Nombre: .....

Fecha: /01/2012 Grupo: 1 □

PRÁCTICA 1
INFORMÁTICA BÁSICA, CODIFICACIONES

Entre los objetivos de la práctica se encuentra el trabajar con algunas de las distintas normas de codificación disponibles en nuestro sistema. Veremos el resultado de guardar ciertas informaciones en nuestro sistema con respecto a alguna codificación, qué sucede cuando intentamos recuperarlas con codificaciones distintas, formas de cambiar las páginas de códigos o las posibilidades y consecuencias de cambiar la configuración regional del sistema.

1. En esta práctica vamos a utilizar información de al menos las siguientes direcciones de Internet. Abre cada una de ellas y observa el tipo de información que contienen.

¿Qué es ASCII? Tabla ASCII y manejo de binario, hexadecimal y decimal:

Tablas de códigos:
http://www.microsoft.com/globaldev/reference/cphome.mspx
¿Qué es Unicode?
http://es.wikipedia.org/wiki/Unicode
http://msdn.microsoft.com/es-es/goglobal/bb688113.aspx
¿Qué es UTF-8?
http://es.wikipedia.org/wiki/UTF-8
Consorcio Unicode:
http://www.unicode.org/
Lista de caracteres Unicode (de los 65536 primeros caracteres):
http://en.wikibooks.org/wiki/Unicode/Character_reference/0000-0FFF
Calculadora de representación de los caracteres Unicode en codificación UTF-8:
http://www.ltg.ed.ac.uk/~richard/utf-8.cgi

2. Crea una carpeta con el nombre "Practica" en "Mis Documentos", donde guardaremos los ficheros de esta práctica.

3. Ejecuta el bloc de notas ("notepad") y escribe en él el siguiente texto. Escríbelo también en tu informe de prácticas (Practica01.txt):

```
Esta fila es sólo de caracteres ASCII?
¿Esta línea tiene algún carácter no ASCII?
Ñandú, garçon, cigüeña
```

4. Guarda el fichero anterior ("Guardar como") con cuatro codificaciones distintas, ANSI, Unicode (UTF-16), Unicode Big Endian (UTF-16BE) y UTF-8 con los nombres "textoAnsi", "textoUTF16", "textoUTF16BE" y "textoUTF8", respectivamente (observa los "tipos de formato" en que sabe codificar información "notepad" y anótalos en tu informe de prácticas).

5. Ejecuta el programa "wordpad". Escribe el mismo texto anterior. Entra en el menú "Archivo: Guardar como" y observa los "tipos de formato" (no los tipos de fichero) en que sabe escribir. Anótalos en tu informe de prácticas. Guarda el archivo como "Documento de texto formato MSDOS" con el nombre "textoDos".

6. Entra en la carpeta "Practica" y, pulsando sobre cada uno de los archivos con el botón derecho y seleccionando "Propiedades", observa cuánto ocupan. Anota en tu informe de prácticas el nombre y el tamaño de cada versión del fichero. Cuenta el número de caracteres que hay en el fichero original (no olvides los saltos de línea) ¿Qué ficheros ocupan menos? ¿Cuál es la diferencia entre las distintas codificaciones? ¿Qué ficheros ocupan más? ¿Cuál de las codificaciones usadas es de longitud variable? (Ten en cuenta que las codificaciones de Unicode puede incluir al principio del fichero el BOM, http://es.wikipedia.org/wiki/Unicode#Esquemas_de_codificaci.C3.B3n).

Llamamos "longitud de palabra" o "longitud de carácter" a la cantidad de memoria que ocupa cada carácter propio de una codificación cuando el mismo es almacenado en memoria. La longitud de carácter en las distintas codificaciones suele ser siempre un múltiplo de 8 bits, por razones históricas, y se puede distinguir en general entre codificaciones con longitud de palabra fija (todos los caracteres ocuparán el mismo espacio en memoria) o variable (cada carácter puede ocupar una cantidad de memoria distinta).

7. Comprueba si "Pspad" está instalado en tu ordenador. Si no lo está, descárgalo e instálalo de http://www.pspad.com/es/download.php. Ejecuta Pspad. Vamos a "Archivo: Abrir en hexadecimal" y abrimos cada uno de los 5 ficheros que hemos guardado anteriormente. Trata de responder a las siguientes preguntas (ten en cuenta que dos símbolos hexadecimales pueden representar la misma información que un byte). ¿Qué longitud de palabra en bytes tiene cada una de las codificaciones MSDOS, ANSI, UTF-8 y UTF16 en sus dos versiones (relaciona tu respuesta con el número de caracteres en el fichero)? ¿Qué símbolos del texto tienen la misma representación en las cinco codificaciones (trata de hacer dos grandes bloques, los caracteres ASCII, los 128 primeros de las tablas, y el resto)? ¿Qué símbolos no? Sobre los símbolos que no comparten representación, ¿tienen la misma representación en UTF-16 y en UTF-8? ¿Qué sucedería si guardas un fichero con la codificación UTF-8 y lo abres (o lo visualizas) con la codificación UTF-16 (o Unicode)? ¿Se vería correctamente?

