontuação, acentos e espaços, transformando **ç** em **c** e colocando todas as letras em maiúsculas. Unindo os ficheiros no fim para se obter uma sequência longa de caracteres.

No ficheiro de código `code.py` são apresentados os diferentes contadores com um destes a ser o exato, `exact_counter`. Este é o mais básico, havendo duas implementações uma para verificar a ocorrência de caracteres estranhos e a outra, mais eficaz, em que se inicia um dicionário a 0 para cada uma das letras e se contam todas as ocorrências. Podendo-se assim comparar os resultados de outros contadores a este método exato.

II. Contador de Csuros

O contador de Csuros foi apresentado por Miklós Csuros em [10]. Em que descreve uma técnica para contar até n utilizando $\log\log n + O(1)$ *bits*, partindo do princípio do contador de Morris.

Utilizando um parâmetro $M = 2^d$ onde d é um número natural, sendo este parâmetro o que define o que se pretende favorecer, poupança de memória, menor valor de d ou precisão de contagem, um em desfavorecimento do outro. Para se efetuar um aumento neste contador que se encontra com o valor X primeiramente define-se um valor $t = \lfloor X/M \rfloor$ e enquanto este for superior a 0 efetua-se um processo de seleção aleatória seguindo uma distribuição uniforme entre duas variáveis. Ocorrendo uma destas variáveis o contador não sofre alteração, dando-se a outra então o valor de t é reduzido em 1 unidade e o contador é aumentado em 1.

Assim, o valor do contador será $X = 2^d \cdot t + u \Leftrightarrow u = X - 2^d \cdot t$, onde u retrata os d *bits* mais baixos que são utilizados para contabilizar o valor real, $f(X)$, igual a $f(X) = (M+u)\cdot 2^t - M$. Deste modo, o contador armazena X utilizando d+log log $n + O(1)$ *bits*. O desvio padrão desta estimativa pode ser medido com $\frac{c}{\sqrt{Mn}}$, onde c é um valor entre 0.58 e 0.61, assintoticamente.

Em `csuros` efetua-se uma implementação de um contador para cada uma das letras. Dado que se pretendem efetuar as contagens de todas as letras, inicia-se um dicionário, `dic`, com cada letra a ser a chave e o valor de 0 associado ao valor do contador. Fazendo-se uso de outro dicionário, suplementar, `t` para armazenar os valores de t apresentados anteriormente. No final efetua-se a conta que permite obter o valor de $f(X)$ sendo que devolvem-se os 3 conjuntos, X,t e $f(X)$ por esta mesma ordem.

III. Space-Saving

Este algoritmo determinístico foi proposto por Ahmed Metwally, Divyankant Agrawald e Amr El Abbadi em [11], partindo do princípio de se monitorizarem apenas m elementos. Caso já se esteja a monitorizar um elemento, e, este ao surgir o seu contador é simplesmente incrementado. Na situação de e não ser monitorizado então o contador com menos contagens estimadas passa a contar e e é incrementado em uma unidade, encontrando-se a implementação em `space_saving`.

Desta forma a memória necessária é da ordem de $O(k)$ em que k representa os pares, valor do contador e item a contar. Algumas das propriedades esperadas será que o valor mínimo de um dos contadores será menor ou igual a $\lfloor \frac{N}{m} \rfloor$ assumindo que existem mais que m elementos distintos no conjunto de N dados. Retratando assim o limite superior na sobre-estimação efetuada. Também se verifica que um elemento a analisar com frequência superior a esta é garantido que é contabilizado [11], [12].

A implementação deste método faz uso de um dicionário em que se inserem e se retiram as chaves consoante o elemento e a ser monitorizado. Para a escrita no ficheiro de análise são adicionados os valores de 0 ás chaves que não são monitorizadas para apenas se ter uma forma padrão para se poder analisar mais facilmente.

