TITULO DO TRABALHO

Hugo Veríssimo - 124348 - hugoverissimo@ua.pt

***Abstract* – abstrato em ingles**

***Resumo* – abstrato em pt resumo**

I. Introdução

A análise de texto é uma área de estudo fundamental, com diversas aplicações tais como análise de sentimentos ou de opiniões, personalização da experiência do utilizador, recomendação de conteúdo, entre outras [1]. Uma das tarefas centrais nesta área é a identificação da frequência de palavras em grandes volumes de texto, tal como livros, bases de dados ou redes sociais, de modo extrair informações relevantes sobre o conteúdo e estrutura dos textos em análise.

Contudo, a identificação precisa da frequência de palavras em textos de larga escala apresenta desafios significativos, especialmente em termos de memória. Métodos de contagem precisa, que mantêm o registo exato da contagem de cada palavra, revelam-se ineficientes devido ao elevado consumo de memória. Há assim a necessidade do estudo de métodos mais eficientes e escaláveis, principalmente em situações em que os dados estão em constante fluxo, como em *streams* de dados. Neste contexto, algoritmos de contagem aproximada e identificação de itens frequentes têm vindo a ganhar destaque, uma vez que permitem a identificação de palavras mais frequentes de forma eficiente e com uma margem de erro controlada [2].

Este relatório visa explorar três abordagens para este problema: contadores exatos, contadores aproximados e identificação de itens frequentes em *streams* de dados. Para cada uma destas abordagens, será apresentado um algoritmo e .....

II. Metodologia da Análise

Para realizar a análise de frequência de palavras, foram selecionados três livros, a partir livraria online *Project Gutenberg* [3], nomeadamente: *Pinocchio: The Tale of a Puppet* (em inglês, EN), *Le avventure di Pinocchio: Storia di un burattino* (em italiano, IT) e *Pinocchion seikkailut: Kertomus marioneteista* (em finlandês, FI). Estes livros foram selecionados por serem traduções do mesmo livro original, conhecido em português como *As Aventuras de Pinóquio*, de Carlo Collodi. A escolha destes livros permite a comparação da frequência de palavras em diferentes idiomas, bem como a análise de semelhanças e diferenças entre as traduções.

Numa primeira fase, os ficheiros de texto descarregados a partir do *Project Gutenberg* foram processados removendo informações irrelevantes, como metadados e licenças, palavras insignificantes e sinais de pontuação. Para além disso todas as palavras foram convertidas para minúsculas e lematizadas. Estas transformações são fundamentais, de modo a simplificar o texto e concentrar a análise nas palavras mais relevantes, garantindo uma avaliação mais precisa e eficiente da frequência de termos. É importante referir que estas transformações foram realizadas com recurso à biblioteca *spaCy*, através do *Python*.

implementar algortimos, analise dos dados, correr ns quantas vezes, ...

numero total de palavras que cada book tem ? n = ?

III. Contadores Exatos

Quanto aos contadores exatos, tal como o nome indica, este tipo de técnica é exate, resultando numa contagem precisa da frequência de palavras, no contexto em causa. O algoritmo apresentado de seguida, designado por *Contador Exato*, é um exemplo de um contador exato, que percorre o texto processado e regista a frequência de cada palavra num dicionário. Este algoritmo é eficiente em termos de precisão, uma vez que mantém um registo exato da contagem de cada palavra, no entanto, revela-se ineficiente em termos de memória, especialmente em situações em que o volume de texto é elevado.

Algoritmo 1 Contador Exato

Entrada: texto processado (`T`)
Saída: dicionário onde as palavras são as chaves e os valores são as suas frequências (`D`)

```
1: D ← empty dictionary
2: words ← list of words from T
3: for each word in words do
4:     if word ∉ D then
5:         D[word] ← 0
6:     end if
7:     D[word] ← D[word] + 1
8: end for
9: return D
```

Atendendo à complexidade espacial, no pior caso, onde todas as palavras que constituem o texto `T` são distintas, a mesma é dada por $O(|\texttt{words}|)$, onde $|\texttt{words}|$ representa o número de palavras no texto. Isto acontece pelo facto do dicionário `D` conter uma entrada para cada palavra distinta no texto. NAO SEI SE ESTA CERTO, NS SE TENHO DE CONTABILIZAR O TAMANHO DE WORDS AO EM INVES DE SER SO O D ou REFERIR QUE O QUE IMPORTA É O TAMANHO DO DICIONARIO? TALVEZ FALAR DOS

DOIS e dps no fim dizer q o words deve ser ignorado pq é o texto e o que importa é o dicionario? idk
comparison of the memory (complexity ?) of the algorithms
.....
através da aplicacao do algoritmo de contagem exata, foi possível identificar as 10 palavras mais frequentes em cada um dos livros analisados. A Tabela I apresenta as palavras mais frequentes em cada idioma, juntamente com o número de ocorrências de cada palavra.