8. Nos quedamos solo con los archivos en sus codificaciones MSDOS, ANSI y UTF-8. ¿Qué símbolos comparten representación (es decir, mismo número hexadecimal)? ¿Cuáles no?

9. Observa ahora las versiones en hexadecimal de los ficheros "textoUTF16" y "textoUTF16BE". Ambos están almacenados en la misma codificación (Unicode, o UTF-16). ¿Por qué sus versiones guardadas son diferentes? ¿Serán compatibles ambas representaciones? Sus diferencias están relacionadas con la idea de la "endianness" de las familias de procesadores. Puedes leer en http://es.wikipedia.org/wiki/UTF-

16#Esquemas_de_codificaci.C3.B3n_y_BOM y http://es.wikipedia.org/wiki/Endianness una explicación más detallada.

10. Utiliza Pspad para "Abrir en Hexadecimal" los ficheros "textoUTF16", "textoUTF16BE" y "textoUTF8" y comprueba la existencia en los mismos de BOM. Comprueba en http://es.wikipedia.org/wiki/Marca_de_orden_de_bytes_(BOM) la utilidad del BOM, y apunta en tu informe de prácticas la codificación hexadecimal del mismo en cada una de las codificaciones planteadas.

11. Ejecuta un navegador (Explorer, Firefox…). Abre una página al azar. ¿En qué codificación se han guardado los caracteres de dicha página? Intenta observar el código fuente de la misma y comprobarlo. Ahora abre el menú "Página: Codificación" en Explorer o "Firefox: Desarrollador Web: Codificación de caracteres" en Firefox. Observa todas las codificaciones que contiene. ¿Qué pasa si cambiamos la codificación con respecto a la cual estamos leyendo la página? ¿Qué caracteres cambian? ¿Cuáles se siguen mostrando bien? Entre las distintas versiones, prueba también a observar la página web con respecto a la codificación UTF-16 y explica en tu informe de prácticas el resultado obtenido.

12. Trabajaremos ahora con la introducción de caracteres que no están en el teclado. Esto lo podemos hacer a través del "Mapa de caracteres", que encontrarás en "Accesorios – Herramientas del Sistema". Observa las distintas tablas de codificaciones de que dispone tu sistema, y sus caracteres correspondientes. ¿De cuántos caracteres dispone "DOS - Europa Occidental" (tabla 850)? ¿Y "Windows - Occidental" (tabla 1252)? (Ten en cuenta que sólo se muestran los caracteres correspondientes al tipo de letra que hayas elegido, lo que evita que algunos caracteres de control propios de las tablas se muestren) ¿Y Unicode? Relaciona la respuesta con las longitudes de palabra correspondiente (recuerda que un "byte" permite almacenar $2^8 = 256$ valores distintos).

Selecciona la tabla de códigos "DOS – Europa Occidental" y alguno de los caracteres especiales de la misma (entre el 129 y el 256). Cópialo en tu fichero "textoDOS.txt" y en "Practica01.txt". Ejecuta una consola de comandos de MSDOS (puedes escribir "cmd" en la ventana "Buscar programas y archivos"). Muestra el fichero "textoDOS.txt" por medio del comando "type". ¿Se muestran correctamente los caracteres? Existe otra forma de insertar caracteres especiales (en realidad, cualquier carácter). La puedes encontrar en el Anexo de la práctica. Compruébala en la propia consola de MSDOS y en el fichero "textoDOS.txt" (con "wordpad") incluyendo algunos caracteres que no estén en el teclado.

13. Unicode, dentro de sus distintos rangos de caracteres, admite lo que se conoce como caracteres privados. Los mismos son caracteres que distintos desarrolladores de software u organizaciones pueden generar libremente para distribuir entre sus usuarios. Ejecuta desde la barra de programas de tu sistema operativo la herramienta "Editor de caracteres privados". Selecciona posiciones dentro de la misma y crea tus propios caracteres. En el menú "Archivo" puedes vincularlos con los tipos de fuentes que elijas (por

defecto, vamos a vincular los caracteres creados para que estén disponibles con todas las fuentes).

Ahora, en tu "Mapa de caracteres", selecciona como juego de caracteres "Unicode" y como fuente "Todas las fuentes (Caracteres privados)". Deberían aparecer los caracteres que has diseñado. Selecciona alguno de ellos y cópialo en el fichero "textoUnicode.txt".

Puedes encontrar en http://en.wikipedia.org/wiki/Plane_(Unicode)#Private_Use_Area_planes una descripción más detallada de cuáles son los planos que reserva Unicode para caracteres privados.

14. Desde la propia consola de MSDOS podemos modificar la página de caracteres que se está usando en la consola. Para ello debes utilizar el comando "chcp". Alguna de las tablas estándar que suelen estar instaladas son 850 (tabla por defecto en la consola para nuestra región, "DOS – Europa Occidental"), 437 ("DOS – USA"), 1252. Comprueba su disponibilidad y la capacidad para introducir en la consola caracteres especiales por medio de la combinación "Alt + codificación decimal" (puedes encontrar una lista más detallada de tablas y los caracteres que ocupan cada posición en http://msdn.microsoft.com/en-us/library/cc195051.aspx y más detalles sobre la introducción de caracteres por medio de su codificación decimal en el Anexo al final de la práctica).