IV. Análise da compilação das obras de Luís Vaz de Camões e Eça de Queirós

Primeiramente podem-se verificar quais as letras mais frequentes neste ficheiro da combinação de todos os textos. Na Fig. 1 é possível verificar quais as letras mais frequentes e a diferença relativa que estas apresentam. Verificando-se como se poderia até especular que a letra mais frequente seria a letra **A**. Seguindo-se as letra **E** e **O**, ou seja vogais. As menos frequentes são as que não fazem parte do alfabeto português original, **W**, **K** e **Y**. Estas não foram confirmadas se de facto foram um erro de limpeza dos ficheiros. Dado que não se ficou apenas com o início da redação de cada texto, ficaram ainda porções de texto relativas ao livro do qual foi retirado. Removendo-se assim do inicio e do fim de cada texto a informação e texto legal relativo ao *Project Gunteberg*, e não do texto em si que está disponível. Também de notar que os pares de letras **D** e **N**, **U** e **T**, **Q** e **G** e **B** e **F** apresentam valores muito próximos entre si.

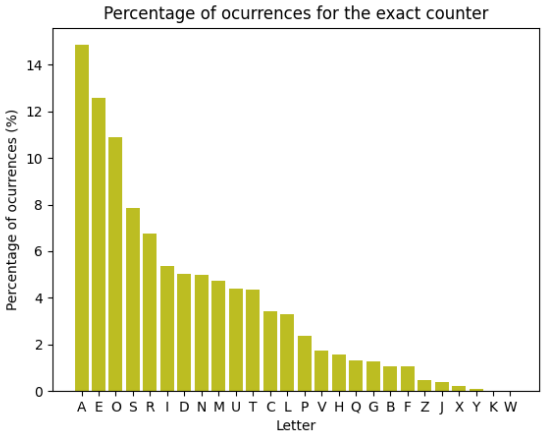

Fig. 1: Percentagem de ocorrências de cada uma das letras para o contador exato para a combinação de todos os textos.

A. Csuros

Analisando um contador probabilístico como o de Csuros é essencial que se repita o seu algoritmo n diversas vezes para se o poder estudar eficazmente. Uma destas medidas é a média, $\overline{x}$ que se obtém com $\frac{1}{n}\sum_{i=1}^{n} x_i$, onde x_i retrata o valor de um contador para cada uma das repetições, sendo este valor médio o resultado que é considerado e apresentado quando não dita mais informação.

Outras medidas estatísticas de interesse relativas ás diversas repetições do processo será o desvio máximo, *maximal_deviation*, que é igual a $max(|x_i - \overline{x}|)$. O valor médio dos desvios, *mean_absolute_deviation*, será igual a $\frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|$. Também relevante se torna a necessidade de verificar o desvio padrão, *standard_deviation*, igual a $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$. Podendo-se verificar também os *outliers* com os valores mínimos

e máximos obtidos.

Também essencial é efetuarem-se as comparações com os valores reais, x, uma vez que se sabem os valores exatos. Deste modo, pode-se medir a exatidão, *exatidao*, que é igual a $\frac{|\overline{x}-x|}{x}$ e o erro relativo, *relative_error*, que é igual a $\frac{|x_i-x|}{x}$, podendo-se estudar o valor máximo, mínimo e médio.

No ficheiro `all.xlsx` encontram-se diversas estatísticas aqui apresentadas e até outras com as páginas para diferentes valores de d, entre 0 e 12, ou seja para M entre 1, contador de Morris, e 4096. Tendo repetido o algoritmo de Csuros 50 vezes. Na página *analise_geral* deste ficheiro é onde se encontram em maior detalhe a combinação de diferentes estatísticas que são apresentadas agora em seguida.

Uma das estatísticas em análise é a exatidão dos resultados médios obtidos, encontrando-se esta na Tabela I. Nesta apresentaram-se os resultados para as 3 letras mais vistas e para as 3 menos contadas pelo contador exato. Como se pode concluir, para as de menos contagens, com um incremento de d o erro passa a ser nulo visto que o contador é exato para estes valores menores, como se esperava pelo princípio do algoritmo.

Constata-se também que para a situação em que este é o contador de Morris a exatidão aparenta ser menor, ou seja, os valores médios obtidos para cada uma das letras é mais próximo que para os d mais elevados. Contudo, verificando os valores para as restantes letras verifica-se que não é possível retirar esta conclusão.

TABELA I: Tabela com os valores da exatidão, entre o valor médio das experiências e o valor exato para o contador de Csuros. Para a combinação de todos os ficheiros, mostrando-se as três letras mais frequentes e as três menos frequentes segundo o contador exato.

