Projecto 3: Contagem de palavras mais frequentes

Samuel Vieira, 89055

***Abstract* – The frequenct items problem is approached in this project. The aim of this project is to count the number of words from a given text, explore what happens if we change the parameters of the function containing the lossy algorithm and compare with the exact count results.**

***Resumo* – O problema dos items frequentes é abordado neste projeto. O objetivo deste projeto é contar o número de palavras de um determinado texto, explorar o que acontece se mudar os parâmetros da função que contém o algoritmo Lossy e comparar estes resultados com os reultados do algoritmo de contagem exata.**

***Keywords* – Words, datastream, Frequency counts, lossy counting**

***Palavras chave* – Palavras, "datastream", contagem de frequência, contagem Lossy**

I. Introdução

Dados do tipo *datastream*, são bastante usados em áreas de telecomunicações. Esta "datastream" pode ser visto como uma colecção de dados transmitidos continuamente. No contexto deste trabalho, será feita a contagem de uma "stream" de palavras, que serão os nossos items em análise, figura 1. Este problema é bastante famoso na área de "datastream". [1,2,3,4]

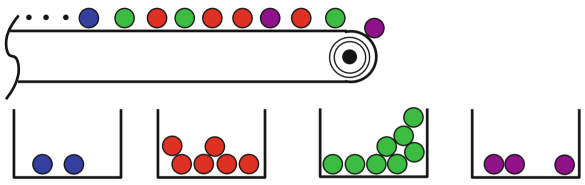

Fig. 1 - Esquema de uma "stream" de items cuja as diferentes frequencias de cada item forma uma certa distribuição.

Na secção II é feita uma breve abordagem à explicação do algoritmo Lossy, na secção III irá ser feita uma análise dos dados, assim como uma breve discussão. Finalmente na ultima secção faz-se um breve resumo do trabalho e retira-se pequenas conclusões.

II. Algoritmo de contagem Lossy

O algoritmo de contagem Lossy, ver algoritmo 1, consiste na divisão de uma "stream" em várias partes. Cada parte da "stream" é denominado por "bucket" e este tem uma largura (que é o inverso do erro) e um índice associado ao mesmo. Este índice é definido como o tamanho atual da "stream" dividido pela largura dos "buckets" [1,2].

Cada iteração feita por cada palavra na "stream", é adicionada uma entrada ao dicionário caso a palavra não exista, e incrementa-se o contador associado a essa palavra caso esta já esteja presente no dicionário [1,2].

Cada vez que o resto da divisão entre o tamanho atual da "datastream" e largura do "bucket" for igual a zero, todas as entradas que tenham a sua contagem menor ou igual ao valor delta, são eliminadas e o valor do índice do "bucket" é atualizado [1,2].

Algorithm 1 Lossy

```
bucketId ← 1
currentLenght ← 1
entries ← um dicionário vazio
bucketWidth ← 1/erro

for word in datastream do
    if word in entries then
        entries[word][0] ← entries[word][0] + 1
    else
        entries[word] ← [1, bucketId-1]

    if currentLenght mod bucketWidth is 0 then
        bucketId ← currentLenght/bucketWidth
        oldEntries ← cópia de entries
        for word in oldEntries do
            count ← oldEntries[word][0]
            delta ← bucketId-oldEntries[word][1]
            if count less or equal delta then
                the entry is deleted
        currentLenght ← currentLenght + 1
```

III. Implementação das bibliotecas

A. Biblioteca Book

Esta biblioteca contém três funções: *Load()*, *FilterLetters()* e *RemoveStopWords()*. A função *Load()* carrega o texto do ficheiro que vai ser analisado, retorna o seu conteúdo numa string. A função *FilterLetters()* faz o tratamento do texto em sí, removendo pontuações, acentos entre outros e converte todas as letras para minúsculas. A ultima função remove "stop-words" (tal como "the", "i", "me", ect.)

B. Biblioteca WordCounters

Nesta biblioteca foram inseridas as funções que fazem a contagem de palavras. A primeira função, *Exact()*,

que faz a contagem exata das palavras do texto enquanto que a função *Lossy()* faz uso do algoritmo Lossy para fazer a contagem de palavras dentro de uma lista de palavras.

IV. ANÁLISE E DISCUSSÃO DOS DADOS

A. Contagem exata

O autor da obra literária, analisada neste trabalho foi escrita por William Shakespeare com o título: A tragédia de António e Cleópatra com 26556 palavras (versão inglesa) incluindo as "stop-words". As linguagens escolhidas para analise foram inglês, francês e finlandês.

Antes de fazer a remoção das "stop-words", comparou-se as palavras mais comuns em inglês [5] e na obra literária em análise. Na tabela I é possível fazer essa comparação.