TABELA I: CAPTION top10 palavras mais frequentes em cada idioma

EN		IT		FI	
Palavra	#	Palavra	#	Palavra	#
pinocchio	457	pinocchio	460	pinocchio	443
say	282	il	386	sanoa	258
little	238	dire	282	saada	143
puppet	209	si	251	alkaa	134
come	141	burattino	225	tehdä	134
boy	140	volere	167	marionetti	131
like	133	vedere	152	poika	81
good	131	andare	134	huutaa	81
poor	127	povero	134	nähdä	80
go	116	ragazzo	126	kysyä	77

como seria de esperar, a palavra "pinocchio"é a mais frequente em todos os idiomas, uma vez que se trata do nome do protagonista do livro. Para além disso, é possível observar algumas semelhanças entre os idiomas, nomeadamente a presença de palavras que têm o mesmo siginificado em diferentes idiomas, como "puppet"e "burattino"(marioneta em italiano), ...
para além disso, esta analise tambem permite a analise da quantidade de palavras distintas Fig. 2 em cada idioma, bem como a distribuição da frequência de palavras Fig. 1, o que pode ser útil para a comparação de diferentes traduções de um mesmo livro, por exemplo.

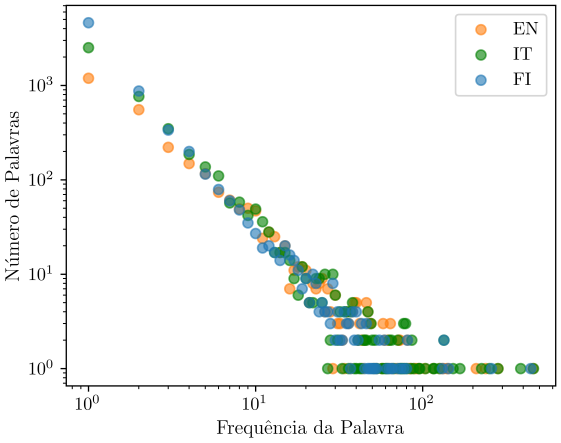

Fig. 1: distribuicao da frequência de palavras em cada idioma

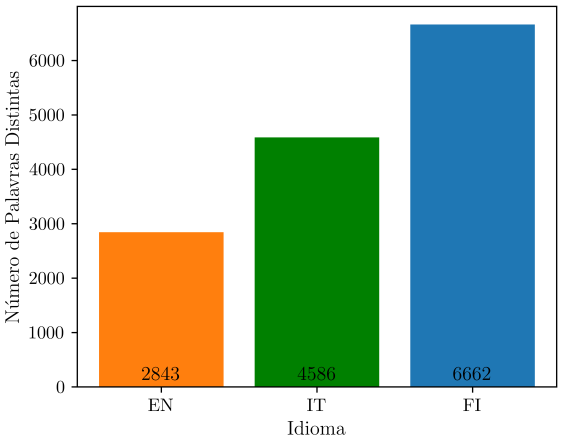

Fig. 2: numero de palavras distintas em cada idioma

a diferenca pode ter a ver com o desempenho do spacy a lematizar as palavras em diferentes idiomas, ou com a propria complexidade da lingua

IV. CONTADORES APROXIMADOS

Os contadores aproximados, também conhecidos por contadores probabilísticos, inventados por Robert Morris, são algoritmos especializados em realizar estimativas eficientes de contagens, utilizando quantidades reduzidas de memória em comparação com métodos tradicionais, através de técnicas baseadas em probabilidade. Estes contadores são particularmente úteis em situações em que a precisão exata da contagem não é crítica, permitindo uma redução significativa no uso de memória, sem comprometer a qualidade da análise [4].

Um exemplo de um contador aproximado é o algoritmo de contagem aproximada apresentado de seguida, designado por *Contador Aproximado*. Neste exemplo, o algoritmo utiliza uma probabilidade de contagem fixa de 1/16. Para cada evento, neste caso, cada palavra do texto, é gerado um número aleatório entre 0 e 1, e a palavra é contada se o número gerado for menor que a probabilidade de contagem. Isto permite que a contagem seja realizada de forma aproximada, com uma margem de erro controlada, e um menor uso de memória.