Exatidão	A	E	O	...	Y	K	W
d0	4,349149	7,368657	2,083644	...	2,832627	1,073593	11,53719
d1	11,02004	10,19617	12,11079	...	16,27648	7,168831	9,950413
d2	15,40516	23,18247	10,86422	...	16,5339	18,51948	28,94215
d3	17,34946	16,39998	8,810857	...	18,98305	17,4026	14,5124
d4	16,112	10,22213	23,24942	...	16,46504	17,39394	16,49587
d5	11,05867	14,91223	10,72223	...	9,84322	12,67532	11,98347
d6	17,7454	12,66841	24,77438	...	24,91102	5,792208	7,157025
d7	15,51319	12,26391	21,37453	...	17,2839	10,88312	0
d8	14,94149	8,573459	26,29008	...	15,15148	0	0
d9	23,54248	10,2066	22,46196	...	14,17479	0	0
d10	15,7383	13,75356	26,82397	...	13,73623	0	0
d11	18,94109	14,87725	18,05671	...	0	0	0
d12	14,56337	13,56524	19,29568	...	0	0	0

Na Fig. 2 apresenta-se o número de contagens médio para cada um dos valores de d assim como a média das contagens para todos os valores de d. Podendo-se comparar estes aos valores exatos que são apresentados em cor diferente. Daqui verifica-se que no geral a média de todos os valores obtidos são uma subestimação do resultado real medido com o contador exato. Sendo que também ocorrem sobre-estimações para alguns d.

Na Figura 3 apresenta-se o gráfico com o valor do erro médio relativo para cada um dos valores de d. Daqui constata-se que o erro relativo médio tem tendência a decrescer com o aumento do valor numérico de d. Sendo no entanto uma descida não constante. De notar que para as letras com menor frequência as barras desaparecem com o aumento de d devido a este se tornar exato e não se verificar aí qualquer erro.

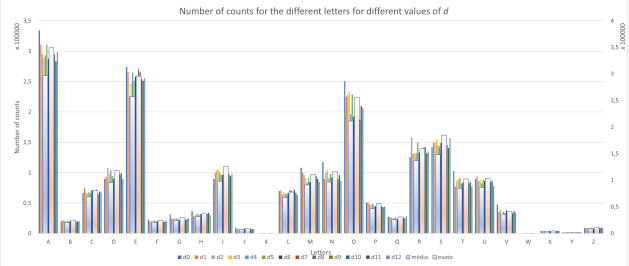

Fig. 2: Valores das contagens médias dos contadores de Csuros para cada uma das letras em cada d.Com o valor médio de todos os d em comparação ao valor exato para o ficheiro de todas as combinações.

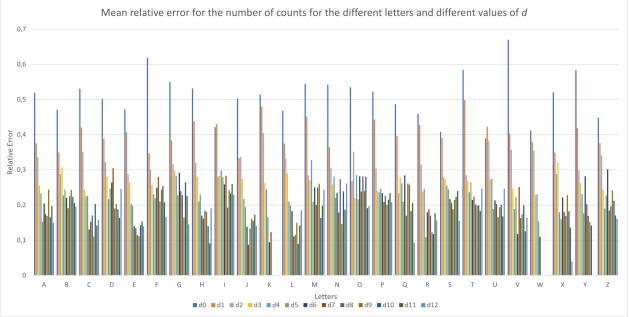

Fig. 3: Valores dos erros relativos entre o valor exato e o valor médio do contador de Csuros para cada letra para diferentes d, para o conjunto de todos os ficheiros.

Atentando ao valor máximo do erro relativo, Fig. 4, verifica-se que estes são bastante consideráveis. Pelo que se torna então essencial repetir o algoritmo diversas vezes pois o erro que pode ser cometido é significativo. Sendo que o maior erro é normalmente obtido com o contador de Morris ou com o contador em que $d = 1$.

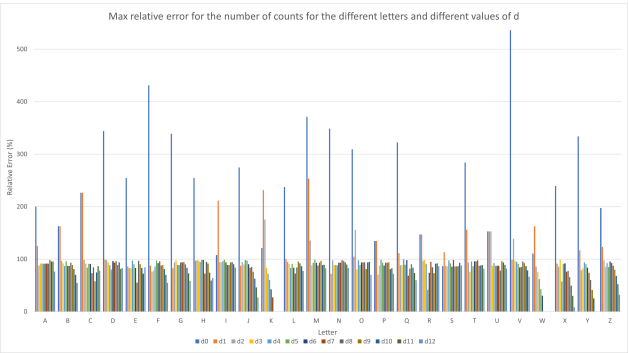

Fig. 4: Valores do erro relativo máximo para cada letra para diferentes d, para o conjunto de todos os ficheiros.

