

Nombre:

EJERCICIOS 1

TEMA 1. CODIFICACIONES DE CARACTERES

1. Nos encontramos con el siguiente repertorio de caracteres:

{a,e,i,o,u,A,E,I,O,U,SP,b,c,d,f,g,h,j,k,l,B,C,D,F,G,H,J,K,L,CR}

Observa que el mismo contiene el espacio en blanco, SP, y el retorno de carro o carácter de nueva línea, CR.

- a) Propón una tabla de caracteres para el mismo.
- b) Propón una regla de codificación de los mismos (de todos los caracteres), cuya longitud de palabra sea 2 bytes. Puedes utilizar sistema binario o hexadecimal.
- c) Con respecto a esa regla de codificación, escribe la representación en hexadecimal del contenido de un fichero de texto que contenga la siguiente cadena de caracteres:

*Coge el agua debajo
Ulula la hoja*

- d) ¿Cuánto ocuparía (en bytes o en bits) el anterior fichero de texto (suponiendo que no se le añadan caracteres de control) al guardarlo en un dispositivo de almacenamiento?

2. Escribimos en un editor de texto la siguiente frase:

En España hay cigüeñas pero no ñandús

- a) ¿Puede guardarse el siguiente texto con codificación ASCII (128 caracteres)?
- b) ¿Puede guardarse en codificación ISO-8859-1 (ISO-Latin 1)? ¿Cuánto ocuparía y por qué?
- c) ¿Puede guardarse en codificación UTF-16? ¿Cuánto ocuparía y por qué?

3. Alguien nos plantea un problema que le ha surgido con un fichero que le ha mandado un compañero. Lee el archivo con un editor de texto y ve lo siguiente:

Đand• ±and+ Cig³e±a Cig_eĐa

Su autor le jura que escribió lo siguiente:

Ñandú ñandÚ Cigüeña CigÜeÑa

Ayudado por la siguiente información explica detalladamente qué ha sucedido.

Tabla MSDOS 850:

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
80	Ç 00C7	ü 00FC	é 00E9	ã 00E2	ä 00E4	à 00E0	å 00E5	ç 00E7	ê 00EA	ë 00EB	è 00E8	ì 00EF	í 00EE	î 00EC	Ä 00C4	Å 00C5
90	É 00C9	æ 00E6	Æ 00C6	ô 00F4	ö 00F6	ò 00F2	û 00FB	ù 00F9	ÿ 00FF	Ö 00D6	Ü 00DC	ø 00F8	£ 00A3	Ø 00D8	× 00D7	f 0192
A0	á 00E1	í 00ED	ó 00F3	ú 00FA	ñ 00F1	Ñ 00D1	ª 00AA	º 00BA	¿ 00BF	® 00AE	¬ 00AC	¼ 00BD	½ 00BC	¡ 00A1	« 00AB	» 00BB
B0	2591	2592	2593	2502	2524	Á 00C1	Ã 00C2	À 00C0	© 00A9	¶ 2563	2551	¶ 2557	¶ 255D	¢ 00A2	¥ 00A5	¿ 2510
C0	L 2514	l 2534	T 252C	t 251C	— 2500	† 253C	ã 00E3	Ä 00C3	ℒ 255A	℔ 2554	℔ 2569	℔ 2566	℔ 2560	= 2550	† 256C	» 00A4
D0	ø 00F0	Ð 00D0	Ê 00CA	Ë 00CB	È 00C8	ı 0131	Í 00CD	Î 00CE	İ 00CF	Ј 2518	Г 250C	■ 2588	■ 2584	ı 00A6	İ 00CC	■ 2580
E0	Ó 00D3	ß 00DF	Ö 00D4	Ò 00D2	ø 00F5	Õ 00D5	µ 00B5	þ 00FE	ƒ 00DE	Ú 00DA	Û 00DB	Ü 00D9	Ý 00FD	Ý 00DD	— 00AF	‘ 00B4
F0	— 00AD	± 00B1	≡ 2017	¾ 00BE	¶ 00B6	÷ 00A7	÷ 00F7	° 00B8	° 00B0	° 00A8	° 00B7	° 00B9	° 00B3	° 00B2	25A0	NBSP 00A0

Tabla Windows 1252:

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
80	€ 20AC		ı 201A	f 0192	" 201E	... 2026	† 2020	‡ 2021	~ 02C6	% 2030	Š 0160	< 2039	Œ 0152		Ž 017D	
90		\ 2018	/ 2019	" 201C	" 201D	• 2022	— 2013	— 2014	~ 02DC	™ 2122	Š 0161	> 203A	œ 0153		ž 017E	Ÿ 0178
A0	NBSP 00A0	ı 00A1	¢ 00A2	£ 00A3	¤ 00A4	¥ 00A5	¦ 00A6	§ 00A7	¨ 00A8	© 00A9	ª 00AA	« 00AB	¬ 00AC	­ 00AD	® 00AE	¯ 00AF
B0	° 00B0	± 00B1	² 00B2	³ 00B3	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	¹ 00B9	º 00BA	» 00BB	¼ 00BC	½ 00BD	¾ 00BE	¿ 00BF
C0	À 00C0	Á 00C1	Â 00C2	Ã 00C3	Ä 00C4	Å 00C5	Æ 00C6	Ç 00C7	È 00C8	É 00C9	Ê 00CA	Ë 00CB	Ì 00CC	Í 00CD	Î 00CE	Ï 00CF
D0	Ð 00D0	Ñ 00D1	Ò 00D2	Ó 00D3	Ô 00D4	Õ 00D5	Ö 00D6	× 00D7	Ø 00D8	Ù 00D9	Ú 00DA	Û 00DB	Ü 00DC	Ý 00DD	Þ 00DE	ß 00DF
E0	à 00E0	á 00E1	â 00E2	ã 00E3	ä 00E4	å 00E5	æ 00E6	ç 00E7	è 00E8	é 00E9	ê 00EA	ë 00EB	ì 00EC	í 00ED	î 00EE	ï 00EF
F0	ø 00F0	ñ 00F1	ò 00F2	ó 00F3	ô 00F4	õ 00F5	ö 00F6	÷ 00F7	ø 00F8	ù 00F9	ú 00FA	û 00FB	ü 00FC	ý 00FD	þ 00FE	ÿ 00FF

4. Define brevemente los siguientes conceptos y contesta a las preguntas formuladas:

a) Unicode.

b) UTF8 y UTF16. ¿Cuáles son sus longitudes de palabra? ¿Son compatibles (es decir, existen caracteres que comparten la misma representación en ambas codificaciones)? ¿Hay alguna diferencia en el rango de caracteres que pueden representar?

c) Observa el siguiente comentario extraído de un foro donde se trata de explicar las diferencias entre ISO-8859-1 y UTF8 a la hora de ser utilizados para construir páginas web (compara las afirmaciones vertidas con tu propia experiencia y tus conocimientos):

“Ventajas y desventajas de ISO-8859-1:

Todos los editores de texto, editores de html manejan sin problemas esta codificación con lo que te ahorras dolores de cabeza. Lo malo son los acentos, ya que si bien los puedes escribir tal cual no se verán correctamente y tendrás que usar ´ y demás.”

Realiza un comentario crítico sobre la anterior opinión, en especial de la segunda afirmación.

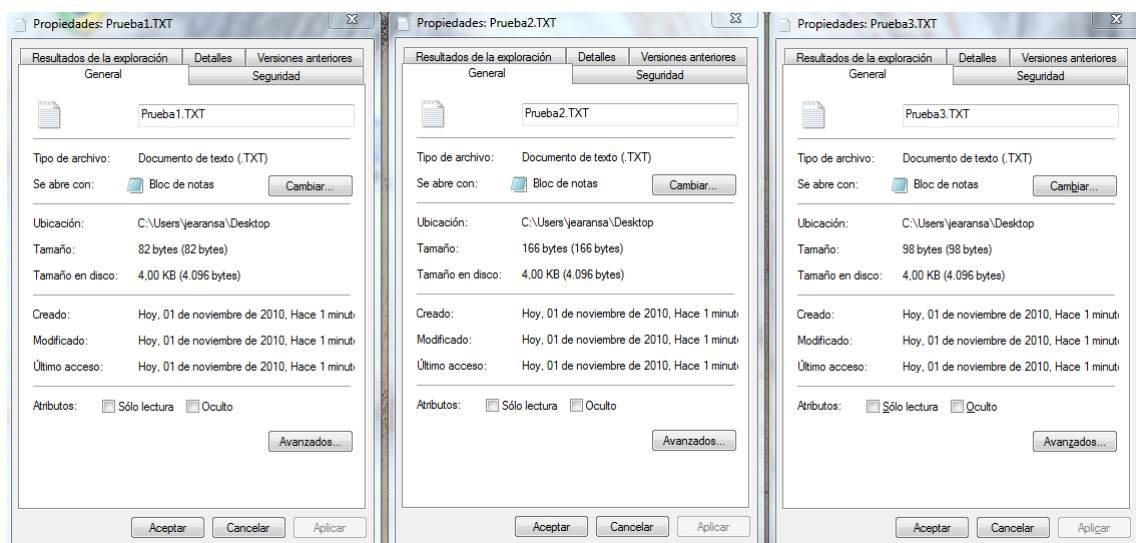
d) Observa ahora lo que el mismo individuo comenta sobre UTF8 (sic):

