
UNIVERSIDADE DE COIMBRA

**Engenharia Informática**
**Teoria da Informação**

Trabalho Pratico nr 1

- Entropia
- Redundância
- Informação mútua

**Projeto elaborado por:**

Cíntia Dalila Luís Cumbane nr 2020244607
Edson Fernando Alage nr 2021244423

# Entropia

É definida em termos do conjunto das mensagens que a fonte pode produzir.

Seja uma fonte discreta X com M símbolos diferentes e estatisticamente independentes.

Quando o símbolo de ordem j é transmitido a informação transportada é $I_j = -\log_2 P_j$ bits.

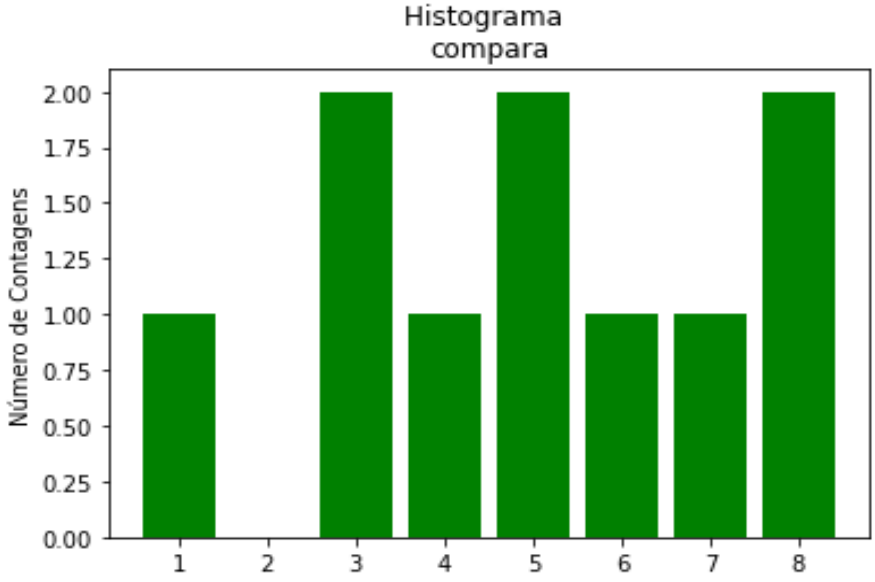
A informação média associada aos M símbolos da fonte X é a média **ponderada** das auto-informações de cada símbolo. A essa informação média por símbolo da fonte chama-se entropia e designa-se por H(X):

$$H(X) = \sum_{j=1}^{M} P_j I_j = -\sum_{j=1}^{M} P_j \log_2 P_j \text{ bits/símbolo}$$

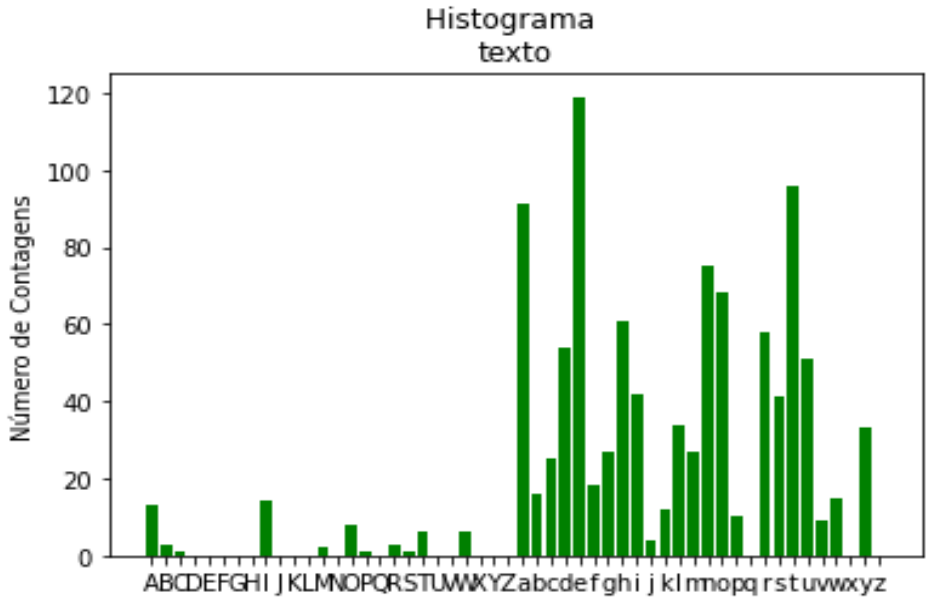
O que é que significa a **entropia** de uma fonte? Significa que:

Embora não possamos prever qual o símbolo que a fonte irá produzir a seguir, em média esperamos obter H bits de informação por símbolo, ou NH bits numa mensagem de N símbolos, se N for elevado.