Na Fig. 5 apresenta-se o desvio padrão entre as 50 repetições do contador de Csuros para cada um dos valores de d estudados. Daqui é possível constatar-se

que quanto maior o valor do contador em si maior estes erros são pelo que a análise deve ser feita de forma relativa, Fig. 6. Nesta já é possível confirmar-se que usualmente os menores valores de d levam a que haja uma dispersão maior entre os valores medidos nas diferentes experiências. Também é possível evidenciar-se que a dispersão segue uma tendência decrescente com o aumento de d quando este está mesmo quase a tornar-se um contador exato.

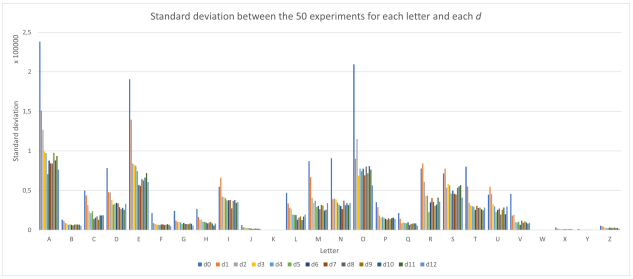

Fig. 5: Desvio padrão das 50 experiências para o contador de Csuros, para o conjunto de todos os ficheiros.

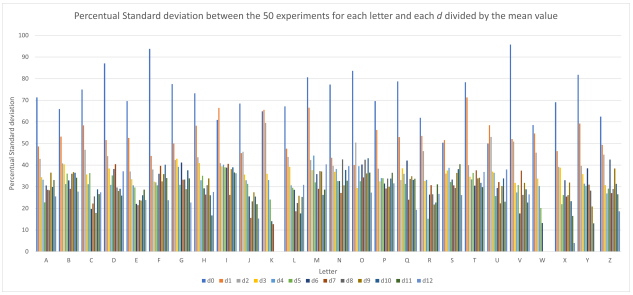

Fig. 6: Desvio padrão percentual das 50 experiências para o contador de Csuros, para o conjunto de todos os ficheiros.

Concretizando a média das 50 experiências dos contadores de Csuros para cada letra é possível ordenar os mesmos para verificar qual a ordem das letras mais frequentes, Tabela II. Aqui apresenta-se por ordem decrescente, de cima para baixo, as letras mais frequentes para cada valor de d.

Na melhor das situações ocorre apenas a troca de um par de letras e no pior caso a troca de 14 letras. Efetuando ainda a média dos valores médios para cada d, coluna *média*, verifica-se que ocorre a melhor situação como em $d = 12$. Contudo, a letra que tem sempre maior número de contagens permanece inalterada, a letra **A**. Para as letras com menores contagens sucede-se a mesma permanência, indo de acordo com o esperado do algoritmo. Havendo contadores com valores exatos relativamente próximos já seria de esperar que a ordem obtida pudesse variar.

B. Space-Saving

Neste caso o objetivo passa por verificar quais as letras com maior contagem. Na Tabela III verifica-se por ordem decrescente, de cima para baixo, a ordem das

TABELA II: Tabela com a ordem decrescente obtida para o valor médio de cada um dos valores de d assim como para o valor médio de todos os d. A amarelo são apresentadas as células que se encontram erradas por comparação à ordem exata, para a combinação de todos os ficheiros.