“ventajas y desventajas de UTF8:

Olvidate del lenguaje que lo escribas (salvo japones y demás idiomas idiogramáticos que usan UTF-16) porque te lo va a coger tal cual, podrás escribir directamente los acentos y siempre se verán correctamente.”

Realiza un nuevo comentario crítico sobre las dos afirmaciones que contiene el mismo.

5. Acabamos de guardar un texto en nuestro ordenador en tres ficheros distintos de texto. Nos encontramos con las siguientes propiedades de los mismos:



- a) ¿Por qué los ficheros tienen distintos tamaños almacenando el mismo texto?
- b) Alguien nos dice que el texto en Prueba1.TXT ha sido guardado con una codificación de longitud de palabra 1. A cambio de su información nos formula las siguientes preguntas ¿Qué codificación hemos usado, más probablemente, en Prueba3.TXT? ¿Y en Prueba2.TXT? ¿Cuántos caracteres “especiales” (no ASCII) hemos usado en el texto como máximo?

6. Responde cuáles de las siguientes afirmaciones son verdaderas o falsas. Todas las respuestas deben ir acompañadas de una justificación razonada para ser consideradas correctas:

- a) Las codificaciones de longitud de palabra 1 byte pueden representar un máximo de 256 caracteres.
- b) Todas las codificaciones de la familia ISO – 8859 representan el mismo repertorio de caracteres, aunque estos puedan encontrarse en distintas posiciones de sus correspondientes tablas.
- c) El repertorio de caracteres de UTF – 16 y UTF – 16BE es mayor que el de UTF – 8.

7. El código fuente de la página web de la asignatura de nuestro compañero, a través del navegador, se ve de la siguiente forma:

```

1 <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
2 "http://www.w3.org/TR/html4/strict.dtd">
3 <html>
4 <head>
5 <meta http-equiv="content-type" content="text/html; charset=utf-8">
6 <meta name="keywords" content="Sistemas Informáticos, Grado en Ingeniería Informática,
7 Universidad de La Rioja, tu nombre">
8 <meta name="description" content="En esta página se pueden encontrar enlaces a las
9 soluciones de las prácticas de Sistemas Informáticos para el año 2010.
10 También contendrá un enlace al trabajo de la asignatura. Todo el
11 contenido ha sido elaborado por nombre.">
12 <meta name="Author" content="Tu nombre">
13 <meta name="Date" content="30, Septiembre 2010">
14 <meta name="robots" content="all">
15 <meta name="generator" content="PSPad editor, www.pspad.com">
16 <title>Página web de Tu Nombre</title>
17 </head>
18 <body>
19 <p>
20 Bienvenido a la página web para la asignatura Sistemas Informáticos de Tu Nombre.
21 </p>
22 <p>
23 A continuación puedes ver un listado con enlaces a los informes de las prácticas de la
24 misma:
25 </p>
26 <ul>
27 <li><a href="informes/Practica01.txt">Práctica 01</a></li>
28 <li><a href="informes/Practica02.txt">Práctica 02</a></li>
29 <li>Añade tantas entradas como necesites </li>
30 </ul>
31 </body>
32 </html>

```

Él nos asegura que, como se puede observar, ha definido bien el “meta” correspondiente al “charset”. Hazle un diagnóstico de los errores que ha cometido, y cómo puede solucionarlo.

