Most Frequent Letters – 3rd Assigment

Sobral, Pedro – 98491 – sobral@ua.pt

*Resumo* – Este relatório procura mostrar através de diferentes aborgagens algorítmicas como contar o número de vezes que uma determinada letra aparece numa determinada obra literária. Todos os algoritmos foram escritos em Python (3.10), e serão feitas analises sobre a complexidade computacional, os erros relativos e absolutos das contagens, entre outros tipos de comparação entre os algoritmos.

*Abstract* – This report seeks to through different algorithmic approaches, how to count the number of times a given letter appears in a given literary work. All algorithms were written in Python (3.10), and analyzes will be made on computational complexity, relative and absolute errors of counts, among other types of comparison between algorithms.

# Introdução

Através deste trabalho é pedido que se façam comparações em diferentes nível e modos às 3 abordagens distintas da frequência das letras em determinadas obras literárias. Além deste relatório, foram criados os ficheiros de código em Python, e os ficheiros que guardam os resultados da implementação dos algoritmos. Em relação aos algoritmos desenvolvidos, como já referido há 3 tipos de abordagens, uma que conta o numero exato de letras, uma abordagem que faz uma contagem aproximada, e uma outra que faz a contagem de acordo com o LossyCount.

De modo a correr o algoritmo, basta correr o seguinte comando dentro da pasta /src, ou ler o README do projeto. O ficheiro “process\_data.py” serve para adaptar as obras literárias para o formato que o enunciado do projeto pede, e o ficheiro “main.py” para correr os algoritmos e obter os resultados.

$ python3 process\_data.py

$ python3 main.py

# II. Preprocessamento dos Dados

Todas as obras literárias foram consumidas do Projeto Guttenber REFERENCIA, e foram analisadas as seguintes obras:

- Alice no País das Maravilhas em Francês, Alemão, e em Inglês.

- Uma obra do filosofo Friedrich Nietzsche em Inglês.

Em todas as obras referidas anteriormente o cabeçalho dos ficheiros foi retirado, foram removidas as stopwords, os acentos, e as letras foram todas convertidas para maiúscula.

Assim feito, os dados (obras literárias) estão prontas para serem analisados pelos algoritmos de contagem de letras.

# III. Algoritmo de contagem exata

A abordagem para a criação deste algoritmo foi a seguinte:

Criar um dicionário onde o par chave/valor seja letra/numero de ocorrências, desta forma é criada uma iteração sobre as obras literárias de modo a passar por todas as letras e assim poder fazer a contagem certa das letras.

# IV. Algoritmo de contagem aproximada

Este tipo de algoritmo permite contar um número gigante de letras usando poucos recursos a nível de memória. Para esta implementação há um incremento de 1/sqrt(2)^k para os eventos. [1]

# V. Algoritmo Lossy-Count

O algoritmo Lossy-count é direcionado para identificar elementos num data stream cuja frequência excede o threshold inserido pelo utilizador. A frequência neste caso de letras nem sempre é inteiramente precisa, no entanto tem um limite de erro, o threshold tal como referido anterirormente. [2]

# V. Resultados

A nível de resultados, os mesmos serão dividos primeiramente numa avaliação sobre cada obra literária, e posteriormente será feita uma analise comparativa em relação as várias obras, e que conclusões se podem tirar por exemplo da mesma obra, mas escrita em línguas diferentes.

1. Alice no País das maravilhas em Francês.

Calendar

Description automatically generated with low confidence

A picture containing calendar

Description automatically generated

Tabela 2 - Erros Absolutos e Relativos

Analisando a Tabela 1, podemos averiguar que o a frequência das letras é igual no Contador Exato e no Contador “Lossy Counter”, relativamente a este útlimo algoritmo o threshold usado foi: 0.00000775. Já quanto ao Algoritmo de contagem aproximada, não se pode dizer o mesmo, a frequência das letras na mesma obra literária é bastante mais pequena comparado com os outros algoritmos.

Tabela 1 - Frequência das Letras

Ainda nesta tabela, podemos fazer a comparação entre as três aproximações, a versão mais rápida é a de contagem exata, com 0.02 segundos, sendo também aquela que nos dá os resultados sempre corretos, o algoritmo que se segue é o algoritmo Lossy-count com 0.029 segundos, e por último o algoritmo de contagem aproximada com um tempo de execução de 0.085 segundos, sendo portanto este o algoritmo mais demorado, sendo aproximadamente 3 vezes mais lento em relação aos outros 2, e sendo excessivamente pior a nível de resultados comparado com os outros algoritmos.

