Análise de Frequência de Palavras

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].

Assim, este relatório visa explorar três abordagens para este problema: contadores exatos, contadores aproximados e identificação de itens frequentes em *streams* de dados.

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, permitindo para além de uma comparação da frequência de palavras em diferentes idiomas, uma 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*.

Posteriormente, foram implementadas as abordagens referidas anteriormente, nomeadamente contadores exatos, contadores aproximados e identificação de itens frequentes em *streams* de dados, com o objetivo de identificar as 10 palavras mais frequentes em cada um dos livros em análise.

Por fim, foram realizadas análises comparativas entre os resultados obtidos com as diferentes abordagens, nomeadamente em termos de .... !!!

Atendendo a estas últimas fases, estas serão apresentadas em detalhe nas secções seguintes, com a apresentação dos resultados obtidos e a respetiva análise.

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
```

Através da aplicacao do algoritmo de contagem exata, foi possível identificar as 10 palavras mais frequentes

em cada um dos livros em análise. A Tabela I apresenta as palavras mais frequentes em cada idioma, juntamente com o número de ocorrências de cada palavra (#).

TABELA I: 10 palavras mais frequentes em cada idioma (livro).

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 as diferentes traduções, nomeadamente a proximidade das frequências de palavras que têm o mesmo significado entre os idiomas. Por exemplo, as palavras "puppet", "burattino"e "marionetti", que significam marioneta, têm frequências semelhantes nos três idiomas, tal como as que significam rapaz ("boy", "ragazzo"e "poika"), dizer ("say", "dire"e "sanoa"), entre outras.

Ademais, atendendo à análise das palavras menos frequentes, é mais complexo identificar semelhanças entre as traduções, uma vez que a frequência de palavras menos frequentes é muito maior, tornando-se complexa a comparação entre os idiomas. Isto é comprovado pela distribuição da frequência de palavras apresentada na Fig. 1, que demonstra um enorme número de palavras cuja contagem pequena. Ainda através da mesma visualização, é possível identificar que a distribuição da frequência de palavras é semelhante entre os diferentes idiomas, com uma grande quantidade de palavras com contagens baixas e uma diminuição exponencial da frequência à medida que a contagem aumenta.

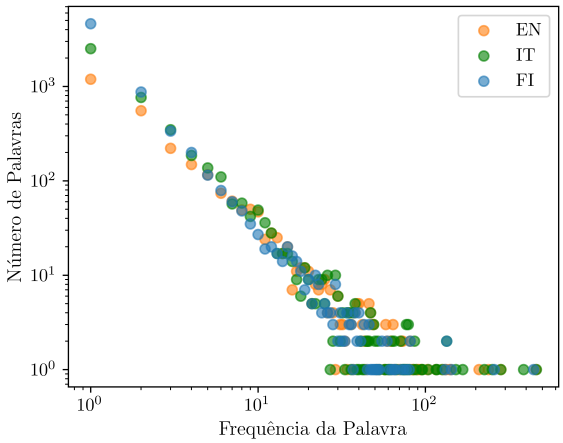

Fig. 1: Distribuição da frequência de palavras, por idioma.

Por fim, ainda através da aplicação do algoritmo de contagem exata, torna-se possível a análise da quantidade de palavras distintas em cada idioma, como apresentado na Fig. 2. Neste caso, é possível observar que a quantidade aumenta de idioma para idioma, inglês (EN) < italiano (IT) < finlandês (FI), o que pode não representar uma distinção entre as traduções, pelo facto deste desequilíbrio poder ser influenciado por fatores como a complexidade da língua ou o desempenho do *spaCy* na lematização das palavras.

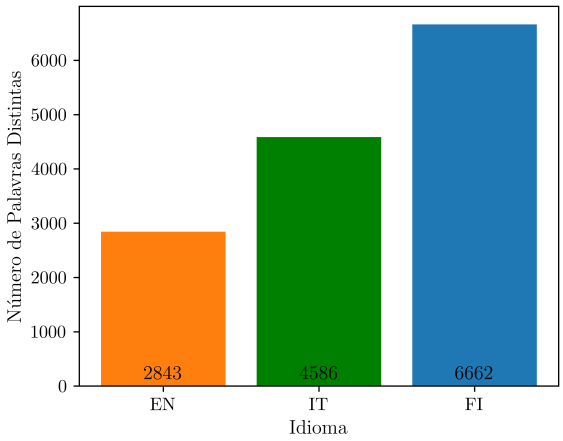

Fig. 2: Número de palavras distintas analisadas pelo algoritmo, em cada idioma.

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, pelo que apenas é estudado um caso particular deste tipo de algoritmos neste relatório. 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
```

Pelo facto deste algoritmo ter uma componente aleatória, para a obtenção dos resultados relativos ao mesmo, esta abordagem foi aplicada a cada livro em análise 20 vezes. Posteriormente, para cada uma das traduções, foi calculada a contagem (#) mínima, média e máxima das palavras, sendo selecionadas as 10 palavras mais frequentes com base na média, de modo a obter uma estimativa das palavras mais frequentes em cada idioma. Estas contagens, tal como a contagem exata ($\#_{\text{real}}$), são apresentadas nas seguintes tabelas (Tabela II a IV). Nestas tabelas, para além das contagens referidas, também é possível verificar um código de cores nas palavras, preto, laranja e vermelho, que indicam se a posição da palavra está correta na ordem das 10 palavras mais frequentes, se está deslocada, ou se não faz parte das 10 mais frequentes de todo, respetivamente.