Algoritmo 2 Contador Aproximado

Entrada: texto processado (`T`)
Saída: dicionário onde as palavras são as chaves e os valores são as suas frequências estimadas (`D`)

```
D ← empty dictionary
words ← list of words from T
for each word in words do
    r ← Uniform(0, 1)
    if r < 1/16 then
        if word ∉ D then
            D[word] ← 0
        end if
        D[word] ← D[word] + 1
    end if
end for
for each word in D do  ▷ Estimate the total count
    D[word] ← D[word] ×16
end for
return D
```

é importante referir q existem outros tipos, diferentes fixed probability, etc.

as contagem sao smp multiplas de 16, pq a probabilidade é 1/16, e dps é multiplicado por 16 para obter a contagem aproximada

complexidade espacial tal tal bla bla

nas tabelas seguintes, por ser um processo aleatorio, apos tere sudo corrido 20 vezes, foi calculada a média, valor max e min de cada palavra, escolhidos as 10 mais frequentes tendo em conta a media, e comparado com a contagem exata. nas tabelas seguintes pode-se observar esses resultados, para cada um dos idiomas analisados. também se repara que as palavras a preto mantiiveram a sua posucao em relacao ao top10 exato, as a laranja estao descoladas mas fazem parte, e as a vermelho não fazem parte do top10 real

TABELA II: top10 dos aproximados sendo o top tendo em conta a media, e media arredondada ao inteiro, e idioma ING

Palavra	$\#_{\text{min}}$	$\#_{\text{média}}$	$\#_{\text{max}}$	$\#_{\text{real}}$
pinocchio	336	439	544	457
say	96	270	400	282
little	144	227	320	238
puppet	128	223	336	209
come	64	148	240	141
boy	80	140	176	140
go	48	132	240	116
poor	48	129	240	127
like	32	128	192	133
good	64	126	224	131

TABELA III: top10 dos aproximados sendo o top tendo em conta a media, e media arredondada ao inteiro, e idioma italiano

Palavra	$\#_{\text{min}}$	$\#_{\text{média}}$	$\#_{\text{max}}$	$\#_{\text{real}}$
pinocchio	240	439	544	460
il	224	379	512	386
dire	160	263	448	282
si	128	252	320	251
burattino	96	223	320	225
volere	96	169	256	167
povero	32	139	240	134
bello	80	133	240	116
vedere	32	131	192	152
andare	80	127	192	134

TABELA IV: top10 dos aproximados sendo o top tendo em conta a media, e media arredondada ao inteiro, e idioma finlandes

Palavra	$\#_{\text{min}}$	$\#_{\text{média}}$	$\#_{\text{max}}$	$\#_{\text{real}}$
pinocchio	320	445	672	443
sanoa	160	266	352	258
saada	48	160	256	143
tehdä	48	136	224	134
marionetti	80	132	224	131
alkaa	64	104	160	134
geppetto	32	87	160	71
huutaa	32	86	128	81
pää	32	80	128	65
olla	16	76	144	61

nota se mais erros quanto menor é a contagem, pq ha maior densidade de palavras, como se viu na contagem exata figura tal

V. Contadores Space-Saving

Pelo facto de muitos processos de geração de dados poderem ser modelados como fluxos de dados, que produzem enormes quantidades de informações simples isoladamente, mas que, em conjunto, formam um todo complexo, torna-se interessante a utilização de métodos que respondam rapidamente a cada nova informação e utilizem recursos muito pequenos em comparação com o volume total de dados. Neste contexto, o algoritmo de contagem *Space-Saving* é uma solução eficiente para a identificação de itens frequentes em *streams* de dados, permitindo acompanhar contagens frequentes de forma eficiente, mesmo sob restrições de memória (número máximo de diferentes palavras a manter, k) [5].

Este algoritmo é exposto no pseudocódigo seguinte:

Algoritmo 3 Contador *Space-Saving* [5]

Entrada:
- texto processado (`T`)
- número máximo de itens a manter (`k`)

Saída: dicionário com a estimativa das `k` palavras mais frequentes e respetivas frequências estimadas (`D`)

```
D ← empty dictionary
words ← list of words from T
for each word in words do
    if word ∈ D then
        D[word] ← D[word] + 1
    else if |D| < k then
        D[word] ← 1
    else
        j ← arg min_{j∈D} D[j]
        D[word] ← D[j] + 1
        D ← D \{j}
    end if
end for
return D
```

ao aplicar o algortimo aos dados ...