A los pocos días nos lo volvemos a encontrar y nos cuenta que, tras varios días sin dormir, se ha dado cuenta de que si le indica al navegador que muestre la página con la codificación “ISO-8859-1”, lo que ve ahora es:

Bienvenido a la página web para la asignatura Sistemas Informáticos de Tu Nombre.

A continuación puedes ver un listado con enlaces a los informes de las prácticas de la misma:

- [Práctica 01](#)
- [Práctica 02](#)
- Añade tantas entradas como necesites

Completa tu primer diagnóstico con una explicación completa de lo sucedido.

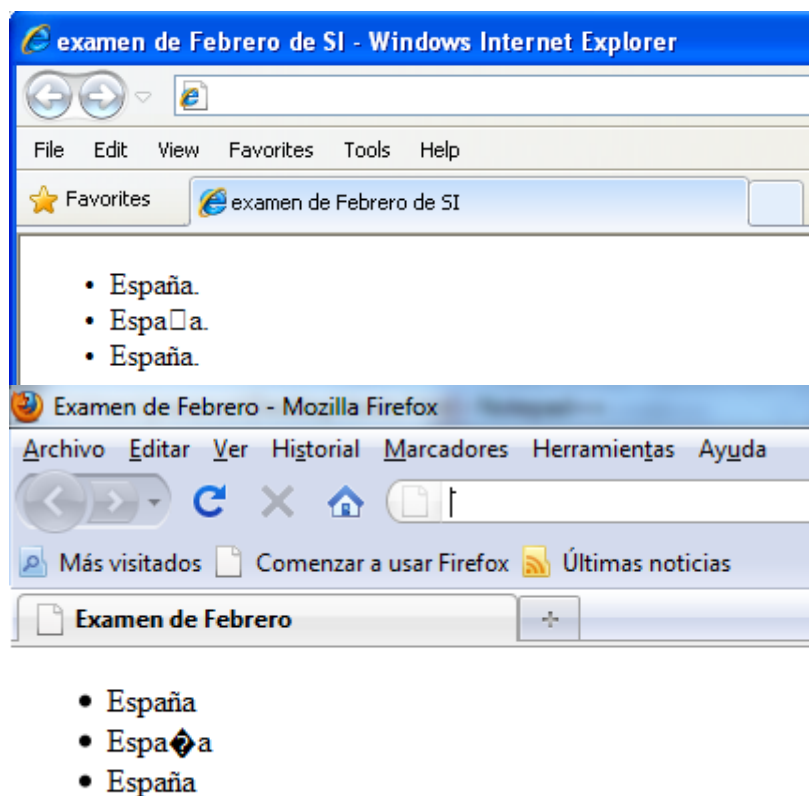
8. Las especificaciones HTML 4.01 indican lo recogido en el punto 5.4. Explica razonadamente dónde y cómo se aprecia su cumplimiento en la representación del siguiente fichero en IE y Mozilla Firefox.

5.4 Caracteres no representables

Un agente de usuario puede no ser capaz de representar correctamente todos los caracteres de un documento, por ejemplo, porque el agente de usuario carezca de una fuente apropiada, porque un carácter tenga un valor que no pueda expresarse con la codificación de caracteres interna del agente de usuario, etc.

Al haber tantas cosas diferentes que pueden hacerse en estos casos, este documento no prescribe ningún comportamiento por defecto. [...] En ausencia de un comportamiento más sofisticado [...], recomendamos el siguiente comportamiento a los agentes de usuario:

1. Adoptar un mecanismo claramente visible, pero no molesto, para alertar al usuario sobre los recursos ausentes.



9. Nos encontramos con el siguiente texto en formato Hexadecimal:

A14D	6520	6865	206C	6576	616E	7461	646F
2063	6173	6920	6120	6C61	7320	3130	2120
0D0A	7920	6D65	2068	6520	7175	6564	6164
6F20	656E	206C	6120	6361	6D61	206D	E173
2064	6520	5472	6573	2043	7561	7274	6F73
2064	6520	486F	7261				

Su autor nos dice que ha usado la siguiente tabla para codificarlo (ISO-8859-1):

