# PRACTICA 3, PABLO MENDOZA

1. Información básica:
    a. Definiciones:
        i. Repertorio de caracteres: Es el conjunto de caracteres y símbolos (sean imprimibles o de instrucción interna) que se busca representar.
        ii. Tabla de caracteres: Es la asignación única de cada carácter que se quieres representar a un valor o código que lo representa.
        iii. Codificación de caracteres: Es la forma de codificar el valor único que representa a un carácter (No el carácter en sí), en un código distinto al usado en tabla, como puede ser binario o Morse.
    b. UTF-8 y UTF-16 son dos formas de codificación de caracteres reconocidas por el repertorio Unicode y por la UTC. Ambos son capaces de representar todos los caracteres Unicode. Además, UTF-16 es un sucesor de UTF-8, debido a ventajas que ofrece sobre este último.
    c. No. UTF-8 usa de 1 a 4 Bytes por carácter Unicode, mientras que UTF-16 usa 2 o 4 (1 o 2 palabras de 16 bits). UTF-8 usa 1 byte para los símbolos comunes (US-ASCII), 2 para los símbolos romances, 3 para completar el Plano básico multilingüe, y 4 para los símbolos de uso poco común y símbolos matemáticos. Sin embargo, UTF-16 usa una palabra de 16 bits para todo el Plano Básico Multilingüe, y 2 para símbolos fuera de este, así que UTF-16 es, con un uso común una codificación de longitud fija.
2. Archivo de texto de Pablo Neruda
    a. Tamaños de archivo por codificaciones:
        i. 117 caracteres en total
        ii. ANSI: 121 Bytes (117+4)
        iii. Unicode y Unicode BE: 244 Bytes (117*2+10)
        iv. UTF-8: 131 Bytes(117+14)
    b. PSPAD:
        i. Longitud de palabra:
            1. ANSI: 1 Byte
            2. Unicode: 2 Bytes (Una palabra de 16 bits)
            3. Unicode BE: 2 Bytes (una palabra de 16 bits, pero al revés que Unicode)
            4. UTF-8: 1 Byte
        ii. No he visto que se represente nada de distinta forma, excepto unos bits al principio del texto (En Unicode y Unicode BE son el BOM, y en UTF-8 son EFBB)
        iii. El BOM en Unicode y Unicode BE están al revés.
        iv. El BOM entre los Unicode dice que, como son símbolos comunes, realmente no ocupan toda la palabra de 16 bits, solo 8, que en Unicode van al final los ceros, y en BE van al principio de la palabra.

3. El charset dice que está en ISO 8859-1
    a.
        i. En Firefox: Herramientas>Opciones>General>En el botón que dice Idiomas
        ii. En Chrome: Menú Hamburguesa> Más herramientas>Codificación
    b. Hay algunos caracteres que no se muestran correctamente debido a que los leemos con codificación distinta en la que se guardaron. Estos caracteres suelen ser caracteres acentuados y la Ñ.
4. Glosario:
    a. ANSI o US-ASCII: Código estándar americano para el intercambio de información: Sistema de codificación que usa 7 bits para representar enteros del 0 al 127 de su tabla de asignación, y un bit para comprobación de errores.
    b. Unicode: Es una gran tabla de caracteres que contiene símbolos de cientos de idiomas y miles de símbolos especiales como los matemáticos, griegos, y musicales. Pretende ser la tabla de caracteres universal.
    c. UTF-8: Es una codificación reconocida por Unicode, capaz de representar todos sus caracteres en 1, 2, 3 o 4 bytes, dependiendo del conjunto donde se encuentre el carácter.
    d. UTF-16: Es una evolución del UTF-8, también reconocido por Unicode, que representa todos sus caracteres en una o dos palabras de 16 bits cada una. Cuando se trata de un carácter de la Tabla Básica Multilingüe, solo usa una palabra, por lo que esta codificación es normalmente fija en longitud de carácter.
    e. BOM: Bit Order Marker: Es un marcador (FEFF o FFFE), que indica, en Unicode, el orden de los bytes, o indica que codificación de UTF se usa en cada momento.