**Teoria da Informação**

Trabalho Prático – TP1

**Trabalho realizado por:**

Alexy Damien Tomé Julien Denis de Almeida nº2019192123

Sofia Botelho Vieira Alves nº2019227240

Sofia Santos Neves nº2019220082

**Professor Orientador:**

Rui Pedro Pinto de Carvalho e Paiva

# Exercícios 1 a 3

## Introdução:

Neste exercício foi-nos pedido para determinar a distribuição estatística através de um histograma e o limite mínimo teórico do número médio de bits por símbolo de cada fonte que nos foi apresentada. Analisando os resultados das funções criadas nas alíneas 1 e 2, conseguimos obter cada um destes resultados para cada fonte.

Relativamente à função correspondente ao **exercício 1**, dada uma fonte de informação P com um determinado alfabeto, foram extraídos os seus dados para um *array* denominado “data”, usando funções como **mpimg.imread(file\_name)** no caso do ficheiro se tratar de uma imagem, **wavfile.read(file\_name)** no caso de se tratar de um ficheiro áudio e ***open*** e ***read*** para os ficheiros texto.

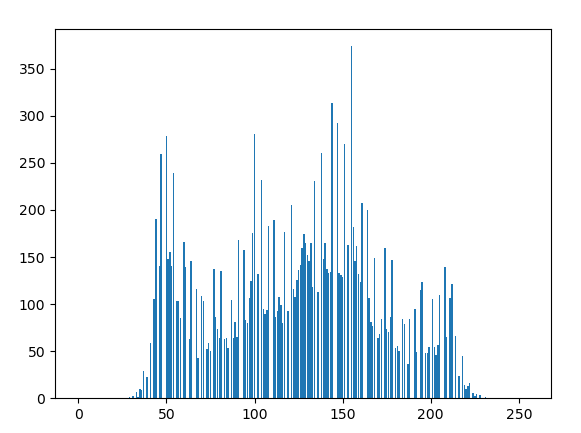
De seguida, foram calculadas as ocorrências de cada símbolo para cada uma das fontes. Para tal, criámos um *array* denominado“ocorrencias” que vaiincrementar o valor de ocorrências de um determinado símbolo do alfabeto à medida que percorre os valores armazenados no array “data”. Calculados estes valores, criámos um histograma que, a partir do alfabeto e das ocorrências, cria uma representação gráfica com a frequência da ocorrência de cada símbolo do alfabeto.

No **exercício 2**, para calcular a entropia de cada fonte criámos uma função “entropia” que recebe como parâmetro o *array* “ocorrências” e retorna a entropia fazendo o somatório do produto entre a probabilidade de ocorrer cada símbolo e o logaritmo de base 2 dessa mesma probabilidade.

## Resultados e análise:

### Imagem: lena.bmp

Uma imagem com pessoa, mulher, vestuário, toucado

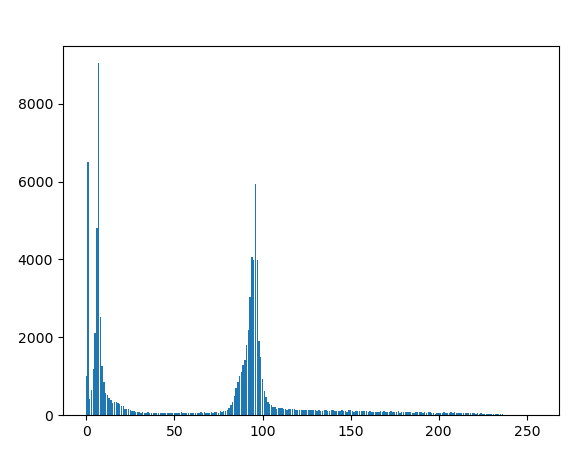
Descrição gerada automaticamente

**Entropia:** 6.915336

Nesta imagem, o elevado nível de entropia deve-se ao facto de haver uma maior quantidade símbolos com valores diferentes, sendo que cada símbolo poderá auferir valores entre 29 e 231, o que irá resultar numa imagem com uma tonalidade mais cinzenta e com menos contraste. Deste modo, a probabilidade de um pixel adjacente veicular a mesma informação será menor, ou seja, a incerteza será maior, resultando numa entropia elevada.

