Martínez Soto Mariana
Grupo 7

Tarea 4:
Codificación de caracteres en unicode.
Unicode

Unicode es un set de caracteres universal, es decir, un estándar en el que se definen todos los caracteres necesarios para la escritura de la mayoría de los idiomas hablados en la actualidad que se usan en la computadora. Su objetivo es ser, y, en gran medida, ya lo ha logrado, un superconjunto de todos los sets de caracteres que se hayan codificado.

El texto que aparece en la computadora o en la Web se compone de caracteres. Los caracteres representan letras del abecedario, signos de puntuación y otros símbolos.

En el pasado, distintas organizaciones han recopilado diferentes sets de caracteres y han creado codificaciones específicas para ellos. Un set puede abarcar tan sólo los idiomas de Europa occidental con base en el latín (sin incluir países de la UE como Bulgaria o Grecia), otro set puede contemplar un idioma específico del Lejano Oriente (como el japonés), y otros pueden ser parte de distintos sets diseñados especialmente para representar otro idioma de algún lugar del mundo.

Lamentablemente, no es posible garantizar que su aplicación particular pueda soportar todas las codificaciones, ni que una determinada codificación pueda soportar todos sus requerimientos para la representación de un cierto idioma. Además, generalmente resulta imposible combinar distintas codificaciones en la misma página web o en una base de datos, por lo que siempre es muy difícil soportar páginas plurilingües si se aplican enfoques "antiguos" cuando se trata de tareas de codificación.

El Consorcio Unicode proporciona un único y extenso set de caracteres que pretende incluir todos los caracteres necesarios para cualquier sistema de escritura del mundo, incluyendo sistemas ancestrales (como el cuneiforme, el gótico y los jeroglíficos egipcios). Hoy resulta fundamental para la arquitectura de la Web y de los sistemas operativos, y las principales aplicaciones y navegadores web incluyen soporte para este elemento. En el Estándar Unicode también se describen las propiedades y algoritmos necesarios para trabajar con caracteres.

Este enfoque facilita mucho el trabajo con sistemas o páginas plurilingües y responde mucho mejor a las necesidades del usuario que la mayoría de los sistemas de codificación tradicionales.

Es importante hacer una clara distinción entre los conceptos "set de caracteres" y "codificación de caracteres".

Martínez Soto Mariana
Grupo 7

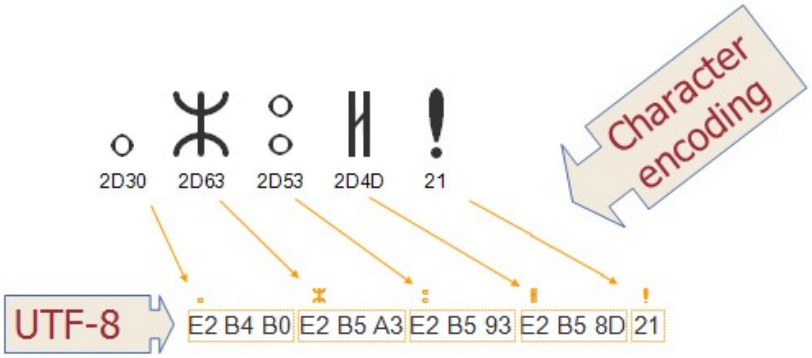
Un set de caracteres o repertorio comprende el grupo de caracteres que se utilizarían para una finalidad específica, ya sea los necesarios para el soporte de los idiomas de Europa Occidental en la computadora, o los que aprendería en el colegio un niño chino en tercer grado (sin relación con la computadora).

Un set de caracteres codificados es un grupo de caracteres en el que se ha asignado un número exclusivo a cada carácter. Las unidades de un set de caracteres codificados se conocen como puntos de código. El valor de un punto de código representa la ubicación de un carácter en el set de caracteres codificados. Por ejemplo, el punto de código para la letra á en el set de caracteres codificados Unicode es 225 en notación decimal, o E1 en notación hexadecimal. (Tenga presente que la notación hexadecimal generalmente se utiliza para hacer referencia a puntos de código y es la que se usará aquí).

En algunos casos, los sets de caracteres codificados se denominan páginas codificadas.

La codificación de caracteres refleja la manera en la que el set de caracteres codificados se convierte a bytes para su procesamiento en la computadora. En la siguiente imagen se muestra cómo se convierten a secuencias de bytes en memoria los caracteres y puntos de código del sistema de escritura Tifinagh (Berber) mediante la codificación UTF-8. Los valores de los puntos de código para cada carácter se enumeran inmediatamente debajo del glifo (es decir, la representación visual) correspondiente a dicho carácter en la parte superior del diagrama. Las flechas indican de qué manera estos elementos se convierten en secuencias de bytes, donde cada byte está representado por un número hexadecimal de dos dígitos. Observe cómo los puntos de código de Tifinagh se convierten en tres bytes mientras que el signo de exclamación se convierte en un solo byte.

Martínez Soto Mariana
Grupo 7



Un set de caracteres, múltiples codificaciones. Muchas normas en materia de codificación de caracteres, como aquellas incluidas en la serie ISO 8859, emplean un solo byte para un determinado carácter y la codificación es una conversión sencilla a la ubicación escalar de los caracteres en el set de caracteres codificados. Por ejemplo, en el set de caracteres codificados de la norma ISO 8859-1, la letra A se encuentra en la ubicación N.° 65 (comenzando por cero) y está codificada para representación en la computadora mediante un byte al que corresponde el valor 65. Esta organización de ISO 8859-1 es fija y no se modifica.

En Unicode, sin embargo, el proceso no es tan sencillo. Mientras que el punto de código para la letra á en el set de caracteres codificados Unicode es siempre 225 (en decimal), en UTF-8 se representa en la computadora mediante dos bytes. En otras palabras, no existe una correspondencia uno a uno entre el valor del set de caracteres codificados y el valor de codificación para este carácter.

Además, en Unicode existen distintas formas de codificar el mismo carácter. Por ejemplo, la letra á se puede representar mediante dos bytes en una codificación y con cuatro bytes, en otra. Los formatos de codificación que se pueden usar con Unicode se denominan UTF-8, UTF-16 y UTF-32.