Exato	d0	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	média
A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
E	E	E	O	E	E	E	E	E	E	E	E	E	E	E
O	O	O	E	O	O	O	O	O	O	O	O	O	O	O
S	S	R	S	S	S	R	S	S	S	S	S	S	S	S
R	R	S	R	R	R	S	R	R	R	R	R	R	R	R
I	N	M	D	I	D	N	I	N	I	N	I	D	I	I
D	M	I	I	N	I	I	N	I	D	I	D	N	D	N
N	T	U	N	D	N	D	D	M	U	D	M	I	N	D
M	D	D	M	M	T	M	U	U	N	U	N	M	M	M
U	I	N	T	T	M	U	M	D	M	M	U	U	U	U
T	U	T	U	U	U	T	T	T	T	T	T	T	T	T
C	L	C	C	L	C	C	C	L	L	C	L	C	C	C
L	C	L	L	C	L	L	L	C	C	L	C	L	L	L
P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
V	V	V	V	V	V	V	V	H	V	V	V	V	V	V
H	H	H	H	H	H	H	H	V	H	H	H	H	H	H
Q	G	Q	Q	Q	Q	Q	Q	Q	G	G	Q	Q	Q	Q
G	Q	G	G	G	G	G	G	G	Q	Q	G	G	G	G
B	F	B	B	F	B	F	B	B	F	B	B	B	F	B
F	B	F	F	B	F	B	F	F	B	F	F	F	B	F
Z	J	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
J	Z	J	J	J	J	J	J	J	J	J	J	J	J	J
X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
K	K	K	K	K	K	K	K	K	K	K	K	K	K	K
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W

letras mais frequentes obtidas com este algoritmo, primeiro para as 3 mais frequentes, seguindo-se das 5 mais frequentes e por fim das 10. Como se pode averiguar, para quando se procuram as 3 mais frequentes obtém-se o mesmo valor no contador pelo que não existe distinção entre estes e na sua ordem, revelando que não se consegue com sucesso fazer este estudo. Dado que neste ficheiro existem $N = 2.352.669$ letras é possível apurar que para as 3 letras mais frequentes $\lfloor \frac{N}{3} \rfloor = 784223$, assim como para as 5 $\lfloor \frac{N}{5} \rfloor = 470533$ e para as 10 mais frequentes $\lfloor \frac{N}{10} \rfloor = 235266$. Confirmando-se para este caso o que tinha sido referido relativamente ao valor do contador mínimo.

Na situação em que se analisam as 5 mais frequentes letras é possível verificar que o valor associado aos contadores continua a ser bastante próximo, não se conseguindo assim fazer uma distinção muito clara entre as letras mais frequentes. Contudo, ao analisarem-se as 10, já se consegue ter uma perceção clara das 4 primeiras letras que já apresentam valores de contadores bastante distintos e na ordem certa ao contador exato. Ainda assim, para as restantes 6 letras estas são bastante próximas e verifica-se que apenas as letras **I** e **N** estão de facto presente nas letras mais frequentes. Regressando ao caso das 5 letras mais frequentes vê-se ainda que, por acaso, a letra **E** acaba por ser também referida como uma das mais frequentes, mas tendo este o valor mínimo do contador ao surgir uma outra letra poderá ser mudado.

As informações relativas a este ponto podem ser verificadas em `all_analise_k_order.xlsx` mas não existe informação adicional à que já foi apresentada.

TABELA III: Tabela com a ordem decrescente das letras mais frequentes e do valor do contador para o algoritmo de *Space-Saving* para quando se procura pelas mais frequentes 3,5 e 10 letras com comparação à ordem exata. A amarelo são apresentadas as letras com uma ordem errada e a verde valores de contadores que sejam iguais, para a combinação de todos os ficheiros.

k3	k3 - count	k5	k5 - count	k10	k10 - count	Exato	Valor Exato
A	784223	A	470535	A	349388	A	349387
L	784223	L	470534	E	295554	E	295554
H	784223	H	470534	O	256136	O	256134
		E	470533	S	207376	S	184520
		V	470533	L	207370	R	159116
				I	207370	I	126282
				H	207369	D	118073
				V	207369	N	116843
				C	207369	M	111262
				N	207368	U	103678

V. Análise da compilação das obras de Luís Vaz de Camões

A inclusão deste ponto não pretende apresentar os mesmos detalhes estatísticos. Apenas apresentar alguma informação adicional para verificar quais os efeitos que a inclusão dos textos de Eça de Queirós apresentam. Desta forma, o ficheiro apresenta também um número de caracteres inferior.

Primeiramente podem-se analisar quais as letras mais frequentes nos textos de Camões, Fig. 7. Aqui verifica-se que ao contrário da combinação de todos os textos, Fig. 1, a letra mais frequente é **E** em vez da letra **A**. Constatando-se que as contagens são mais próximas entre diversos pares como ocorre para **E** e **A**, **N** e **I**, **D** e **M** e **T** e **U**.