As 3,5,10, letras mais frequentes no algoritmo Lossy-Count são respetivamente:

- E, A, T

- E, A, T, I, S

- E, A, T, I, S, R, N, U, L, O

Que comparando com o algoritmo de contagem exata, verifica-se que o Top 3, 5, e 10, são iguais aos do algoritmo Lossy-Count.

Vendo os erros quer relativos quer absolutos do algoritmo [3] de contagem aproximada comparando com os resultados do algoritmo de contagem exata, tal como já visto na tabela 1, a maioria do erro relatico de cada letra é sempre à volta de 99%. Como se pode ver na parte final é possível ver alguns cálculos relativamente aos erros, de entre os quais a média, que a nível absoluto, o algoritmo não conta cerca de 4862 em cada letra, representado um erro relativo de 98.5%.

1. Calendar

   Description automatically generatedAlice no País das Maravilhas em Alemão

Tabela 3 - Frequência das Letras

A picture containing calendar

Description automatically generated

Tabela 4 - Erros Absolutos e Relativos

Analisando a Tabela 3, podemos averiguar que o a frequência das letras é diferente no Contador Exato e no Contador “Lossy Counter”, relativamente a este útlimo algoritmo o threshold usado foi: 0.00000775.

Ainda nesta tabela, podemos fazer a comparação entre as três aproximações, a versão mais rápida é a de contagem exata, com 0.02 segundos, sendo também aquela que nos dá os resultados sempre corretos, o algoritmo que se segue é o algoritmo Lossy-count com 0.027 segundos, e por último o algoritmo de contagem aproximada com um tempo de execução de 0.084 segundos, sendo portanto este o algoritmo mais demorado, sendo aproximadamente 3 vezes mais lento em relação aos outros 2, e sendo excessivamente pior a nível de resultados comparado com os outros algoritmos.

As 3,5,10, letras mais frequentes no algoritmo Lossy-Count são respetivamente:

- E, N, I

- E, N, I, A, R

- E, N, I, A, R, S, T, H, D, U

Que comparando com o algoritmo de contagem exata, verifica-se que o Top 3, 5, e 10, são iguais aos do algoritmo Lossy-Count.

Vendo os erros quer relativos quer absolutos do algoritmo [3] de contagem aproximada comparando com os resultados do algoritmo de contagem exata, tal como já visto na tabela 3, a maioria do erro relatico de cada letra é sempre à volta de 99%. Como se pode ver na parte final é possível ver alguns cálculos relativamente aos erros, de entre os quais a média, que a nível absoluto, o algoritmo não conta cerca de 4947 em cada letra, representado um erro relativo de 97.3%.

1. Alice no País das Maravilhas em Inglês

Calendar

Description automatically generated with medium confidence

Tabela 5 - Frequência das Letras

A picture containing text

Description automatically generated

Tabela 6 - Tabela 4 - Erros Absolutos e Relativos

Analisando a Tabela 5, podemos averiguar que o a frequência das letras é igual no Contador Exato e no Contador “Lossy Counter”, relativamente a este útlimo algoritmo o threshold usado foi: 0.00000775. Já quanto ao Algoritmo de contagem aproximada, não se pode dizer o mesmo, a frequência das letras na mesma obra literária é bastante mais pequena comparado com os outros algoritmos.

Ainda nesta tabela, podemos fazer a comparação entre as três aproximações, a versão mais rápida é a de contagem exata, com 0.011 segundos, sendo também aquela que nos dá os resultados sempre corretos, o algoritmo que se segue é o algoritmo Lossy-count com 0.015 segundos, e por último o algoritmo de contagem aproximada com um tempo de execução de 0.045 segundos, sendo portanto este o algoritmo mais demorado, sendo aproximadamente 3 vezes mais lento em relação aos outros 2, e sendo excessivamente pior a nível de resultados comparado com os outros algoritmos.

As 3,5,10, letras mais frequentes no algoritmo Lossy-Count são respetivamente:

- E, T, I

- E, T, I, A, O

- E, T, I, A, O, N, S, R, L, D

Que comparando com o algoritmo de contagem exata, verifica-se que o Top 3, 5, e 10, são iguais aos do algoritmo Lossy-Count.

