UTF8

A unidade básica de armazenamento usada no UTF-8 é um byte (8 bits) mas o código de um caracter pode ocupar de um a quatro bytes.

Os caracteres que façam parte do conjunto ASCII são armazenados em apenas um byte.

Como o ASCII usa códigos de 7 bits, o bit mais significativo do byte será sempre zero.

Nos caracteres que não fazem parte do conjunto ASCII, o bit mais significativo de todos os bytes empregados é sempre um.

No primeiro byte de de cada código os dois bits mais significativos tem sempre o valor 11.

Nos demais bytes do código, o valor dos dois bits mais siginificativos é sempre 10 e os seis demais bits informam o valor de um grupo de seis bits do código UNICODE do caracter.

O valor do primeiro byte dos códigos que ocupam mais de um byte também informa o número total de bytes empregados:

- Em códigos que ocupam dois bytes, os três bits mais significativos do primeiro byte tem o valor 110 e os cinco demais bits informam os cinco bits mais significativos do código UNICODE do caracter;
- Em códigos que ocupam três bytes, os quatro bits mais significativos do primeiro byte tem o valor 1110 e os quatro demais bits informam os quatro bits mais significativos do código UNICODE do caracter;
- Em códigos que ocupam quatro bytes, os cinco bits mais significativos do primeiro byte tem o valor 11110 e os três demais bits informam os três bits mais significativos do código UNICODE do caracter;

Os códigos que ocupam dois bytes são formados pelos cinco bits informados no primeiro byte e por mais um grupo de seis bits informados no segundo byte, totalizando onze bits, o que permite referenciar os códigos UNICODE com valores menores que 2048. Esses códigos cobrem quase todos os caracteres usados nas línguas que usam o alfabeto latino e também os alfabetos grego, cirílico, copta, armênio, hebraico, árabe, siriaico, Thaana e N´Ko.

Os códigos que ocupam três bytes são formados pelos quatro bits informados no primeiro byte e por mais dois grupos de seis bits informados no segundo e terceiro bytes, totalizando dezesseis bits, o que permite referenciar os códigos UNICODE com valores menores 65536. Esses códigos abrangem toda a definição original do UNICODE (de 16 bits), o que permite codificar quase todos os carateres atualmente em uso, incluindo a maior parte dos caracteres usados no chinês, japonês e coreano.

Os códigos que ocupam quatro bytes são formados pelos três bits informados no primeiro byte e por mais três grupos de seis bits informados no segundo, terceiro e quarto bytes, totalizando vinte e um bits, o que permite referenciar os códigos UNICODE que tem valores menores 2097152. Esses códigos abragem todo o UNICODE atualmente definido, incluindo os caracteres incomuns do chinês, japonês e coreano, diversas línguas históricas, símbolos matemáticos, emoji etc.

A tabela abaixo resume a codificação usada no UTF-8:

Tamanho	Byte 1	Byte 2	Byte 3	Byte 4	UNICODE	Último UNICODE	observações
1	0vvv vvvv				vvv vvvv	0x7F=127	ASCII (94 caracteres apresentáveis)
2	110v vvvv	10xx xxxx			vvv vvxx xxxx	0x7FF= 2047	Línguas alfabéticas
3	1110 vvvv	10xx xxxx	10yy yyyy		vvvv xxxx xxyy yyyy	0xFFFF= 65535	Quase todos os caracteres em uso
4	1111 0vvv	10xx xxxx	10yy yyyy	10zz zzzz	v vvxx xxxx yyyy yyzz zzzz	0x1F FFFF= 2097151	Todos os caracteres