15. Observa que la utilización de distintos Sistemas Operativos (y de su configuración local) puede influir en la tabla de caracteres que utilicen los mismos, y por tanto en cómo se almacene y muestre la información. Si está Linux instalado en tu ordenador, trata de comprobar qué tabla de código usa por defecto (en caso contrario, puedes encontrar la información en http://en.wikipedia.org/wiki/Unicode).

16. Ve a la dirección http://www.aq.upm.es/Departamentos/Fisica/agmartin/webpublico/latex/FAQ-CervanTeX-main.html y entra en cada uno de los enlaces de abajo. ¿Qué sucede al leerlos con el navegador? ¿Por qué? ¿Qué sucederá en otro navegador o sobre otro S.O.? Trata de buscar en las codificaciones de tu navegador la correcta, y comprueba si el fichero se observa adecuadamente.

17. En el ámbito de las codificaciones de caracteres, muchas veces se abusa del lenguaje sin dejar claro a qué codificación nos estamos refiriendo. Dentro del Pspad, en la pestaña Herramientas, pulsa la opción "Tabla ASCII". Observa la tabla obtenida. ¿A qué codificación, de las que hemos visto en la práctica, corresponde la misma? Trata de averiguarlo observando las tablas que tú consideres más probables. Anótalo en tu informe. Aquí, el programa Pspad está realizando un abuso de lenguaje denominando codificación ASCII a una de las numerosas codificaciones ASCII extendidas (http://es.wikipedia.org/wiki/ASCII_extendido), pero sin mencionar cuál.

18. Otras dos tareas habituales cuando trabajamos con distintas codificaciones es la de averiguar en qué codificación está trabajando el

programa que usamos. Genera un fichero con el bloc de notas de nombre "notepad.txt" que contenga caracteres "no ASCII" como los que usamos en el texto del comienzo de la práctica (o "ä", "Ä", "ñ", "Ñ"…). Guárdalo con la opción por defecto de "notepad" sin especificar su codificación. Ábrelo con PsPad en hexadecimal. Trata de inferir en qué codificación ha sido guardado el mismo. Apúntalo en tu informe.

19. De la misma manera, trata de averiguar en qué codificaciones trabaja por defecto WordPad y Pspad. Anótalo en tu informe.

20. Como consecuencia de lo anterior, nos podremos encontrar con situaciones en que tengamos un fichero de texto guardado en una codificación determinada y tengamos que modificar la misma. Vamos a instalar ahora otro editor de textos que facilita esta tarea. Entra en http://notepad-plus-plus.org/ y descarga Notepad++. Instálalo en tu equipo.

21. Ejecuta Notepad++. En la opción "Formato" del mismo observa las opciones que se muestra. Desde el menú "Archivo" abre alguno de los ficheros que has generado al comienzo de la práctica. Desde la opción "Formato" de Notepad++ observa la codificación del fichero que has abierto, y las opciones de conversión que ofrece el mismo. Cambia la codificación de caracteres del fichero entre alguna de las opciones que se ofrecen (por medio de la opción "Convertir en …"). Observa que el mismo aparece como "no guardado". Tras guardarlo, comprueba las propiedades del fichero (en particular su tamaño) y compáralas con las que tenía previamente. También puedes observar su codificación hexadecimal (por ejemplo, en PsPad) para saber si los cambios han surgido efecto.

Anexo – Ayuda

Introducción de caracteres por medio de su posición en la tabla de códigos:
1. Mantener presionada la tecla ALT y, después, presionar las teclas del teclado numérico que representan el valor del código decimal del carácter que se desea introducir.
2. Una vez termine de escribir, suelte la tecla ALT.
3. "notepad" o la consola de MSDOS generan el carácter especificado.

Notas

Si el primer dígito que se escribe es 0, el valor se reconoce como un punto de código, o valor de carácter, del idioma actual (el "idioma actual" lo puedes seleccionar desde la barra de herramientas). Por ejemplo, cuando el idioma actual es el inglés estadounidense (página de códigos 1252: Latin 1 de Windows), al presionar la tecla ALT y escribir después 0163 en el teclado numérico, se produce £, el símbolo de la libra (U+00A3). Cuando el idioma actual es el ruso (página de códigos 1251: cirílico de Windows), la misma secuencia de teclas produce la letra mayúscula cirílica JE (U+0408).

Si el primer dígito que escribe es un número del 1 al 9, el valor se reconoce como un punto de código en la página de códigos OEM del sistema. El resultado difiere en función del idioma del sistema Windows especificado en Configuración regional y de idioma del Panel de control. Por ejemplo, si el idioma del sistema es inglés estadounidense, la página de códigos es 437 (MS-DOS Latín EE.UU.) y si se

presiona la tecla ALT y se escribe después 163 en el teclado numérico, se produce ú (U+00FA, letra minúscula latina U con acento). Si el idioma del sistema es griego (página de códigos OEM 737 MSDOS griego), la misma secuencia produce la letra minúscula griega MU (U+03BC).