### Imagem: ct1.bmp

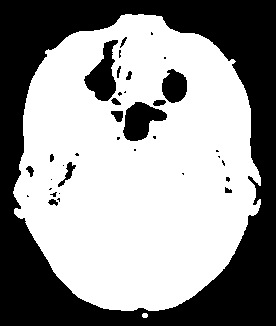
Uma imagem com bolo, sentado, mesa, aceso

Descrição gerada automaticamente

**Entropia:** 5.972234

Relativamente à imagem ct1.bmp, verificamos a presença de dois picos de valores. O primeiro pico corresponde ao fundo preto da imagem, enquanto que o segundo pico corresponde à parte mais cinzenta. Comparando os valores de ct1.bmp com os apresentados na imagem lena.bmp, verificamos que a entropia é menor pois existe uma menor dispersão dos valores, uma vez que estes se concentram mais perto de 0 e de 100. No entanto, a entropia ainda é relativamente alta pois a variância da informação veiculada entre pixéis adjacentes é elevada. Esta variância deve-se ao facto de haver um contraste entre tonalidades que não é suficientemente acentuado para diminuir ainda mais o valor da entropia.

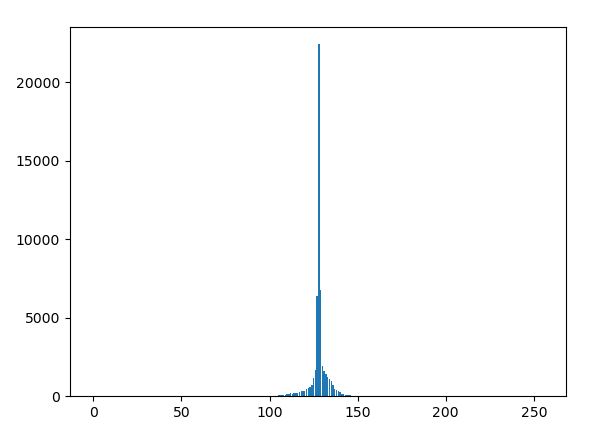
### Imagem: binaria.bmp



**Entropia:** 0.975527

Nesta fonte, o histograma possui apenas duas barras que correspondem aos valores 0 e 255. Isto deve-se ao facto de estarmos perante uma imagem binarizada onde cada pixel irá possuir apenas um de dois valores, 0 ou 255. Deste modo, como possuímos apenas dois valores distintos, a probabilidade de dois pixéis adjacentes terem a mesma informação irá ser bastante elevada, ou seja, haverá uma menor incerteza e, por conseguinte, uma menor entropia.

### Áudio: saxriff.wav



**Entropia:** 3.530989

No ficheiro áudio, nota-se um pico na zona dos valores entre 100 e 150 o que demonstra que a frequência do som nesse intervalo é elevada. Logo, a probabilidade de os pixéis terem a mesma informação é alta, o que se traduz numa incerteza baixa. Consequentemente, é possível ter uma entropia abaixo de metade do número médio de bits usados para codificar esta mesma informação.

### Texto: texto.txt

**Entropia:** 4.196889

São prevalentes as letras minúsculas em relação às letras maiúsculas neste excerto, com maior ocorrência das vogais. (\*)

Este apresenta uma redução para quase metade na entropia devido ao facto de haverem determinadas letras cuja probabilidade de frequência é maior, causando uma menor incerteza, ou seja, há maior certeza de que a próxima letra seja igual a uma das mais prevalentes do gráfico. Contudo, como mais de metade das letras apresentadas têm pouca probabilidade de ocorrer, não é possível baixar a entropia para menos de metade do número médio de bits usados.

(\*) Todas as letras que não fossem alfanuméricas foram excluídas para efeitos do enunciado.

## Compressão de dados:

Depois da análise dos resultados, foi pedido para avaliar se será́ possível comprimir cada uma das fontes de forma não destrutiva. Sabemos que para comprimir cada uma das fontes será necessário reduzir o número de bits usados. No entanto, existe um limite para essa compressão. Esse limite é a **entropia**, que será o limite mínimo de bits necessários para comprimir as fontes sem que **haja perda de informação (lossless).** Assim, analisando os resultados anteriores, conseguimos averiguar que a entropia de cada fonte é sempre inferior aos 8 bits usados por cada fonte para armazenar a informação veiculada, concluindo-se que podemos reduzir esses 8 bits para o nível da entropia, que será a compressão máxima que se pode alcançar de forma não destrutiva.

### Compressão máxima:

A compressão máxima da fonte pode ser calculada através da seguinte fórmula:

* **lena.bmp:** 13,5583%
* **ct1.bmp:** 25,347075%
* **binaria.bmp:** 87,8059125%
* **saxriff.wav:** 55,8626375%
* **texto.txt:** 47,5388875%