TABELA II: Estimativa das 10 palavras mais frequentes, pelo algoritmo de contagem aproximada, para o livro em inglês (IN).

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: Estimativa das 10 palavras mais frequentes, pelo algoritmo de contagem aproximada, para o livro em italiano (IT).

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: Estimativa das 10 palavras mais frequentes, pelo algoritmo de contagem aproximada, para o livro em finlandês (FI).

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

Assim, é possível verificar que este algoritmo é tanto melhor quanto maior é a contagem real da palavra, o que se deve ao facto de haver uma menor densidade de palavras com contagens próximas, como foi verificado na contagem exata (Fig. 1) e ao facto da probabilidade de contagem ser fixa, o que implica que a contagem de

palavras com contagens mais baixas seja menos precisa.

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
```

Pelo facto deste algoritmo ser sensível ao valor de k, torna-se importante a realização de uma análise para a determinação do valor mínimo de k que permita a obtenção de resultados precisos, neste caso das 10 palavras mais frequentes, sem comprometer a eficiência espacial do algoritmo [6]. Na Tabela V pode-se observar um exemplo de aplicação do algoritmo em análise, para um valor k inadequado.

TABELA V: Estimativa das 10 palavras mais frequentes, pelo algoritmo de contagem *Space-Saving*, com $k = 10$, para as traduções em inglês e finlandês.

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

Verifica-se que, quando o k é 10, todas as contagens têm o mesmo valor (k/n, onde n é o número total de palavras), aproximadamente, o que não corresponde à realidade. Isto deve-se ao facto de o k ser demasiado pequeno, e como tal, o algoritmo não consegue manter uma contagem precisa das palavras mais frequentes.

Tendo em vista a determinação do valor mínimo de k, com o objetivo de determinar os 10 itens mais frequentes, foi criada uma medida de contagens significativas, que consiste no número de contagens que estão a uma distância significativa do valor k/n, neste caso, de pelo menos 1%. A Fig. 3 apresenta a quantidade de palavras significativas em função de k, para o idioma italiano, e uma linha horizontal que representa o valor 10, o número de palavras a identificar. Assim, é esperado que o k mínimo seja o valor mais baixo que permita a identificação de pelo menos 10 palavras significativas (primeiro ponto acima da linha). É também importante referir que, ainda a partir da visualização, é possível observar a proporção entre k e o número total de palavras (n) (eixo superior), para cada k, permitindo uma perceção da eficiência espacial do algoritmo.

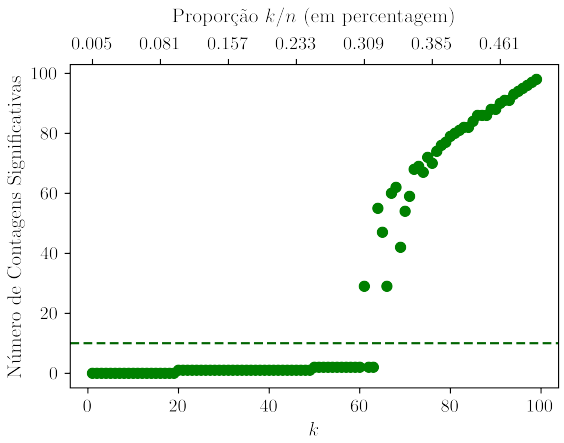

Fig. 3: Número de palavras significativas em função do valor de k, para a tradução em italiano.

Através de uma análise numérica, com a metodologia referida, envolvendo o número de palavras significativas, foi possível determinar o valor mínimo de k para cada uma das traduções, sendo estes apresentados na Tabela VI.

TABELA VI: Valor mínimo de k para que o número de palavras significativas seja superior a 10, para cada tradução.

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

Pelo facto de todos os valores de k obtidos a partir da análise da tabela anterior serem próximos e não representarem um valor significativo de memória, foi selecionado o valor $k = 70$ para a aplicação do algoritmo de contagem *Space-Saving* a todos os livros em análise. A Tabela VII apresenta as 10 palavras mais frequentes em cada um dos livros, com base na aplicação do algoritmo com $k = 70$, juntamente com a sua contagem estimada, e um código de cores nas palavras, que indica se a posição da palavra está correta na ordem das 10 palavras mais frequentes (preto), se está deslocada (laranja), ou se não faz parte das 10 mais frequentes de todo (vermelho).

TABELA VII: Estimativa das 10 palavras mais frequentes, pelo algoritmo de contagem *Space-Saving*, com $k = 70$.

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

Pelo facto dos resultados obtidos com o algoritmo de contagem *Space-Saving* com $k = 70$ não serem suficientemente precisos, devido ao valor de k e por existirem várias contagens muito próximas, nas 10 palavras mais frequentes, pode-se concluir que a medida de contagens significativas poderá ter de ser melhorada, possivelmente aumentando a distância que determina a significância de uma contagem.
AQUI AQUI AQUI AQUI AQUI
Não sendo a determinação desta medida o foco deste estudo, foi selecionado empiricamente um valor de $k = 150$, para .....

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
- analysis of the computational efficiency and limitations of the developed approaches ()
- evaluate the quality of estimates

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.

[6] Fuheng Zhao, Divyakant Agrawal, Amr El Abbadi, e Ahmed Metwally, "Spacesaving ±: an optimal algorithm for frequency estimation and frequent items in the bounded-deletion model", *Proceedings of the VLDB Endowment*, vol. 15, no. 6, pp. 1215–1227, 2022.