TABLE I
FREQUÊNCIA DE PALAVRAS NO LIVRO E DE MODO GERAL (OEC RANK)

Palavra	Frequência no texto	Frequência no geral
the	1º lugar	1º lugar
and	2º lugar	5º lugar
to	3º lugar	3º lugar
i	4º lugar	10º lugar
of	5º lugar	4º lugar

Observando os resultados da tabela, as 5 palavras mais frequentes no texto estão pelo menos dentro do top 10 de palavras mais frequentes na língua inglesa. A palavra "the" é a mais comum na língua inglesa assim como no livro.

Passando agora para a contagem exata, ao remover todas as "stop-words" do texto, os resultados diferem bastante. Na figura 2 observa-se que os nomes das personagens "antony", "cleopatra", "caesar" e "enobarbus" aparecem com bastante frequência.

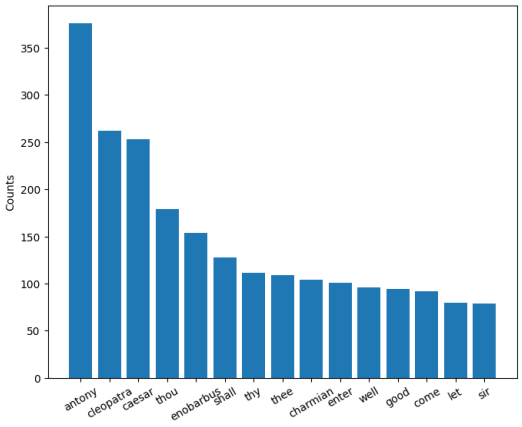

Fig. 2 - As palavras mais comuns no livro em inglês.

Ao comparar a figura 2 com as figuras 3 e 4, nota-se que a distribuição é semelhante à língua inglesa e também verifica-se que existe uma elevada semelhança entre as palavras mais frequentes, o que era de esperar já que se tratam de traduções.

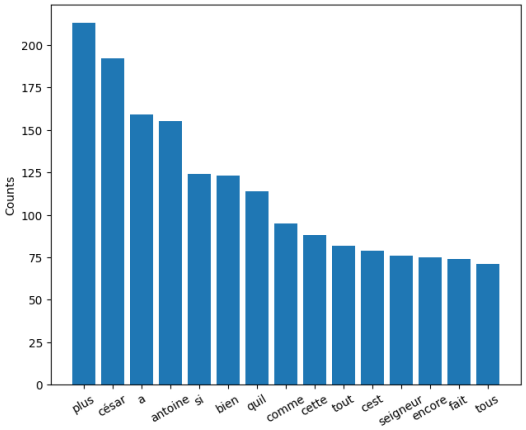

Fig. 3 - As palavras mais comuns no livro em francês.

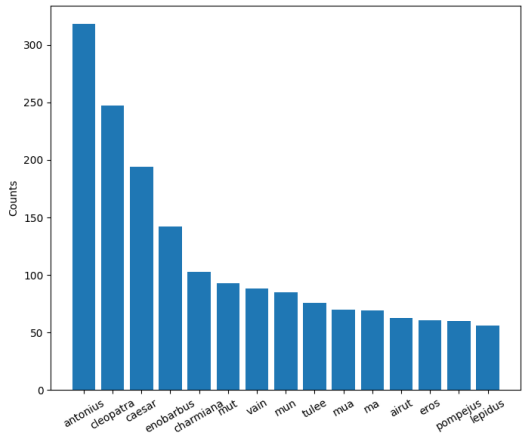

Fig. 4 - As palavras mais comuns no livro em finlandês.

B. Contagem Lossy

Ao contar o número de palavras usando o algoritmo 1 (Lossy), obtiveram-se resultados bastante semelhantes. Em primeiro lugar comparou-se as contagem exata com a contagem resultante do algoritmo Lossy, no entanto não se verificou diferença nas 15 palavras mais frequentes do texto. Isto deve-se ao facto de que o algoritmo nunca eliminou qualquer uma destas palavras ao processar o texto em si. No entanto as letras menos comuns acabam por ser eliminadas de acordo com a condição descrita na introdução do algoritmo.

Para analisar o número diferentes palavras a serem contadas, conta-se o número total de "keys" que o dicionário contém. Isto pode ser verificado na figura 4.