TABELA V: CAPTION top10 com space saving 10 para os idiomas ingles e finlandes

EN		FI	
Palavra	#	Palavra	#
boy	1714	poja	1808
time	1713	tyytyväinen	1807
say	1713	olinpa	1807
great	1713	sentään	1807
complacency	1713	hullunkurinen	1807
ridiculous	1713	näköinen	1807
puppet	1713	marionetti	1807
glad	1713	onnellinen	1807
behave	1713	muuttua	1807
little	1713	oikeaksi	1807

quando o k é 10, ve se que todos os valores têm o mesmo valor aproximadamente, o que nao corresponde a realidade, tanto para os resultados dos idiomas apresentados, como para o italiano. isto deve se ao facto de o k ser muito pequeno, e como tal, o algoritmo nao consegue manter a contagem de todas as palavras.

- https://www.vldb.org/pvldb/vol15/p1215-zhao.pdf
- All items whose true count is ¿ n / k are stored ! : n = total number of items, k = number of counters (k = 20 n chega)
- Smallest count min cannot be larger than e x n: IMPLICA AUMENTAR O K ? SIM!!!!

fazer em % a qnt de k pq é normal ser maior e tal, mas meter o grafico a explicar, posso meter o k = 5 para mostrar os maus resultados e tal e dps meto o grafico e mostro a escolha e tudo mais bla bla

importante dizer q contagens significativas sao aquelas q nao esta tao proximas da media por causa da situacao do n/k +- 1%

MOSTRAR GRAFICO E DIZER Q OSAO OS PONTINHOS E Q A LINHA É TP OS 10 SIGNIFICATIVOS, O QUE É O MAIS RELEVANTE QUE PQ A ANALISE ESTA MAIORITARIAMENTE FOCADA NO TOP 10

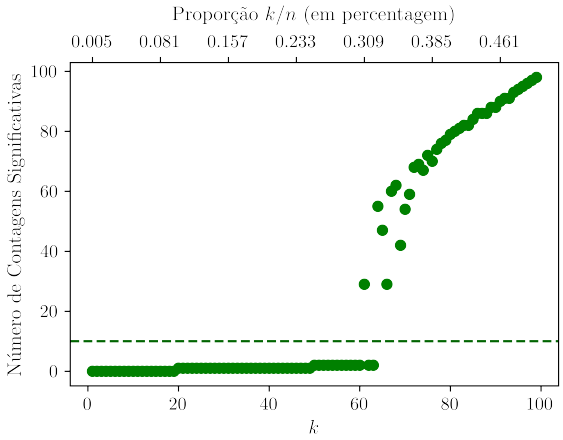

Fig. 3: qnt de palavras significataivas em funcao de k para o idioma italiano

falar tbm das percentagens q estao em cima e q, q mostram q o k é bem menor q o numero total de palavras

...

apos analise do valor minimo de k para que haja uma contagem significativa maior que 10, foi observarvado os seguintes valores minimos

TABELA VI: CAPTION s

EN		IT		FI	
k	k/n	k	k/n	k	k/n
59	0.344%	61	0.309%	67	0.371%

como todos os valores, tanto de k sao proximos, de modo a garantir uma certa coerencia ou igualdade ou criterio na comapracao dos resultados, sera aplicado o space savving a todos os livros com k = 70, visto não ser um valor significativo de memoria e ser suficiente para garantir uma contagem significativa de palavras para os 3 livros

para k = 70 temos (falta meter cores e resultado real? SE O TOP10 ESTIVER CERTO, METER COR NA CONTAGEM COM BASE NO DESVIO ABSOLUTO)

TABELA VII: CAPTION top10 palavras do SS70

EN		IT		FI	
Palavra	#	Palavra	#	Palavra	#
pinocchio	457	pinocchio	466	pinocchio	443
say	288	il	389	sanoa	266
little	270	dire	283	pieni	257
good	244	burattino	283	geppetto	257
snail	243	pesce	280	ihmeellinen	256
fairy	243	si	279	isä	256
new	242	ragazzo	279	vanha	256
boy	242	ciuchino	279	istua	256
look	242	geppetto	278	pää	256
geppetto	242	fata	278	äkkinäinen	256

os resultados n foram os mlhores, ainda ha confusao ali e talvez paavras nao significativas, pelo que eventualmente podem ser procuradas melhores formas em vez do intervalo em torno da media, mas reparase numa melgoria substancial face ao ss5

