Algoritmos Probabilísticos: Contagem dos Itens Mais Frequentes – Ficheiros de Texto de Obras Literárias, com Count-Min Sketch

Ana Sofia Medeiros de Castro Moniz Fernandes | 88739

***Resumo* - O presente relatório tem como propósito a apresentação de uma solução para determinar os itens mais frequentes de um conjunto de dados, usando "The Count-Min Sketch", no âmbito do terceiro trabalho prático da Unidade Curricular Algoritmos Avançados, do Mestrado de Engenharia Informática. Ao longo do documento serão apresentados e explicados os detalhes da solução (onde foram usadas cinco funções de "hash"), bem como alguns testes realizados e as suas interpretações.**

***Abstract* - The present report aims to present a solution to count the most frequent items of a data set, using "The Count-Min Sketch", in the context of the third 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 (where five hash functions were used), as well as some performed tests and its interpretations.**

I. INTRODUÇÃO

No terceiro 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 "Hipótese A-2 – Ficheiros de texto de obras literárias", usando "The Count-Min Sketch", com o intuito de determinar quais as letras mais frequentes presentes em cada ficheiro. Assim sendo, foram usadas cinco diferentes funções de hash, por forma a visualizar vários resultados e realizar comparações. Além disso, foi realizada uma análise da eficiência de cada função de hash.

As obras usadas para realizar este trabalho foram [2]:

- *Hamlet*, de William Shakespeare ;
- *Camping in the Winter Woods - Adventures of Two Boys in the Maine Woods*, de Elmer Russell Gregor;
- *Home-made Electrical Apparatus*, de Alfred Powell Morgan;
- *Never Fire First - A Canadian Northwest Mounted Story*, de James French Dorrance;
- *Peter Pan*, de James M. Barrie
- *Stories of the Railroad*, de John A. Hill

II. "THE COUNT MIN SKETCH"

Como mencionado anteriormente, para realização deste trabalho, foi usado "The Count Min Sketch", uma estrutura de dados probabilística que serve como tabela de frequências de eventos [3] sendo que, neste caso, um evento corresponde à ocorrência de um caracter.

A. Funções de hashing escolhidas

O "The Count Min Sketch" passa, então, por uma estrutura de dados de duas dimensões, que contém *w* colunas e *d* linhas. Associada a cada linha, está uma função de hash - quando um novo tipo de evento é detetado, para cada linha da tabela, aplica-se a função de hash para obter o *index*. Seguidamente, incrementa-se o valor que está presente nessa linha, e nesse *index* (coluna), em uma unidade. [3]

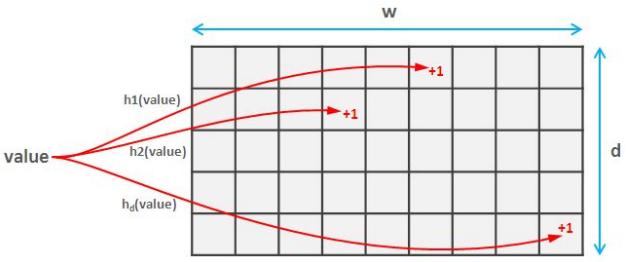

Fig. 1 - Funcionamento do "Count-Min Sketch" [4]

As funções de hash escolhidas para este trabalho, retiradas da biblioteca *hashlib* (em Python 3), foram:

- 'md5'
- 'sha256'
- 'sha1'
- 'blake2s'
- 'dsaEncryption'

III. ANÁLISE DOS RESULTADOS

Para um melhor estudo dos resultados, irão analisar-se os resultados obtidos para o ficheiro de texto "eng_hamlet.txt", que corresponde à obra "Hamlet", na sua língua original (língua inglesa). É de notar que o código usado para o "The Count Min Sketch" foi adaptado do código que se encontra na página da disciplina [5].

A.As 20 letras mais contadas, com 50 colunas e 5 linhas

Comece-se, então, por analisar os resultados para as 20 letras mais contadas, com uma **tabela de 50 colunas e 5 linhas**:

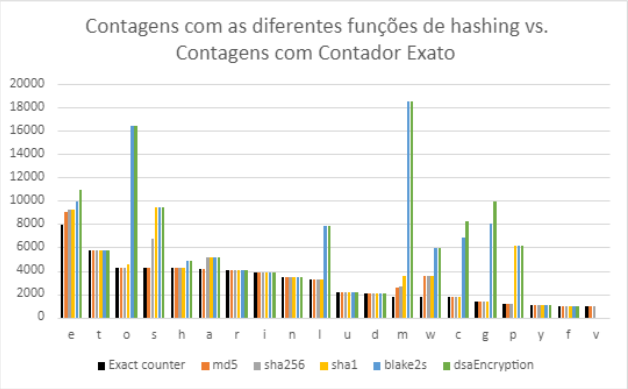

Fig. 2 - Gráfico para comparação das contagens com o contador exato e resultados das contagens para cada função de hash, para o top 20 de caracteres mais contados pelo contador exato

