• U C •

UNIVERSIDADE DE COIMBRA

Teoria da Informação 2019/2020

Trabalho Prático nº 1

Entropia, Redundância e Informação Mútua

Trabalho realizado por:
Christopher Liu:2013150914
João Figueira:2013136460
Marco André Gomes:2010145568

Introdução

Para a realização deste trabalho, o conhecimento dos conceitos de alfabeto e fonte de informação era indispensável.

Uma fonte de informação é um gerador de símbolos pertencentes a um alfabeto onde estão designados um conjuntos de valores possíveis. Cada um desses símbolos tem associada uma probabilidade de ocorrência.

É possível reduzir a quantidade de bits necessários para armazenar imagens através de métodos e compressão, tal como a entropia e o código de Huffman, através da eliminação de bits redundantes de informação. Existem dois métodos:

- Não destrutiva – é possível reconstruir a imagem original antes de ter sido efetuada a compressão
- Destrutiva – no processo de compressão são perdidas características das imagens, o que permite obter graus de compressão mais elevados

Para este trabalho teremos como fontes de informação ficheiros de texto, áudio e imagem fornecidos antecipadamente pelo docente.

O objetivo deste trabalho será o de adquirir sensibilidade para as questões fundamentais da teoria da informação com destaque para a redundância, entropia e informação mútua. Tais questões serão abordadas ao longo dos diversos exercícios.

Exercício 1

Para este exercício foi pedido para que dada uma fonte de informação P com um alfabeto A={a1,...,an} fosse gerado o histograma de ocorrência dos seus símbolos.

Os resultados obtidos, bem como uma breve justificação, para os ficheiros landscape.bmp, MRI.bmp, MRIbin.bmp, soundMono.wav, lyrics.txt.