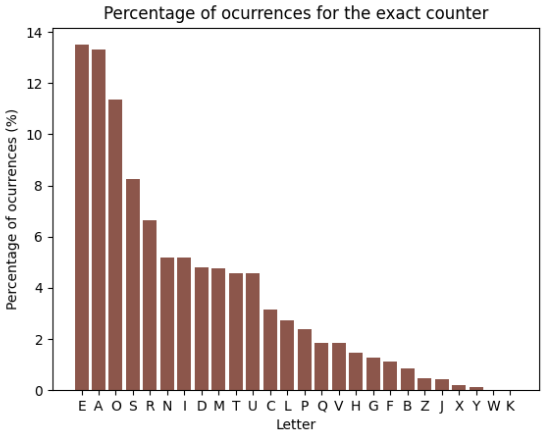

Fig. 7: Percentagem de ocorrências de cada uma das letras para o contador exato para a combinação dos textos de Camões.

A. Csuros

Para este ficheiro utilizou-se também o algoritmo do contador de Csuros repetindo-o 50 vezes como no caso anterior. Na Fig. 8 apresenta-se o valor médio das experiências para cada valor de d. Encontrando-se também no mesmo gráfico os valores médios de todos os d e do valor exato. Verificando-se o mesmo que na combinação de todos os ficheiros que o valor médio acaba por ser uma subestimação do valor real aparentando ocorrerem no entanto menos sobre-estimações.

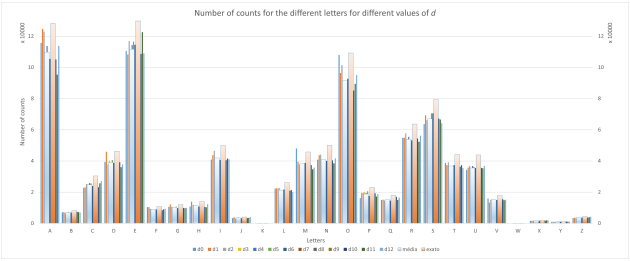

Fig. 8: Valores das contagens médias dos contadores de Csuros para cada uma das letras em cada d.Com o valor médio de todos os d em comparação ao valor exato para o ficheiro das combinações dos textos de Camões.

Analisando o erro relativo, Fig. 9, é possível concluir o mesmo que no ponto anterior que um d menor leva a um erro relativo superior. Dado que o número de caracteres é inferior mais rapidamente existem letras a ficar com um erro relativo nulo, como as letras **K** e **W** com o aumento de d. Em termos numéricos poderá eventualmente considerar-se que os erros relativos aqui medidos foram menores face à combinação de todos os textos.

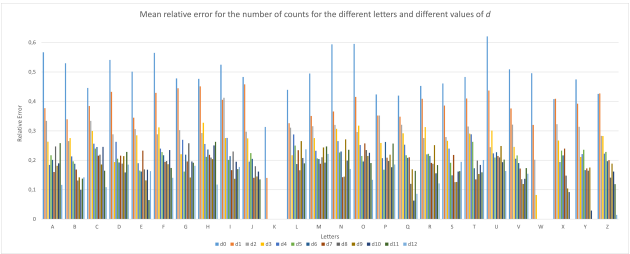

Fig. 9: Valores dos erros relativos entre o valor exato e o valor médio do contador de Csuros para cada letra para diferentes d, para o conjunto das combinações dos textos de Camões.

Na Fig. 10 apresentam-se os desvios padrões percentuais. As conclusões a retirar daqui são as mesmas do ponto anterior da combinação de todos os ficheiros. Que para os valores de d menores existe uma maior variação. Em termos de amplitude a diferença não é muito notória sendo apenas mais pequena para alguns casos.