assim, de forma empirica, foi selecionado um k = 150

TABELA VIII: CAPTION top10 palavras do SS150

EN		IT		FI	
Palavra	#	Palavra	#	Palavra	#
pinocchio	457	pinocchio	461	pinocchio	443
say	284	il	386	sanoa	261
little	239	dire	282	saada	143
puppet	209	si	251	marionetti	138
boy	150	burattino	225	alkaa	136
come	145	volere	168	tehdä	135
good	135	vedere	152	isä	128
donkey	134	ragazzo	145	pieni	121
like	133	bello	138	päivä	119
go	129	andare	137	pois	119

observa se uma melhoria substancial nos resultados, principalemnte relativamente ao idioma ingles. idiomas como o finalndes n tiveram uma melhoria tao significativa, mas ainda assim, a contagem é mais precisa e mais proxima da realidade. esta melhoria nao tao significata para a FI deve se a maior quantiade de palavras distintas, o que torna mais dificil a contagem de palavras significativas, mesmo com um k maior e tambem por causa da contagem mais proxima das palavras q estao mais baixo no rank do top 10

isto releva.... (trade off, limitacoes, etc, cada caso, testes de varios k,....)

VI. Resultados

llalalla

A. analise de memoria

TABELA IX: analise de memoria, em bytes? para cada alg

idioma	Exact	Appro	SS10	SS70	SS150
EN	525112	359432	7984	18832	33264
IT	865568	572312	8064	18992	33664
FI	1332984	884776	8176	20000	35888

B. analise de precisao?

tendo em conta o rank

TABELA X: ranks top10 para cada alg, em ingles

word	exact	aprox	ss10	ss70	ss150
pinocchio	1	1	-	1	1
say	2	2	3	2	2
little	3	3	10	3	3
puppet	4	4	7	20	4
come	5	5	-	-	6
boy	6	6	1	8	5
like	7	9	-	-	9
good	8	10	-	4	7
poor	9	8	-	-	11
go	10	7	-	11	10

TABELA XI: ranks top10 para cada alg, em italiano

word	exact	aprox	ss10	ss70	ss150
pinocchio	1	1	-	1	1
il	2	2	-	2	2
dire	3	3	-	3	3
si	4	4	-	6	4
burattino	5	5	-	4	5
volere	6	6	-	-	6
vedere	7	9	-	-	7
andare	8	10	-	-	10
povero	9	7	-	-	12
ragazzo	10	12	-	7	8

TABELA XII: ranks top10 para cada alg, em finlandes

word	exact	aprox	ss10	ss70	ss150
pinocchio	1	1	-	1	1
sanoa	2	2	-	2	2
saada	3	3	-	47	3
alkaa	4	6	-	24	5
tehdä	5	4	-	25	6
marionetti	6	5	7	19	4
poika	7	14	-	-	17
huutaa	8	8	-	-	135
nähdä	9	12	-	-	13
kysyä	10	11	-	-	14

- NAO TE PREOCUPES TANTO COM COMPLEXIDADE ESPACIAL, TENS DE TER MAIS EM CONTA OS BYTES... talvez fazer analise de bytes dentro de cada seccao e dps no fim fazer uma tabela ou comparacao dos metodos ou assim
- aa - ver erros e isso nos resultados das contagens, de forma quantificada e grafica se der
- analysis of the computational efficiency and limitations of the developed approaches ()
- absolute and relative errors (lowest value, highest value, average value, etc.)
- evaluate the quality of estimates
- same most frequent / less frequent words are identified, and in the same relative order
- most frequent / less frequent words are similar in the text files of the same book in different language
- Compare the performance between themselves and regarding the exact counts.

NAS CONTAGENS IGUAIS HA RANKS DIFERENTES, DEVIA METER IGUAL, ACHO Q ISSO SO INTERSSA NA TABELA AGR NA ULTIMA, MAS CONFIRMAR

VII. Conclusão

conclusaoooo

Bibliografia

[1] Amazon Web Services, "What is text analysis?", 2024, https://aws.amazon.com/what-is/text-analysis/. Accessed: 2024-12-11.

[2] Hongyan Liu, Ying Lu, Jiawei Han, e Jun He, "Error-adaptive and time-aware maintenance of frequency counts over data streams", 2006.

[3] Project Gutenberg Literary Archive Foundation, "Project gutenberg", https://www.gutenberg.org. Accessed: 2024-12-11.

[4] Robert Morris, "Counting large numbers of events in small registers", *Commun. ACM*, vol. 21, no. 10, pp. 840–842, 1978.

[5] Graham Cormode e Marios Hadjieleftheriou, "Finding the frequent items in streams of data", *Commun. ACM*, vol. 52, no. 10, pp. 97–105, 2009.