Vendo os erros quer relativos quer absolutos do algoritmo [3] de contagem aproximada comparando com os resultados do algoritmo de contagem exata, tal como já visto na tabela 5, a maioria do erro relatico de cada letra é sempre à volta de 99%. Como se pode ver na parte final é possível ver alguns cálculos relativamente aos erros, de entre os quais a média, que a nível absoluto, o algoritmo não conta cerca de 2727 em cada letra, representado um erro relativo de 97.9%.

1. Conclusões sobre a obra “Alice no País das Maravilhas” nas diferentes línguas.

Fazendo agora uma comparação entre as 3 diferentes línguas que foram previamente analisadas, podemos logo à partida verificar que a obra em Inglês tem sensivelmente metade das letras das restantes obras nas outras línguas analisadas.

A nível do erro absoluto médio é na versão em inglês, que este é menor, também muito em caso derivado ao número de letras ser sensivelmente metade tal como já referido no segmento anterior. Já relativamente ao erro relatico maior o mesmo é encontrado na obra em Francês.

O erro relativo menor é encontrado na obra em Alemão, e o maior nas obras em Alemão e em Francês sendo o memso de 0.999.

Podemos fazer algumas observações como: A letra mais utilizada nas 3 diferentes linguas é a letra “E”

1. Uma obra do filosofo Friedrich Nietzsche em Inglês.

A picture containing calendar

Description automatically generated

Tabela 7 - Frequência das Letras

Text

Description automatically generated with low confidence

Tabela 8 - Erros Absolutos e Relativos

Analisando a Tabela 7, podemos averiguar que o a frequência das letras não é igual no Contador Exato e no Contador “Lossy Counter”, relativamente a este útlimo algoritmo o threshold usado foi: 0.00000775, no entanto apresenta valores signigicativamente melhores que o algoritmo de contagem aproximada.

Ainda nesta tabela, podemos fazer a comparação entre as três aproximações, a versão mais rápida é a de contagem exata, com 0.032 segundos, sendo também aquela que nos dá os resultados sempre corretos, o algoritmo que se segue é o algoritmo Lossy-count com 0.044 segundos, e por último o algoritmo de contagem aproximada com um tempo de execução de 0.138 segundos, sendo portanto este o algoritmo mais demorado, sendo aproximadamente 3 vezes mais lento em relação aos outros 2, e sendo excessivamente pior a nível de resultados comparado com os outros algoritmos.

As 3,5,10, letras mais frequentes no algoritmo Lossy-Count são respetivamente:

- E, S, I

- E, T, I, N, T

- E, T, I, N, T, A, R, O, L, D

Que comparando com o algoritmo de contagem exata, verifica-se que o Top 3, 5, e 10, são iguais aos do algoritmo Lossy-Count.

Vendo os erros quer relativos quer absolutos do algoritmo [3] de contagem aproximada comparando com os resultados do algoritmo de contagem exata, tal como já visto na tabela 7, a maioria do erro relatico de cada letra é sempre à volta de 99%. Como se pode ver na parte final é possível ver alguns cálculos relativamente aos erros, de entre os quais a média, que a nível absoluto, o algoritmo não conta cerca de 8123 em cada letra, representado um erro relativo de 98.9%.

# VI. Conclusão

Tal como proposto no início do trabalho foram implementados os 3 algoritmos de contagem mencionados e foi feita uma avaliação elaborada dos algoritmos a vários níveis.

O algoritmo Lossy-Count é tendencialmente bom, e será tão bom quanto menor o valor do threshold utilizado, no entanto o valor para uma comparação mais justa o valor foi fixo para todos os testes feitos de modo a ser possível assim comparar de igual forma os resultados obtidos.

Pelos resultados podemos confirmar que o algoritmo de pesquisa por aproximação é de facto mau quer a nível de retornar um bom resultado, quer a nível de tempo de execução comparando com os outros algoritomos, e o algoritmo Lossy-Count é na sua maioria bom, tanto melhor quanto menor for o valor do threshold tal como referido, mas mesmo quando não apresenta resultados iguais aos da contagem exata, apresenta resultados substancialmente superiores aos do algoritmo de contagem por aproximação.

# VII. Referências

[1] - <https://www.geeksforgeeks.org/approximation-algorithms/>

[2] - <https://en.wikipedia.org/wiki/Lossy_Count_Algorithm>

[3] - <https://www.vedantu.com/maths/absolute-and-relative-error>