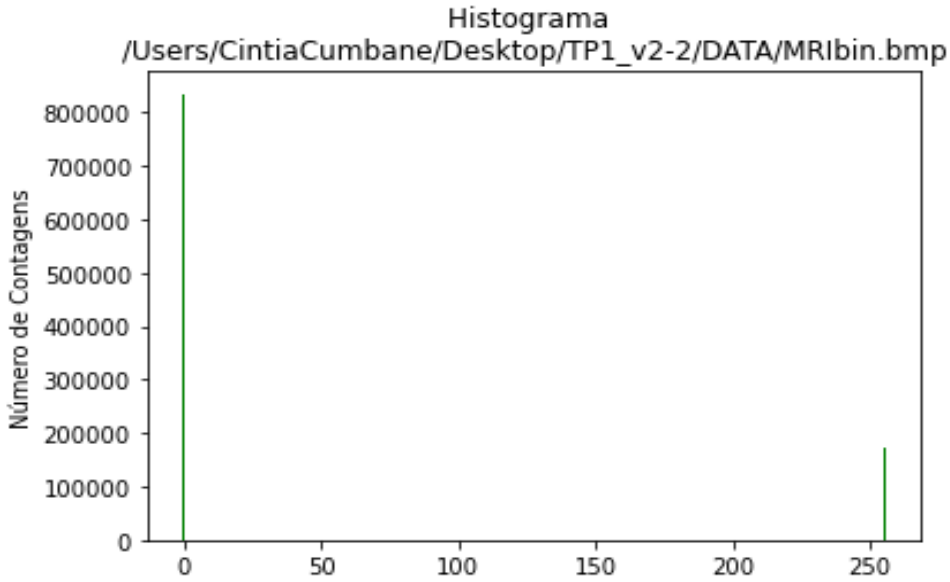
**Exercícios 1,2 e 3**



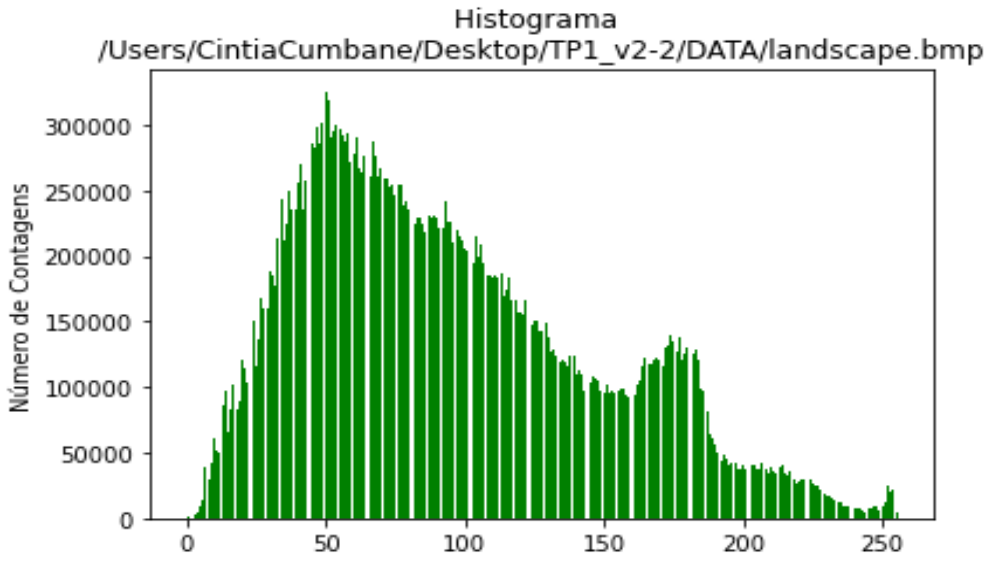

Tem-se este histograma que corresponde a uma determinada fonte de alfabeto a qual criada por nós , simplesmente para ter uma noção de como é que deve ser feito um histograma de ocorrências .