Ao verificar as três linhas da figura nota-se que para as três linguagens existem certas zonas dos gráficos em

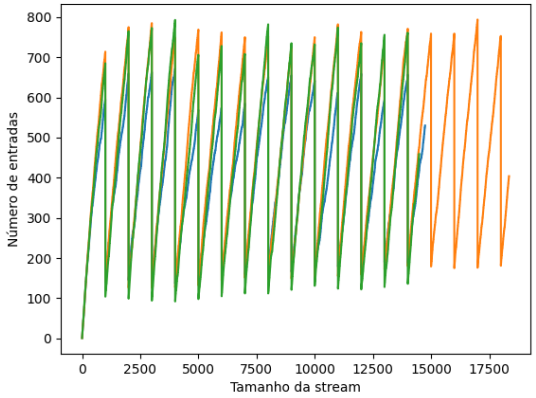

Fig. 5 - Número total de "keys" do dicionário de acordo com o tamanho da stream. Azul - língua inglesa, amarelo - língua francesa, verde - língua finlandesa. Parâmetro erro = 0,001.

que a contagem de palavras diminui abruptamente. Essas zonas em que o número de palavras diminuem correspondem à mudança do índice do "bucket". Alterou-se também o parâmetro erro, que por sua vez também altera o tamanho de cada "bucket". Nas figuras 6, 7 e 8 verifica-se as consequências de alterar o parâmetro erro.

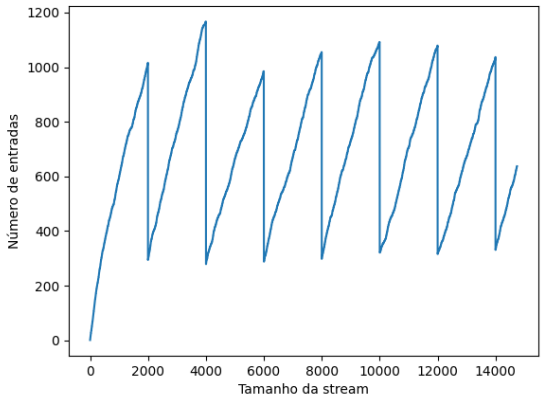

Fig. 6 - Número total de "keys" do dicionário de acordo com o tamanho da stream na língua inglesa. Parâmetro erro = 0,0005.

Como era de esperar com o aumento do parâmetro erro, existe uma diminuição no tamanho de cada "bucket" já que existe um relação de proporcionalidade inversa. Outra observação, é o facto de que existe uma diminuição da precisão quanto maior for o parâmetro do erro (logo "buckets" mais pequenos).

V. Conclusões

Neste projeto comparou-se os resultados da contagem exata de palavras em três textos, com a contagem lossy. Notou-se que os resultados de lossy desviaram-se mais quanto menor for o tamanho dos "buckets" que é controlado através do parâmetro erro definido na função *Lossy*.

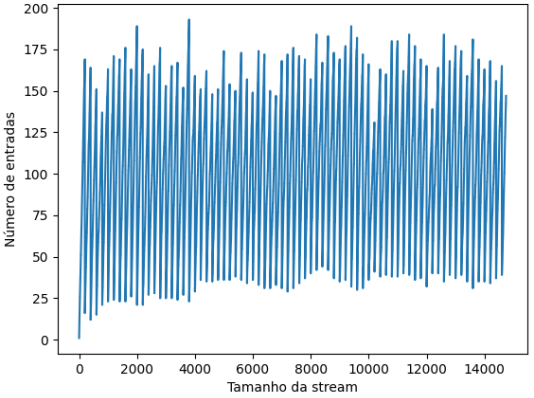

Fig. 7 - Número total de "keys" do dicionário de acordo com o tamanho da stream na língua inglesa. Parâmetro erro = 0,005.

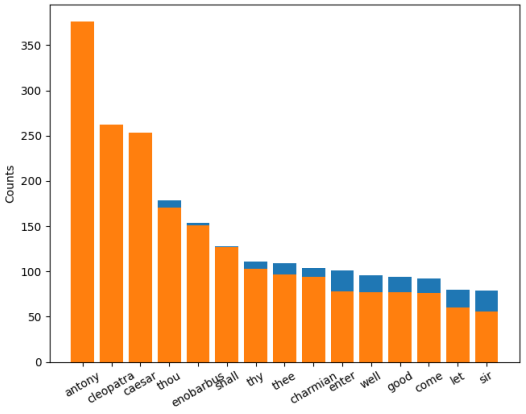

Fig. 8 - As palavras mais comuns no livro em inglês usando a contagem exata (azul) e a contagem com o algoritmo Lossy (laranja).

VI. Referências

[1]Manku, Gurmeet Singh, and Rajeev Motwani. "Approximate frequency counts over data streams." VLDB'02: Proceedings of the 28th International Conference on Very Large Databases. Morgan Kaufmann, 2002.

[2]Cormode, Graham, and Marios Hadjieleftheriou. "Methods for finding frequent items in data streams." The VLDB Journal 19.1 (2010): 3-20.

[3]https://en.wikipedia.org/wiki/Lossy_Count_Algorithm

[4]https://en.wikipedia.org/wiki/Data_stream

[5]https://en.wikipedia.org/wiki/Most_common_words_in_English