Na Tabela IV apresenta-se também pela ordem decrescente as letras mais frequentes medidas para cada d. Aqui verifica-se que existe uma situação em que, por acaso, o valor médio dos contadores para $d = 8$ leva a que a ordem seja exatamente igual à exata. Efetuando a média para os diferentes valores de d chega-se a que ocorra a troca de um par de valores. Dado que es-

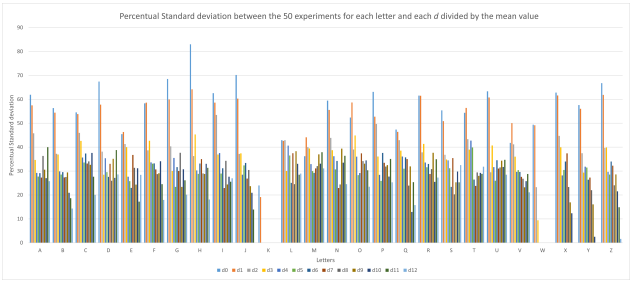

Fig. 10: Desvio padrão percentual das 50 experiências para o contador de Csuros, para o conjunto das combinações dos textos de Camões.

tes valores também apresentavam ser muito próximos poderia já ser de esperar que tal fosse acontecer.

TABELA IV: Tabela com a ordem decrescente obtida para o valor médio de cada um dos valores de d assim como para o valor médio de todos os d. A amarelo são apresentadas as células que se encontram erradas por comparação à ordem exata, para a combinação dos textos de Camões.

Exato	d0	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	média
E	A	A	A	A	E	E	E	A	E	E	E	E	A	E
A	E	E	E	E	A	A	A	E	A	A	A	A	E	A
O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
N	M	D	I	I	I	I	I	N	N	I	I	I	N	I
I	I	N	N	D	N	D	N	I	I	T	N	N	I	N
D	N	I	T	M	D	N	M	T	D	N	D	T	D	D
M	D	M	M	N	M	M	D	M	M	D	M	D	U	M
T	T	T	D	T	U	U	T	D	T	M	T	U	M	T
U	U	U	U	U	T	T	U	U	U	U	U	M	T	U
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Q	V	Q	V	V	V	Q	V	V	Q	V	Q	Q	Q	Q
V	Q	H	Q	Q	Q	V	Q	Q	V	Q	V	V	V	V
H	H	V	H	H	H	H	H	H	H	H	H	H	H	H
G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Z	J	J	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
J	Z	Z	J	J	J	J	J	J	J	J	J	J	J	J
X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
K	K	K	K	K	K	K	K	K	K	K	K	K	K	K

Mais dados estatísticos relativos a esta situação em análise encontram-se em `camoes_analise.xlsx`. Na primeira página, *analise_geral*, encontra-se uma compilação de algumas estatísticas com as restantes páginas a serem referentes a cada um dos d em estudos com as mesmas estatísticas apresentadas em `all_analise.xlsx`.

B. Space-Saving

Procedeu-se então ao uso do algoritmo de *Space-Saving*. Verificando que existem $N = 961.615$ caracteres é possível efetuarem-se as mesmas contas como anteriormente, assim para as 3 letras mais frequentes $\lfloor \frac{N}{3} \rfloor = 320538$, para as 5 $\lfloor \frac{N}{5} \rfloor = 192323$ e para as 10 mais frequentes $\lfloor \frac{N}{10} \rfloor = 96161$. Analisando a Tabela V pode-se confirmar o que se apresentou na introdução e o que se confirmou no ponto anterior, que de facto o valor mínimo é menor ou igual a esta fração. Aqui também se confirma que a ordem a que se chegou para as 3 e 5 letras mais frequentes não tem grande sucesso, com os valores dos contadores muito próximos uns dos outros. Contudo, para as 10 mais frequentes letras verifica-se que as primeiras 4 letras são apresentadas na ordem corretas. Além do mais, apesar da ordem não ser sempre a correta, verifica-se que ao menos algumas das letras mais frequentes são apresentadas como o **T**, o **I**, o **M**, o **R** e o **N**.

TABELA V: Tabela com a ordem decrescente das letras mais frequentes e do valor do contador para o algoritmo de *Space-Saving* para quando se procura pelas mais frequentes 3,5 e 10 letras com comparação à ordem exata. A amarelo são apresentadas as letras com uma ordem errada e a verde valores de contadores que sejam iguais, para a combinação de todos os ficheiros.

k3	k3 - count	k5	k5 - count	k10	k10 - count	Exato	Valor Exato
A	320540	A	192327	E	129771	E	129771
S	320538	T	192323	A	128117	A	128116
T	320537	O	192322	O	109187	O	109185
		S	192322	S	84941	S	79239
		N	192321	T	84935	R	63692
				C	84934	N	49974
				R	84934	I	49840
				I	84933	D	46310
				N	84932	M	45736
				M	84931	T	44072

A análise desta porção encontra-se em `camoes_analise_k_order.xlsx`. Contudo, toda a informação relevante encontra-se já aqui apresentada.

