Sistemas Multimédia

Texto

Professor: Paulo Gomes

Email: paulo.gomes@uportu.pt

Introdução

- O texto é o meio mais utilizado pelos seres humanos para a transmissão de informação.
- Na Interacção Homen-Máquina, um conteúdo textual pode assumir umas das seguintes:
  - **Texto não-formatado** (*Plain Text*) – O número de caracteres é limitado ao *Character Set*; A fonte e a dimensão dos caracteres são fixas e apenas existe uma.

Introdução (Cont.)

- **Texto formatado** (*Rich Text*) – Existem vários fontes e dimensões de caracteres. A representação de texto formatado recorre a formatos para documentos.

- **Hipertexto** – O hipertexto define-se como texto não-linear. Este formato de representação possui um mecanismo de navegação (*links*) que permite navegar entre documentos de texto (nós).

Representação de texto

- O texto possui uma natureza dupla:

  - **Um conteúdo léxico** – é a parte do texto que representa o seu significado ou a sua semântica, como por exemplo o carácter abstracto "J";
  - **Uma aparência** – atributo superficial que afecta a aparência e a facilidade com que o texto é lido pelo utilizador. A aparência não altera o seu significado. Como por exemplo "J", "j", "**J**", "*J*", etc.

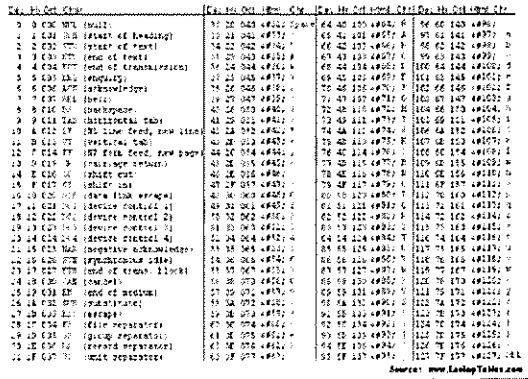
Representação do Conteúdo Textual

- Os alfabetos digitais utilizados incluem caracteres abstractos (maiúsculas e minúsculas), números, sinais de pontuação, símbolos e símbolos matemáticos.

- Para que se possa representar texto no formato digital é necessário definir o mapeamento (por intermédio de um *Character Set ou conjunto de códigos*) entre os caracteres abstractos de um dado alfabeto e um código através do qual o carácter é representado na forma digital.

Representação do Conteúdo Textual (Cont.)

- O primeiro *Character Set* normalizado (1970) foi o ASCII (*American Standard Code for Information Interchange*) com 7 bits.

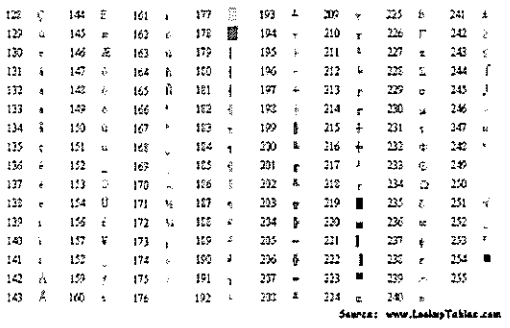


Representação do Conteúdo Textual (Cont.)

- O ASCII foi desenvolvido nos EUA para servir *Character Set* do alfabeto inglês.
- Quando o ASCII foi adoptado como norma internacional pela ISO (*International Organization for Standardization*), designação ISO 646 (ano 1972), foi ampliado com um conjunto de variantes nacionais de modo a suportar um conjunto de caracteres acentuados e símbolos associados a outros idiomas.
- Esta ampliação foi conseguida com a criação de um *Character Set* com 8 bits. (Conseguida ?)

Representação do Conteúdo Textual (Cont.)

- Dados que 8 bits são insuficientes para representar todos os caracteres de todos os idiomas tornou-se necessário criar variantes regionais.
- Exemplo ISO 8859-1 ou ISO Latin1 contêm códigos para os idiomas da Europa Ocidental



Representação do Conteúdo Textual (Cont.)

- A quantidade de caracteres diferentes que permite representar o ASCII de 8 bits não é suficiente para permitir trabalhar com múltiplos idiomas simultaneamente.
- A ISO criou em 1991 a norma 10646, com 32 bits para representar 4.294.967.296 caracteres distintos.

Representação do Conteúdo Textual (Cont.)

- Em simultâneo, as empresas *Adobe*, *Apple*, *Microsoft*, HP, IBM, *Oracle*, SAP, SUN e a *Unisys* criaram um *Character Set* de 16 bits com a capacidade de representar 65.536 caracteres distintos.
- Este *Character Set* foi apelidado de UNICODE (usado nas linguagens *markup* HTML, XML e Java).

Representação da aparência (Cont.)

- A apresentação de texto requer que cada código correspondente a um carácter abstracto seja mapeado na representação gráfica desse carácter.
- A representação visual do carácter abstracto é designada por glifo, podendo um carácter abstracto ser representado por vários glifos.
- A representação gráfica de um carácter pode introduzir a alteração da forma e dimensão do carácter, mas nunca a sua identidade.

