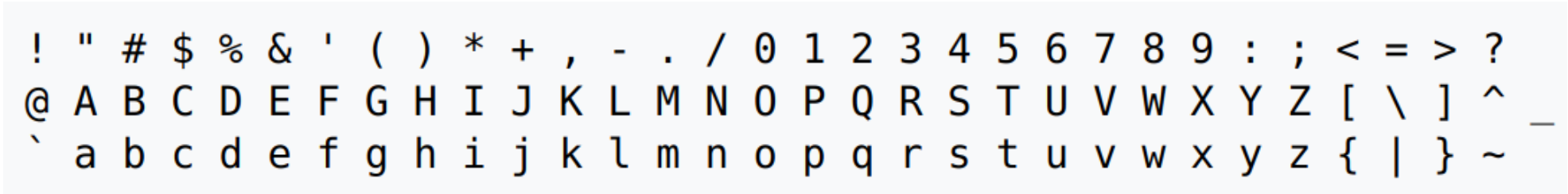
Codificación de caracteres

Sistemas informáticos

ASCII



- *American Standard Code for Information Interchange*
- Standard ASCII:
  - 7 bits de longitud con un bit extra de paridad
  - 128 caracteres
  - Está orientado al inglés
  - No se pueden representar tildes, la ñ...
- Extended ASCII:
  - Añade un bit más que permite la representación de otros caracteres vinculados a otros idiomas que usan el alfabeto latino
  - 256 caracteres

UNICODE

- Es un estándar de codificación de caracteres diseñado para facilitar el tratamiento informático, transmisión y visualización de textos de numerosos idiomas y disciplinas técnicas, además de textos clásicos de lenguas muertas.
- El término Unicode proviene de los tres objetivos perseguidos: universalidad, uniformidad y unicidad.
- La versión 13.0 contiene 143924 caracteres provenientes de alfabetos, sistemas ideográficos y colecciones de símbolos.
- No hay un número fijo de bits para representar un caracter, dependerá del caracter a representar y de la forma de codificación. En cualquier caso todos los caracteres se representarán con entre 8 y 32 bits.
- Existen tres formas de codificiación Unicode: UTF-8, UTF-16 y UTF-32.

UTF-8

- En esta forma de codificación cada carácter ocupará como mínimo un byte (8 bits) de forma que los caracteres existentes en ASCII extendido ocuparán 8 bits y otros ocuparán más.
- Esta forma de codificación es ventajosa para textos que utilizan principalmente caracteres ASCII ya que permiten codificarlos con tan solo 8 bits.
- El resto de caracteres se representarán con un código de 32 bits

UTF-16

- En esta forma de codificación cada carácter ocupará como mínimo dos bytes (16 bits)
- Este formato es más adecuado cuando la mayor parte de los caracteres del texto no están incluidos en la codificación ASCII.
- La mayoría de los caracteres más utilizados ASCII y no ASCII se codificarán en 2 bytes (16 bits) mientras que el resto de caracteres se representarán con un código de 32 bits.

UTF-32

- En este formato todos los caracteres ocupan 32 bits.
- La ventaja de este formato es que todos los caracteres tienen la misma longitud en bits y eso puede resultar práctico para ciertas funciones de procesamiento.

Otros formatos de codificación

- ISO 8859-1 Europa occidental
- ISO 8859-2 Europa occidental y Centroeuropa (checo, polaco, croata, rumano, esloveno, …)
- ISO 8859-3 Europa occidental y Europa del Sur
- ISO 8859-4 Europa occidental y países bálticos (lituano, estonio y lapón)
- ISO 8859-5 alfabeto cirílico
- ISO 8859-6 árabe
- ISO 8859-7 griego
- ISO 8859-8 Hebreo
- ISO 8859-9 Europa occidental con el juego de caracteres turco
- ISO 8859-10 Europa occidental con juegos de caracteres nórdicos, incluyendo el de Islandia.
- ISO 8859-11 tailandés
- ISO 8859-13 idiomas bálticos y polaco
- ISO 8859-14 idiomas celtas (gaélico irlandés, escocés, galés)
- ISO 8859-15 Añade el símbolo de Euro y otros a ISO 8859-1
- ISO 8859-16 idiomas centroeuropeos (polaco, checo, esloveno, eslovaco, húngaro, albano, rumano, alemán e italiano)