VI. Conclusão

Com a realização deste trabalho foi possível estudar a forma como é possível utilizar diferentes tipos de contadores consoante o que se pretende efetuar.

Para a análise do contador de Csuros verificou-se que o incremento de d leva a melhores resultados à custa de necessitar de mais memória. Inclusive a este estudo teria sido de interesse verificar como este algoritmo se iria comportar para valores ainda maiores de d assim como para variações no número de repetições. Uma desvantagem deste método acaba também por ser que a necessidade de repetir o processo várias vezes que leva a que seja necessário um maior tempo de computação. Contudo, para casos em que a memória se torna um fator limitante este acaba por ser uma melhor opção, com melhor desempenho que o contador de Morris para valores baixos de contagens, por exemplo. No que diz respeito à ordem das letras mais frequentes verifica-se que este consegue em média responder com sucesso.

O algoritmo de *Space-Saving* teve um sucesso limitado, constando-se um melhor desempenho para quando se procuravam mais letras. No entanto, este consegue colocar nas mais frequentes de facto as letras mais frequentes mesmo não estando na ordem correta. Contudo, por vezes, isto acaba por depender da ordem pela qual o texto aparece.

Relativamente aos dois ficheiros em análise, é possível verificar-se que a escrita de Eça de Queirós é bastante diferente da de Luís Vaz de Camões, pelo menos na

frequência das letras empregues. Verificando-se no entanto que as principais letras continuam a ser as mesmas apenas por ordem diferente.

Desta forma verifica-se que os contadores não exatos aplicados apresentam a sua relevância conseguindo obter respostas que podem ser suficientemente boas para a aplicação em causa.

Bibliografia

[1] Gundersen, G. Approximate Counting with Morris's Algorithm. *Approximate Counting With Morris's Algorithm.* (2019,11), https://gregorygundersen.com/blog/2019/11/11/morris-algorithm/

[2] J. Cichon and W. Macyna, "Approximate Counters for Flash Memory,"2011 IEEE 17th International Conference on Embedded and Real-Time Computing Systems and Applications, 2011, pp. 185-189, doi: 10.1109/RTCSA.2011.81.

[3] M. Rahman and M. A. Hasan, "Approximate triangle counting algorithms on multi-cores,"2013 IEEE International Conference on Big Data, 2013, pp. 127-133, doi: 10.1109/BigData.2013.6691744.

[4] Camões Os Lusíadas by Luís de Camões. *Project Gutenberg.* (2002,7), https://www.gutenberg.org/ebooks/3333

[5] Bocage & Camões A Morte de D. Ignez de Castro by Manuel Maria Barbosa du Bocage and Luís de Camões. *Project Gutenberg.* (2007,10), https://www.gutenberg.org/ebooks/23110

[6] Camões Obras completas de luis de camões, Tomo II by Luís de Camões. *Project Gutenberg.* (2010,3), https://www.gutenberg.org/ebooks/31509

[7] Camões Obras completas de luis de camões, tomo III by Luís de Camões. *Project Gutenberg.* (2011,8), https://www.gutenberg.org/ebooks/37192

[8] Queirós, E. A Relíquia by Eça de Queirós. *Project Gutenberg.* (2006,1), `https://www.gutenberg.org/ebooks/17515`

[9] Queirós, E. Os Maias: Episodios da vida romantica by Eça de Queirós. *Project Gutenberg.* (2012,8), `https://www.gutenberg.org/ebooks/40409`

[10] Csuros, M. Approximate counting with a floating-point counter. (arXiv,2009), `https://arxiv.org/abs/0904.3062`

[11] Metwally, A., Agrawal, D. & Abbadi, A. Efficient Computation of Frequent and Top-k Elements in Data Streams. *The Honkong University Of Science And Technology.* (2008), `https://cse.hkust.edu.hk/~raywong/comp5331/References/EfficientComputationOfFrequentAndTop-kElementsInDataStreams.pdf`

[12] Cormode, G. & Hadjieleftheriou, M. Finding the frequent items in streams of data - DIMACS. *Dimacs.* (2008), `http://dimacs.rutgers.edu/~graham/pubs/papers/freqcacm.pdf`