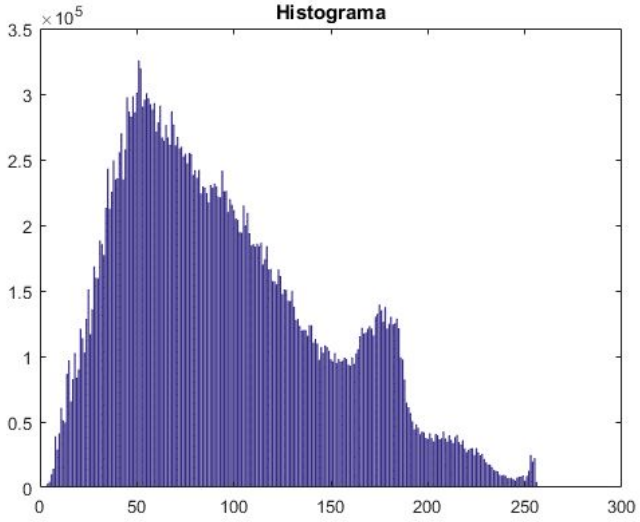

Figura 1: Histograma de landscape.bmp

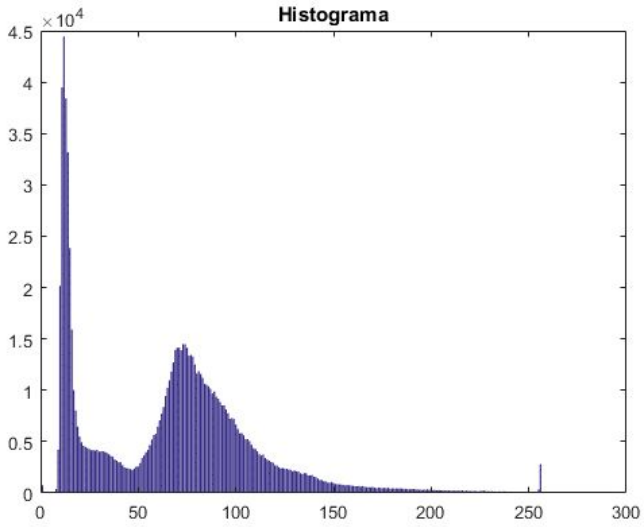

Figura 2:Histograma de MRI.bmp

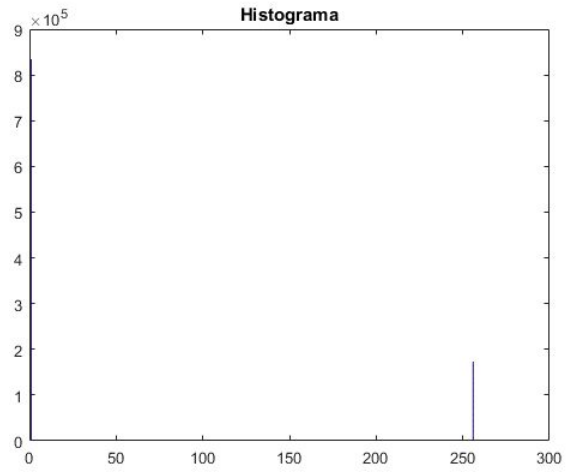

Figura 3:Histograma de MRIbin.bmp

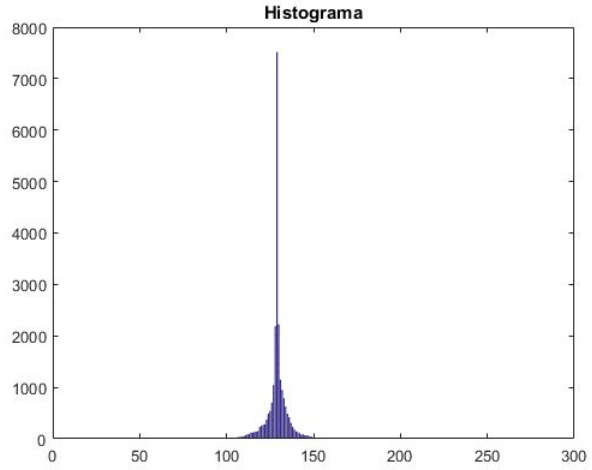

Figura 4:Histograma de soundMono.wav

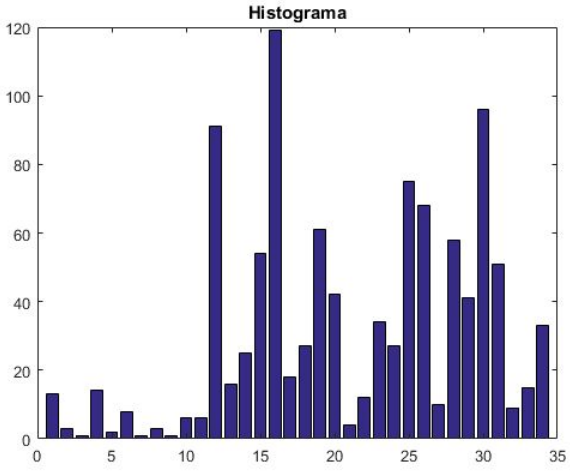

Figura 5:Histograma de lyrics.txt

Após analisar os diferentes gráficos obtidos a partir das fontes de informação fornecidas, podemos concluir que estes se encontram correctos, uma vez que demonstram uma correcta distribuição de valores ao longo dos histogramas:

• landscape.bmp: Imagem a incidir em diferentes tons de cinzento com predominância para os mais escuros, daí os valores apresentados no histograma estarem mais próximos do valor 0.

• MRI.bmp:Existe uma incidência muito maior na cor perto do preto e nos cinzentos daí vermos valores muito elevados perto do 0 e entre os 50 e 100.

• MRIbin.bmp: Imagem binária apenas com valores representantes das cores preto(0) e branco(255).É de notar também que a entropia desta imagem é extremamente baixa devido pois só é necessário um bit para representar duas cores.

• soundMono.wav: Gráfico densamente propagado com um pico que corresponde ao ponto intermédio do gráfico.

• lyrics.txt: Neste histograma é demonstrado que existem certos valores, correspondentes às maiúsculas que têm valores menores, e existe uma maior predominância nas minúsculas especialmente nas vogais que são as que mais aparecem no texto.

Exercício 2

Para este exercício, foi pedido para que fosse criada uma rotina que, dada uma fonte de informação P, com um determinado alfabeto, devolve-se o limite mínimo teórico para o número médio de bits por símbolo, ou seja, para calcular a entropia da fonte de informação.