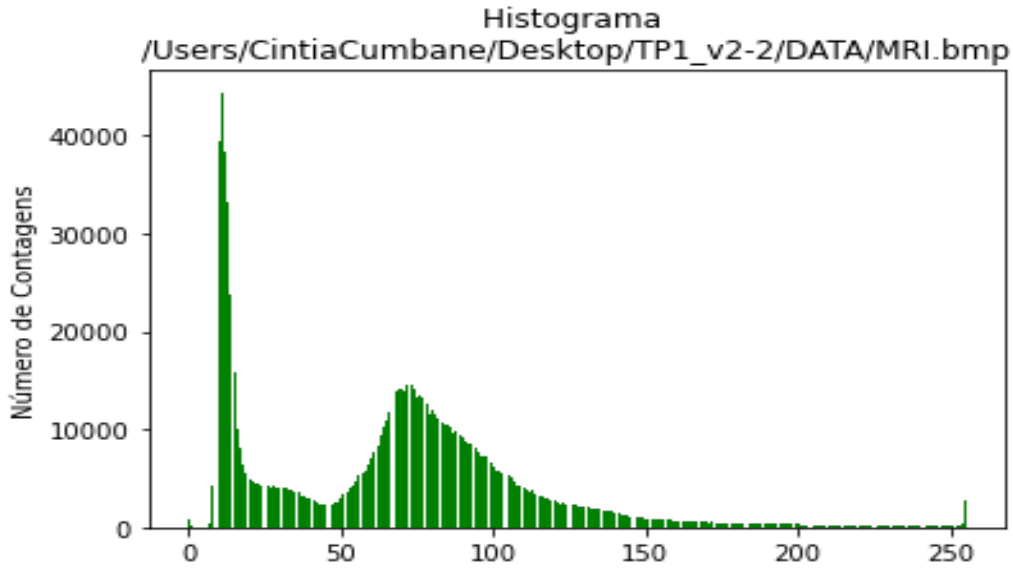
Tem-se este histograma que corresponde ao ficheiro de texto “lyrics.txt”, considerando todas as letras do alfabeto incluindo as minúscula , sem contemplar os carateres especiais, podemos verificar que a letra mais frequente de todas é a letra **e ,** as duas mais comuns são: **a, t**. De tal forma conseguimos notar mais presença do alfabeto minúsculo em relação ao maiúsculo .



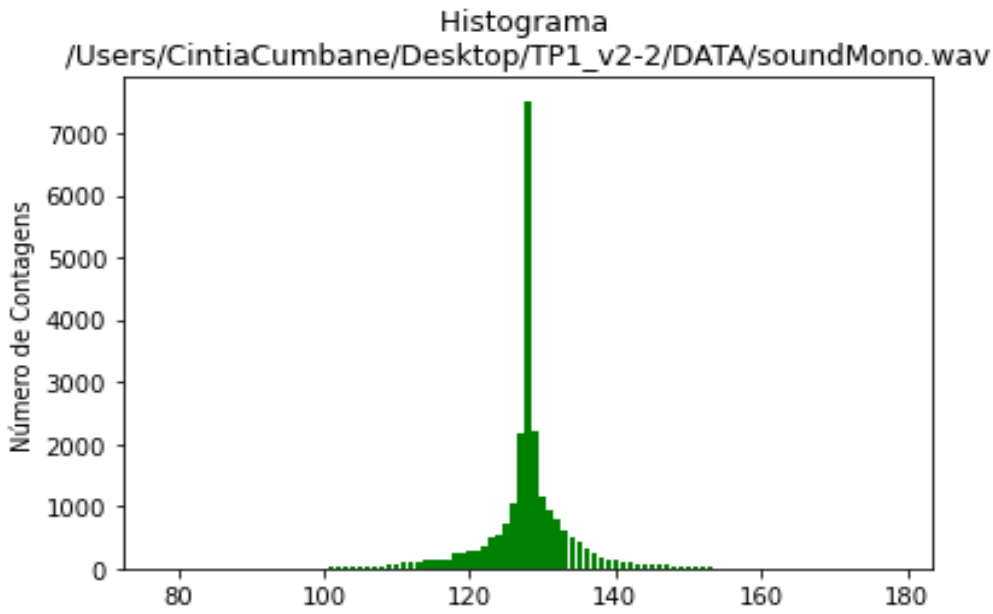

Olhando aqui para este histograma é de notar que em (0,255) somente toma dois valores dos quais 0 e 1 , contudo não há dúvidas de que é assim que o histograma deve apresentar-se .




Relativamente a este histograma da  para perceber que trata-se uma imagem muito mais colorida , de tal forma que o histograma aparece assim




Neste histograma conseguimos ver uma ligeira mistura de cores , mas também conforme se pode ver  que a cor cinza é a mais frequente .




E por fim tem-se este histograma o qual podemos verificar uma simetria em relação a ocorrência dos seus símbolos .

Na tabela abaixo seguem-se os valores da entropia(em bits) e a entropia máxima para cada fonte de informação

| **Ficheiro** | **Entropia em Bits** | **Entropia Máxima** | **Taxa de Compressão** |
|---|---|---|---|
| **soundMono.wav** | **0.0003090270151401479** | **log2 256 = 8** | **99.9961** |
| **MRIbin.bmp** | **0.11242695838839044** | **log2 256 = 8** | **98.5946** |
| **landscape.bmp** | **4.2274462392398387e-05** | **log2 256 = 8** | **99.9994** |
| **MRI.bmp** | **0.003900239747200491** | **log2 256 = 8** | **99.951** |
| **lyrics.txt** | **0.03939514445381911** | **log2 52** = 5.7 | **99.30885** |

Para os valores da entropia máxima , tem-se : considerando os elementos equiprovaveis, é dada por ***log2 N***, onde **N** é o número de símbolos do alfabeto e: **taxa de compressão =**

$\frac{Entropia\ maxima - Entropia\ em\ Bits}{Entropia\ maxima} \times 100$. O alfabeto considerado, para as imagens, foi (0 á 255), ou smelhor 256 símbolos. Para o ficheiro de texto, optou-se, apenas considerar todas as letras do alfabeto (26 maiúsculas e minúsculas) sem contar com carateres especiais.