Char	Couting
e	7937
t	5794
o	4282
s	4277
h	4258
a	4186
r	4026
i	3848
n	3501
l	3217
u	2203
d	2058
m	1787
w	1780
c	1717
g	1353
p	1172
y	1059
f	942
v	911

Fig. 3 - 20 caracteres mais contados pelo contador exato

Hash function	Count min sketch top 20	Count min sketch counting
md5	e	8996
md5	t	5794
md5	o	4282
md5	s	4277
md5	h	4258
md5	a	4186
md5	r	4026
md5	i	3848
md5	w	3567
md5	n	3501
md5	l	3217
md5	m	2566
md5	u	2203
md5	d	2058
md5	c	1717
md5	g	1353
md5	p	1172
md5	y	1059
md5	f	942
md5	v	911

Fig. 4 - 20 caracteres mais contados pelo “Count Min Sketch”, usando a função de hash “md5”

sha256	e	9200
sha256	s	6740
sha256	t	5794
sha256	a	5128
sha256	o	4282
sha256	h	4258
sha256	r	4026
sha256	i	3848
sha256	w	3567
sha256	n	3501
sha256	l	3217
sha256	m	2698
sha256	u	2203
sha256	d	2058
sha256	c	1717
sha256	g	1353
sha256	p	1172
sha256	y	1059
sha256	f	942
sha256	v	911

Fig. 5 - 20 caracteres mais contados pelo “Count Min Sketch”, usando a função de hash “sha256”

sha1	s	9428
sha1	e	9200
sha1	k	7853
sha1	p	6108
sha1	t	5794
sha1	a	5128
sha1	o	4567
sha1	h	4258
sha1	r	4026
sha1	i	3848
sha1	m	3567
sha1	w	3567
sha1	n	3501
sha1	l	3217
sha1	u	2203
sha1	d	2058
sha1	c	1717
sha1	g	1353
sha1	y	1059
sha1	f	942

Fig. 6 - 20 caracteres mais contados pelo “Count Min Sketch”, usando a função de hash “sha1”

blake2s	m	18505
blake2s	o	16405
blake2s	e	9900
blake2s	s	9428
blake2s	g	7994
blake2s	l	7891
blake2s	k	7853
blake2s	c	6817
blake2s	p	6108
blake2s	w	5966
blake2s	t	5794
blake2s	a	5128
blake2s	h	4828
blake2s	r	4026
blake2s	i	3848
blake2s	n	3501
blake2s	u	2203
blake2s	d	2058
blake2s	y	1059
blake2s	f	942

Fig. 7 - 20 caracteres mais contados pelo "Count Min Sketch", usando a função de hash "blake2s"

dsaEncryption	m	18505
dsaEncryption	o	16405
dsaEncryption	e	10950
dsaEncryption	g	9914
dsaEncryption	s	9428
dsaEncryption	c	8242
dsaEncryption	l	7891
dsaEncryption	k	7853
dsaEncryption	p	6108
dsaEncryption	w	5966
dsaEncryption	t	5794
dsaEncryption	a	5128
dsaEncryption	h	4828
dsaEncryption	r	4026
dsaEncryption	i	3848
dsaEncryption	n	3501
dsaEncryption	u	2203
dsaEncryption	d	2058
dsaEncryption	y	1059
dsaEncryption	f	942

Fig. 8 - 20 caracteres mais contados pelo "Count Min Sketch", usando a função de hash "dsaEncryption"

É de notar que, neste relatório, além de se incluir o gráfico, obtou-se por incluir também as tabelas dos valores respetivos, por forma a poder ter-se uma melhor noção dos valores contabilizados.

Através da análise do gráfico (figura 2), onde constam os valores observados nas tabelas seguintes (figuras 3-8), é possível observar-se que:

• As contagens obtidas com as funções de hash "blake2s" e "dsaEncryption" são sempre muito maiores do que as contagens reais;

• No top 20 de caracteres mais contados, das funções de hash "sha1", "blake2s" e "dsaEncryption", não consta a letra "v" (no gráfico da figura 2, o valor da contagem desta letra, das respetivas funções de hash, encontra-se a 0);

• As contagens obtidas com as funções de hash "md5" e "sha256" são, no geral, bastante aproximadas da contagem exata.