Para as fontes de informação fornecidas os resultados para a entropia são os seguintes:

Fonte de informação	Entropia (bits / símbolo)
landscape.bmp	*7.606914*
MRI.bmp	*6.860542*
MRIbin.bmp	*0.661080*
soundMono.wav	*4.065729*
lyrics.txt	*4.410705*

Tabela 1: Valores de Entropia

Exercício 3

Para obter o rácio de compressão não destrutiva utilizamos a expressão: $\frac{\log_2(\#alfabeto)}{H(A)}$

Fonte de informação	Entropia (bits / símbolo)	Rácio de compressão
landscape.bmp	*7.606914*	*log2(256)/H(landscape.bmp) = 1.05167483161*
MRI.bmp	*6.860542*	*log2(256)/H(MRIbin.bmp) = 1.16608862682*
MRIbin.bmp	*0.661080*	*log2(2)/H(MRIbin.bmp) = 1.51267622678*
soundMono.wav	*4.065729*	*log2(257)/H(soundMono.wav) = 1.96905021195*
lyrics.txt	*4.410705*	*log2(36)/H(lyrics.txt) = 1.17213121291*

Tabela 2: Valores de rácio de compressão não destrutiva

Assim, pela tabela 2 podemos concluir que é possível comprimir cada uma das fontes de forma não destrutiva.

Exercício 4

A codificação de Huffman usa as probabilidades de ocorrência dos símbolos no conjunto de dados a ser comprimido para determinar o número de bits necessário para a representação de cada símbolo. Ao contrário da Entropia que nos dá um valor teórico, o valor dado pela codificação de Huffman será o valor mínimo físico alcançável.

Utilizando as rotinas de codificação de Huffman obtivemos um novo valor médio de bits / símbolo.

Fonte de informação	Entropia (bits / símbolo)	Huffman (bits / símbolo)
landscape.bmp	*7.606914*	*7.629301*
MRI.bmp	*6.860542*	*6.890996*
MRIbin.bmp	*0.661080*	*1*
soundMono.wav	*4.065729*	*4.110714*
lyrics.txt	*4.410705*	*4.443487*

Tabela 3: Comparação entre Entropia e codificação de Huffman

Ao analisar os dados, verificamos que a codificação de Huffman e a entropia tem valores muito próximos. A entropia codifica ficheiros pressupondo que cada símbolo ocupa o mesmo número de bits por outro lado, já nos códigos de Huffman, os símbolos que ocorrem menos vezes ocupam menos bits na codificação (otimização de código está dependente do comprimento dos ramos da árvore de Huffman).

É possível reduzir a variância. Para tal apenas temos que colocar os símbolos com ordem mais elevada primeiro, ou seja, símbolos com maior probabilidade de ocorrência primeiro.

Exercício 5

Após o agrupamento de símbolos dois a dois e o cálculo da nova entropia obtivemos os seguintes resultados:

Fonte de informação	Entropia (bits / símbolo)	Entropia Agrupada (bits / símbolo)
landscape.bmp	*7.606914*	*6.277266*
MRI.bmp	*6.860542*	*5.226929*
MRIbin.bmp	*0.661080*	*0.400694*
soundMono.wav	*4.065729*	*3.310798*
lyrics.txt	*4.410705*	*3.652180*

Tabela 4: Comparação entre Entropia e a Entropia agrupada

O que podemos observar neste exercício é que ao fazermos o agrupamento de símbolos em sequências de dois símbolos obtemos um valor mínimo teórico para o número médio de bits por símbolo inferior tal como apresenta a fórmula $H(X,Y) \leq H(X) + H(Y)$, pois os símbolos agrupados têm maior probabilidade de ocorrer. Logo, são precisos menos bits para codificar os símbolos agrupados.

Quanto aos histogramas notámos que estão mais dispersas as ocorrências dos símbolos, o que por si comprova os resultados obtidos no cálculo das entropias.

Exercício 6

Neste problema tivemos problemas na implementação pois não estava a dar os resultados de acordo com os fornecidos.