El Bloc de Notas puede codificar en ANSI, Unicode, Unicode BE y UTF8.

Wordpad guarda en los formatos: RTF, XML, ODF, Documento de texto, Documento de texto Unicode, Documento de texto - formato de MSDOS

El texto original está compuesto de 102 caracteres según Microsoft Office Word. Su tamaño es:

- Texto ANSI: 106 bytes
- Texto Unicode: 214 bytes
- Texto UnicodeBE: 214 bytes
- Texto DOS: 106 bytes
- Texto UTF8: 119 bytes

Por tanto en ANSI y DOS se obtiene el menor tamaño, en oposición a ambos Unicode que ocupan los que más.

- ASCII, es la más básica, de 128 caracteres
- ANSI y DOS tiene una longitud de palabra de 1 byte y 256 caracteres
- UTF8 tiene una longitud de palabra variable de 1, 2, 3 o 4 bytes
- UTF16 tiene una longitud de palabra variable de 2 o 4 bytes

Además UTF8 y UTF16 incluyen una marca inicial que les distingue (BOM: Byte Order Mark).

Los caracteres básicos presentes en ASCII presentan la misma codificación en las distintas codificaciones, al ser éstas expansiones de la primera. Esto se cumple a excepción de Unicode y UnicodeBE que se al usar una longitud de palabra mayor se usan 2 bytes en vez de 1 por defecto.

Las tildes y demás símbolos particulares no comparten codificación. Entre Unicode y UTF8 no he encontrado similitud, al mostrarse el archivo UTF8 con símbolos extraños.

En cuanto a las diferencias entre UTF8 y Unicode (UTF 16) es grande, de hecho un fichero UTF8 abierto en Unicode se vuelve completamente ilegible, al leerse de 2 en 2 bytes en vez de 1 en 1.

Las versiones de Unicode llevan un BOM al inicio para expresar cómo ordenan los bytes y así especificar si se usa LE (Little Endian) o BE (Big Endian). Ambas codificaciones son similares, pero intercambian el orden de los bytes al codificar.

El navegador depende de la página cambia de codificación ISO-8859, UTF8... Al cambiar entre codificaciones se observan cambios en símbolos particulares como letras con acentos, la letra ñ... Al cambiar la página a UTF16, esta cambia completamente, llegando incluso a dejar de ser visibles las imágenes...

DOS-Europa Occidental 850 dispone de 256 caracteres, el límite de 1 byte.

Windows - Occidental 1252 dispone de 256 caracteres, el límite de 1 byte.

Unicode en su versión 6.0 (última disponible a fecha de hoy) dispone de 109.449 caracteres, aunque en 4 bytes se pueden almacenar teóricamente hasta 4.294.967.296

Al copiar caracteres de DOS al archivo “textoDOS.txt” y abrirlo con la consola de comando de MSDOS los caracteres son visibles.

Al entrar en:

http://www.aq.upm.es/Departamentos/Fisica/agmartin/webpublico/latex/FA

Se puede ver el FAQ en diversas codificaciones. Estos solo son vistos adecuadamente solo cuando se abren en la correcta codificación. Si por el contrario se usa otra, se puede ver como ciertos símbolos particulares no se muestran como deberían.

Si el navegador o SO que abre el archivo no soporta la codificación no se verá correctamente. En mi caso he podido ver todas las codificaciones sin problema en Firefox 7 bajo Windows 7, Mac OS X 10.7 y Ubuntu 11.04, a excepción de CP437, que no he encontrado su tabla de codificación, ni en Firefox, ni en otros navegadores (Opera y Safari).