Representação da aparência (Cont.)

- A utilização de glifos obedece a uma estrutura organizada denominada de fontes.
- Os vários glifos pertencentes a uma fonte partilham um conjunto de características gráficas (forma e dimensão) harmoniosas.
- *"É possível encarar a fonte como o mecanismo que faz o mapeamento de caracteres abstractos em glifos, do mesmo modo que encaramos um conjunto de caracteres como o mecanismo que faz o mapeamento de caracteres abstractos em códigos (para armazenamento de texto digital." [Ribeiro04]*

Representação da aparência (Cont.)

- As fontes podem ser classificadas de acordo com as seguintes dimensões:
  - **Fontes mono-espaçadas e fontes proporcionais** – Nas fontes mono-espaçadas cada carácter ocupa o mesmo espaço horizontal, exemplo: Courier e Courier New. Nas fontes proporcionais o espaço ocupado por cada carácter depende da sua largura da sua forma, como por exemplo: Times New Roman, Helvetica, Arial.

Representação da aparência (Cont.)

- **Fontes com *serif* e fontes sem *serif*** – Os serif são traços minúsculos que existem nas extremidades dos caracteres de fontes com *serif* (fontes romanas), como por exemplo: Times New Roman, New York. São exemplos de fontes sem *serif*: Helvetica, Arial.

Representação da aparência (Cont.)

- **Fontes com forma vertical e fontes itálicas** – Nas fontes com forma vertical os caracteres possuem formas com linhas verticais. Nas fontes com forma itálica os caracteres possuem formas com linhas verticais com inclinação para a direita. A maioria das fontes itálicas constitui variações ou acompanha as fontes com forma vertical. Existem fontes com forma itálica que são concebidas para possuírem apenas esta forma, exemplo: *Monotype Corsiva.*

Representação da aparência (Cont.)

- **Fontes pesadas e fontes leves** – Classifica as fontes de acordo com a espessura do traço usado para representar um carácter. Texto com traços grossos possui uma aparência mais sólida, escura e pesada (designadas de negrito ou *bold*). A fontes negrito á semelhança das fontes itálicas, também constituem versões de fontes leves.

Representação da aparência (Cont.)

- **Recomendações:**
  - As fontes proporcionais produzem texto mais legível e fácil de ler do que texto de fontes mono-espaçadas.
  - As fontes sem *serif* são mais adequadas para títulos, títulos de janelas e itens em menus.
  - As fontes com forma itálicas são usadas quando se pretende texto com uma aparência humanizada.

Representação da aparência (Cont.)

- **Recomendações:**
  - As fontes negrito são usadas para assinalar títulos, cabeçalhos ou para destacar conceitos chave. A leitura de texto com fonte negrito é cansativa.
  - Fontes para texto contínuo (exige períodos longos de leitura) devem ser discretas de modo a permitirem uma leitura fácil e menos cansativa possível.

Representação da aparência (Cont.)

- A descrição das características das fontes assenta em medições cuja unidade mais comum e o ponto (pt). Um ponto corresponde aproximadamente 0,3528 mm e é utilizado para medir as dimensões dos caracteres, altura entre o topo do carácter mais alto e o fundo do carácter mais baixo.
- Existem dois tipos de tecnologias para armazenar as imagens dos glifos de uma fonte. As imagens podem ser armazenadas no ficheiro da fonte correspondente sob a forma de gráficos vectoriais (*outline*) ou imagens *bitmap* (*bitmapped*).

Disposição do conteúdo textual

- *"O segundo aspecto associado à aparência do texto, ou formatação, está relacionado com o modo como os caracteres se combinam em palavras, frases, linhas, parágrafos e outras unidades de divisão de documentos de texto, tais como secções e capítulos (...), modo como o conteúdo do texto se dispõe no ecrã ou página a imprimir – o seu layout."* [Ribeiro04]

Disposição do conteúdo textual

- Os documentos formatados possuem uma estrutura interna. Os conjuntos de regras que descrevem tais estruturas são designados por formatos para documentos. Existem dois tipos de formatos para documentos de texto:
  - **Formato de descrição de estrutura** – contêm marcas de controlo adicionadas ao corpo do texto.

Disposição do conteúdo textual (Cont.)

- **Formato de descrição de páginas** – baseiam-se numa linguagem de programação para descrever as páginas de um documento em termos de comandos. Estes comandos podem ser gerados pelo processador de texto e são interpretados por um processador localizado numa impressora ou por uma aplicação. Exemplo: PDF e PS (*Postscript*).

Compressão de Texto

- Os métodos de compressão utilizados para texto apenas removem a redundância existente num dado documento de texto.
- Os métodos de compressão mais comuns são:
  - Codificação Huffman
  - Codificação LZW (*Lempel-Ziv and Welsh*).

