L1 Informatique – EEEA
Logique Combinatoire et Séquentielle
TD n°4
Codage des informations alphanumériques

Exercice 1 – Codage ASCII

Lorsque vous appuyez sur la touche d'un clavier, celui-ci génère un code associé au caractère indiqué sur la touche. Le code le plus utilisé actuellement pour coder les caractères alphanumériques est le code ASCII. C'est aussi ce code qui est utilisé pour coder les informations sauvegardées dans un fichier texte simple ou pour s'échanger des messages texte par courrier électronique.

La figure 1 donne la définition de ce code. On doit lire ce tableau de la façon suivante : la lettre B est dans le tableau à la ligne 0x40 et à la colonne 2 : son code est 0x42 (remarque la notation 0x signifie simplement que les caractères suivants sont des caractères hexadécimaux : le code ASCII de la lettre B est donc $(42)_{16}$).

code	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0x00	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	NP	CR	SO	SI
0x10	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
0x20	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
0x30	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
0x40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0x50	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
0x60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
0x70	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

FIGURE 1 – Table des codes ASCII

Que signifie le message suivant ?

4C 43 53 20 3A 20 46 69 6E 20 64 65 20 6C 61 20 73 26 65 61 63 75 74 65 3B 61 6E 63 65 20 64 65 20 54 44 20 21

Exercice 2 – Codage UTF-8

Le codage UTF-8 est récemment devenu un standard pour le codage des caractères dans tous les alphabets répertoriés, sans pour autant multiplier la taille des fichiers. Il repose sur un codage des caractères sur 1, 2, 3 ou 4 octets. Le principe simplifié est le suivant :

- Tous les caractères codés sur 1 octet commencent par 0.
- Les autres caractères sont codés sur 2, 3 ou 4 octets avec le schéma suivant :
  - Les premiers bits du premier octet indiquent sur combien de bits est codé le caractère.
    - Si les premiers bits du premier octet sont 110, le caractère est codé sur 2 octets
    - Si les premiers bits du premier octet sont 1110, le caractère est codé sur 3 octets
    - Si les premiers bits du premier octet sont 11110, le caractère est codé sur 4 octets
  - Les octets suivants commencent tous par la séquence de bits 10.
  - Tous les bits non fixés précédemment participent au codage des caractères.

1. Combien de caractères peuvent être codés en UTF-8 sur 1 seul octet ?
2. Sachant qu'un fichier encodé en ASCII peut être décodé en UTF-8, qu'en déduisez-vous concernant les caractères d'UTF-8 pouvant être codés sur 1 octet.
3. Sachant que les caractères de code 128 à 255 du code ASCII étendu (aussi appelé ISO-8859-1) sont codés sur deux octets dans le codage UTF-8, quelle est approximativement la taille en octet d'un fichier encodé en UTF-8 de 1000 caractères dont 1 caractère sur 20 en moyenne est accentué ?
4. Combien de caractères différents permet de coder UTF-8 ?
5. Dans le codage Unicode, tous les caractères sont codés sur deux octets.
   (a) Combien de caractères permet de coder Unicode ?
   (b) Quel est l'inconvénient d'Unicode par rapport à UTF-8 ?
6. Expliquer en quoi UTF-8 est robuste aux erreurs de transmission conduisant à un octet supprimé ou à un octet modifié.