Estudando cada resultado de entropia obtidos para cada uma das fontes de informação, pode-se constatar que a entropia da imagem ***landscape..bmp*** é a mais elevada, em relação a entropia das outras imagens Estes resultados vão de acordo com o que nos já havíamos feito , quer pela observação das imagens, quer pela observação dos histogramas correspondentes.

Questão: *Será possível comprimir cada uma das fontes de forma não destrutiva? Se sim, qual a compressão máxima que se consegue alcançar? Justifique.*

Sim, é possível. deve-se usar um algoritmo de compressão *lossless*, ou seja, sem perdas de informação. Para obter a compreensão sem perdas geralmente são usadas duas estratégias, a codificação de redundância mínima e o método do dicionário. A redundância mínima consiste em representar os símbolos que aparecem com mais frequência utilizando menos bits, um exemplo de algoritmo que utiliza essa tática é a codificação de huffman

**Exercício 4**

Eis a tabela de valores da entropia e variância pela codificação de Huffman.

| Nome do ficheiro: | Entropia em bits: | Variância : |
|---|---|---|
| soundMono.wav | 0.0747126436781609 | 16.198404120153334 |
| MRIbin.bmp | 0.8285195114685732 | 0.048726157009380086 |
| landscape.bmp | 7.654361824417141e-05 | 0.09580706504871782 |
| MRI.bmp | 0.007516631913414755 | 18.703009756988813 |
| lyrics.txt | 0.0006384880602732729 | 0.404754577212735 |

## A Codificação de Huffman

A codificação de Huffman é um método de computação que usa as probabilidades de ocorrência dos símbolos no conjunto de dados a ser compactado para determinar códigos de tamanho variável para cada símbolo.

O valor da **variância** do tamanho do código de Huffman exprime a relação entre a probabilidade de ocorrência dos símbolos com o tamanho do seu código: quanto maior o número de símbolos que ocorrem com maior probabilidade, menor o comprimento da cadeia que codifica o símbolo. Conforme vemos a variância é máxima para o ficheiro de som, uma vez que o ficheiro contém um número substancial de símbolos menos frequentes, codificados com um maior comprimento. No extremo oposto, a variância é

mínima para o ficheiro da imagem binária, pois este apenas apresenta dois símbolos apenas.

Questão : *Será possível reduzir-se a variância? Se sim, como pode ser feito em que circunstância será útil?*

Sim! é possível sim, "os códigos de Huffmana não são únicos"

**Redução da variância**

Se ao formar um novo conjunto de probabilidades decrescentes houver probabilidades iguais as que resultam de agrupamentos devem ser colocadas o mais alto possível. Desse modo reduz-se a variância V dos comprimentos ni das palavras de código. No entanto o comprimento médio N mantém-se.

**A variância mínima tem vantagens:**

- o ritmo de produção de bits é mais uniforme (o que é conveniente para o preenchimento de "buffers");
- há uma maior imunidade (resistência) a erros do canal, na descodificação.

**Exercício 5**

Esta tabela remete-se a entropia pelo agrupamento de símbolos dois a dois

| **Nome do ficheiro:** | **Entropia pelo Agrupamento de dois símbolos (em bits )** |
|---|---|
| **soundMono.wav** | **3.6521800559086772** |
| **MRIbin.bmp** | **0.4028703654483534** |
| **landscape.bmp** | **6.2040596930972045** |
| **MRI.bmp** | **5.193602970171974** |
| **lyrics.txt** | **3.3109957723669114** |

Como se pode constatar pela tabela , a entropia dos ficheiros aumenta , reparemos para o seguinte "soundMono.wav, MRI.bmp e lyrics.txt"aumenta quando se consideram agrupamentos binários de símbolos, o que nos levou a pensar nas cadeias de Markov.Houve um ganho ao nível da entropia.

**Conclusão**

Concluímos que , com este trabalho foi possível aprofundar criar mais exercício e buscar mais conhecimento no que diz respeito ao tema e todo o aprendizado que foi dado nas aulas teóricas, Assim como na proposta desde projeto o que concretiza um excelente desafio . Porém Não conseguimos terminar a ultima questão do projeto.