B.Comparação do total de caracteres contado com cada função de hash

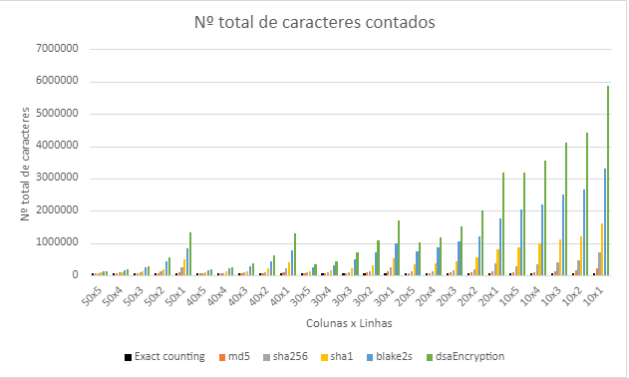

Fig. 9 - Gráfico para comparação das contagens realizadas pelo "Count Min Sketch", em cada função de hash, com as contagens exatas

Counting for file eng_hamlet.txt

Exact counting: 62201

Colums	Rows	md5	sha256	sha1	blake2s	dsaEncryption
50	5	65826	69567	85377	132237	136632
50	4	65826	81315	98127	165894	170664
50	3	65826	93876	124239	243803	263434
50	2	65826	117661	176655	436395	554497
50	1	105492	256536	486864	828464	1340864
40	5	63355	74001	98133	168511	190705
40	4	63446	77298	110978	213380	235574
40	3	65984	87420	136287	287997	366036
40	2	72376	104495	211934	444108	605137
40	1	88193	228920	411572	772202	1313147
20	5	66061	110581	328063	740790	1010647
20	4	73214	119945	382822	874629	1163704
20	3	81732	159072	443303	1056312	1505470
20	2	87861	173886	549716	1196602	2001574
20	1	130373	383870	813182	1766612	3196757
30	5	63244	79949	135503	252561	345781
30	4	64688	94846	140645	297822	427134
30	3	70145	106374	202414	510022	725923
30	2	83552	131749	298240	706151	1093521
30	1	124430	255680	526916	993916	1694416
10	5	97022	290841	863547	2026705	3172908
10	4	104235	325328	988883	2196575	3562580
10	3	122629	402550	1101612	2487887	4099918
10	2	140203	466334	1216522	2660041	4412006
10	1	225394	722761	1604275	3309065	5866250

Fig. 10 - Contagens realizadas pelo "Count Min Sketch", em cada função de hash

Através da análise do gráfico (figura 9), onde constam os valores observados na tabela (figuras 10), nota-se que:

• Quanto menor o número de linhas e de colunas, mais os resultados obtidos pelo "Cout Min Sketch" vão diferir da contagem exata;

• As contagens obtidas com as funções de hash "sha1", "blake2s" e "dsaEncryption" são as que mais diferem da contagem exata.

C.Comparação do tempo de execução de cada função de hash

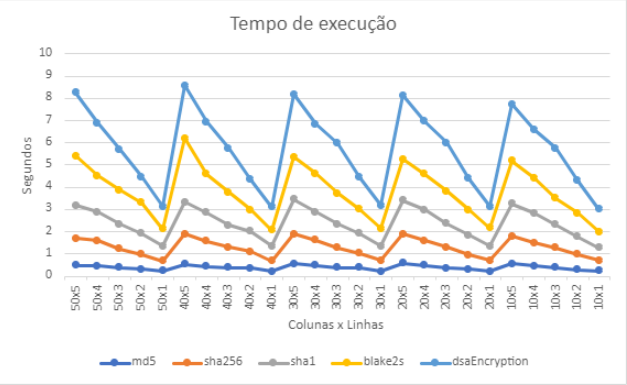

Fig. 10 - Comparação dos tempos de execução

Relativamente aos tempos de execução, pode observar-se (Figura 10) que a função de hashing mais rápida é a "md5", e a mais lenta é a "dsaEncryption". Além disso, pode concluir-se que, quanto menor for o número de linhas, menor será o tempo de execução (independentemente do número de colunas).

IV. CONCLUSÃO

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

- As melhores performances verificam-se quando ocorre um baixo número de linhas. No entanto, isto não é benéfico para a contagem em si, uma vez que quanto menor for o número de linhas e de colunas, mais as contagens diferem da contagem exata ;
- A função de hash com melhor performance é a "md5", sendo também ela a que apresenta valores de contagem mais próximos dos valores exatos;
- Quanto maior for o número de linhas e de colunas, mais os valores das contagens obtidos por cada função de hash se aproximam do valor exato.

REFERÊNCIAS

[1] J. Madeira, "Data Streams : Sketches", Aveiro, 2021
[2] "Project Guttenberg", [Online]. Available: https://tinyurl.com/qfmmkma [Acedido em 5 de fevereiro de 2021]
[3] "Count–min sketch", [Online]. Available: https://tinyurl.com/qfbjol5 [Acedido em 5 de fevereiro de 2021]
[4] "Explaining the count sketch algorithm", [Online]. Available: https://tinyurl.com/y8rezer3 [Acedido em 5 de fevereiro de 2021]
[5] "CountMinSketch", [Online]. Available: https://tinyurl.com/y36v9ptt [Acedido em 5 de fevereiro de 2021]