8 bit ASCII codes

The space character appears in place of unprintable characters

0	32	64 @	96 ' 128	160	192 À	224 à
1	33 !	65 A	97 a 129	161 ¡	193 Á	225 á
2	34 "	66 B	98 b 130	162 ¢	194 Â	226 â
3	35 #	67 C	99 c 131	163 £	195 Ã	227 ã
4	36 \$	68 D	100 d 132	164 ¤	196 Ä	228 ä
5	37 %	69 E	101 e 133	165 ¥	197 Å	229 å
6	38 &	70 F	102 f 134	166 ¦	198 Æ	230 æ
7	39 '	71 G	103 g 135	167 §	199 Ç	231 ç
8	40 (	72 H	104 h 136	168 ¨	200 È	232 è
9	41)	73 I	105 i 137	169 ©	201 É	233 é
10	42 *	74 J	106 j 138	170 ®	202 Ê	234 ê
11	43 +	75 K	107 k 139	171 «	203 Ë	235 ë
12	44 ,	76 L	108 l 140	172 ¬	204 Ì	236 ì
13	45 -	77 M	109 m 141	173 -	205 Í	237 í
14	46 .	78 N	110 n 142	174 ®	206 Î	238 î
15	47 /	79 O	111 o 143	175 ¯	207 Ï	239 ï
16	48 0	80 P	112 p 144	176 °	208 Ð	240 ð
17	49 1	81 Q	113 q 145	177 ±	209 Ñ	241 ñ
18	50 2	82 R	114 r 146	178 º	210 Ò	242 ò
19	51 3	83 S	115 s 147	179 ¸	211 Ó	243 ó
20	52 4	84 T	116 t 148	180 ´	212 Ô	244 ô
21	53 5	85 U	117 u 149	181 µ	213 Õ	245 õ
22	54 6	86 V	118 v 150	182 ¶	214 Ö	246 ö
23	55 7	87 W	119 w 151	183 ·	215 ×	247 ÷
24	56 8	88 X	120 x 152	184 ,	216 Ø	248 ø
25	57 9	89 Y	121 y 153	185 ´	217 Ù	249 ù
26	58 :	90 Z	122 z 154	186 ¸	218 Ú	250 ú
27	59 ;	91 [	123 { 155	187 »	219 Û	251 û
28	60 <	92 \	124 156	188 ¼	220 Ü	252 ü
29	61 =	93]	125 } 157	189 ½	221 Ý	253 ý
30	62 >	94 ^	126 ~ 158	190 ¾	222 Þ	254 þ
31	63 ?	95 _	127 159	191 ¿	223 ß	255 ÿ

Recupera el texto original (la posición 32 de la tabla corresponde con el espacio en blanco, y las posiciones 13 + 10 con el salto de línea propio de los sistemas "Windows").

10. Tenemos la siguiente descripción de la regla de codificación de UTF8 para caracteres Unicode:

- Los caracteres que en Unicode ocupan las posiciones (en Hexadecimal) entre 000000 hasta 00007F (dicho de otra forma, los 128 caracteres ASCII), se codifican en binario por medio de la secuencia de bits "0xxxxxxx", donde "xxxxxxx" debe ser reemplazado por la codificación en binario de la posición en la tabla Unicode del carácter deseado;
- Los caracteres que en Unicode ocupan las posiciones (en Hexadecimal) entre 000080 (posición 129) hasta 0007FF (posición 2048) se codifican por medio de dos bytes de la forma "110xxxxx 10xxxxxx", donde "xxxxx" y "xxxxxx" son sustituidos por la codificación en binario de la posición en la tabla Unicode del carácter deseado.

Aunque no forma parte de la regla, se sobreentiende que los bits sobrantes (por la izquierda) deben completarse con ceros.

A partir de la anterior información, muestra la representación binaria en UTF - 8 de los siguientes caracteres:

- a) μ
- b) T
- c) æ
- d) &
- e) Ä

La tabla Unicode tiene las siguientes posiciones:

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0000	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
0010	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
0020	SP	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
0030	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
0040	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0050	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
0060	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
0070	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL
0080	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
0090	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
00A0	NBSP	ı	ç	£	¤	¥	¦	§	¨	©	ª	«	¬	SHY	®	¯
00B0	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
00C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
00D0	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
00E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
00F0	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ
U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0100	Ā	ā	Ă	ă	Ą	ą	Ć	ć	Č	č	Ĉ	ĉ	Ċ	ċ	Ď	ď