Compressão de Texto – *Huffman*

- Codificação por entropia; comprime sem perdas cadeia de símbolos independentemente do seu significado.
- Na codificação Huffman (codificação estatística) são atribuído menos bits a símbolos que registam maior frequência e mais bits a símbolos de menor frequência.

Compressão de Texto – *Huffman* (Cont.)

- **Exemplo:**
  - Suponha a existência de um ficheiro de texto com 1000 caracteres, estes caracteres são e, ç, x e z.
  - A probabilidade de ocorrência de e, ç, x e z são 0.8, 0.16, 0.02, 0.02 respectivamente.
  - Quanto ocupa o ficheiro de texto?

Compressão de Texto – *Huffman* (Cont.)

- Usando a codificação *Huffman*, os diferentes caracteres usam distintas dimensões de bits.
- Assim, é usado 1 bit para representar e, 2 bits para ç, 3 bits para x e z. Neste caso de exemplo o ficheiro de texto ocuparia após a codificação *Huffman* um total:
  - 1000*(1*0.8+2*0.16+3*0.02+3*0.02)=1240 bits.

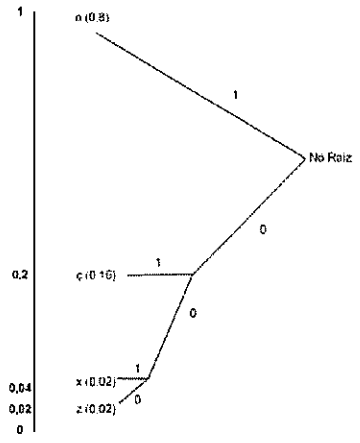
Compressão de Texto – *Huffman* (Cont.)

- As regras para atribuir bits (códigos) aos símbolos são denominadas de *codebook*.
- Os *codebooks* são normalmente expressos em tabelas: w(e)=1, w(ç)=01, w(x)=001 e w(z)=000.

Compressão de Texto – *Huffman* (Cont.)

- ***Procedimentos:***
  1. Coloque todos os símbolos ao longo da linha de probabilidade acumulativas.
  2. Selecciona-se os dois símbolos de menor probabilidade e cria-se um nó pai para formar dois ramos na árvore.
  3. O novo nó pai formado possui a soma das frequências dos símbolos ramos.
  4. Repita os passos 2 e 3 até que todos os símbolos sejam inseridos na árvore. O último nó é denominado de nó raiz.
  5. Partindo do nó raiz, atribua o bit 1 ao ramo de maior probabilidade e o bit 0 ao ramo de menor probabilidade de cada nó.
  6. O código para cada símbolo resulta da adição dos códigos ao longo dos ramos da árvore.

Compressão de Texto – *Huffman* (Cont.)



Compressão de Texto – *Huffman* (Cont.)

- No lado do descodificador, este apenas realiza uma simples verificação na tabela.
- Portando, o descodificador necessita da tabela *Huffman* usada no codificador.
- Esta tabela faz parte do fluxo de dados ou previamente conhecida pelo descodificador (tabelas padrão para vídeo e áudio).

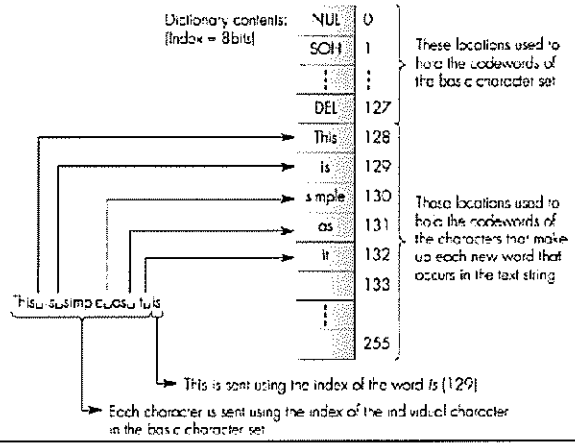
Compressão de Texto - *Lempel-Ziv-Welch* (LZW)

- A codificação LZW tem como base a construção de um dicionário de palavras a partir do fluxo de texto de entrada.
- Quando uma nova palavra é encontrada, o codificador LZW adiciona-a ao dicionário e a palavra é substituída por um *token* que identifica a posição desta no dicionário.
- Se a palavra já se encontra registada no dicionário, ela é substituída pelo *token* de posição no dicionário.

Compressão de Texto – LZW (Cont.)

- Inicialmente o dicionário (codificador e descodificador) contém apenas os caracteres básicos (ASCII por exemplo).
- O dicionário é criado de modo que espera-se ter um número máximo de palavras, sendo que as primeiras são os caracteres básicos e as demais palavras adicionadas dinamicamente ao dicionário.
- Somente palavras contendo letras são armazenadas no dicionário.

Compressão de Texto – LZW (Cont.)



Compressão de Texto – LZW (Cont.)

- Dicionários podem iniciar com um tamanho reduzido e ir aumentando o número de verbetes dinamicamente.