UNICODE

UNICODE DEFINES A LARGE TABLE OF CODE POINTS, THAT CAN BE USED FOR ALL SORTS OF LETTERS AND SYMBOLS.

BYTE ORDER MARK

THE BYTE ORDER MARK (BOM) IS A UNICODE CHARACTER.

U+FEFF is its unicode code point.

↑ UNICODE ↑ HEXADECIMAL POSITION

Code points

characters ARE REFERRED TO BY THEIR "UNICODE CODE POINT"

UNICODE CODE POINTS ARE WRITTEN IN HEXADECIMAL PRECEDED BY A "U+"

ENCODING

THERE ARE SEVERAL WAYS TO ENCODE UNICODE CODE POINTS INTO BITS. E.G.

UTF-8, UTF-16 } VARIABLE LENGTH ENCODINGS

UTF-32 } FIXED LENGTH → EACH CHAR TAKES UP 32 BITS

A → UTF-8 (8 BITS) ~ VARIES

€ → UTF-8 (24 BITS) /

EXERCISE 1

ENCODE THE EURO SYMBOL AS A UTF-8 CHARACTER. ITS CODE POINT IS U+20AC

TO DO THIS WE WILL NEED TO USE A TABLE.

UNICODE TABLE

BITS OF CODE POINT	UNICODE CODE POINT RANGE: FIRST CODE POINT	UNICODE CODE POINT RANGE: LAST CODE POINT	BYTES IN SEQUENCE	BYTE 1	BYTE 2	BYTE 3	BYTE 4
7	U+0000	U+007F	1	0xxxxxxx			
11	U+0080	U+07FF	2	110xxxxx	10xxxxxx		
16	U+0800	U+FFFF	3	1110xxxx	10xxxxxx	10xxxxxx	
21	U+10000	U+1FFFF	4	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

EXERCISE 2

ENCODE THE CHARACTER A AS A UTF-8 CHARACTER. ITS CODE POINT IS

WE WILL USE A TABLE

U+0041

↓ ↓

0100 0001

FOR AN ALPHABETIC CHARACTER E.G. A...Z a...z

THIS IS ALL THAT IS NEEDED.

0100 0001 IS THE 1 BYTE IT TAKES UP.

EXPRESS IN HEX: 41h

EXERCISE 1

ENCODE THE EURO SYMBOL AS A UTF-8 CHARACTER.
ITS CODE POINT IS U+20AC.

EXPRESS YOUR ANSWER IN hex for convenience.

(i) USE TABLE, LOOK UP CODE POINT – CHECK WHAT RANGE IT IS IN.

USE 3RD LINE FROM TABLE: => 3 BYTES NEEDED

BYTE 1	BYTE 2	BYTE 3
1110xxxx	10xxxxxx	10xxxxxx
AAAA	BBBBBB	cccccc

20AC

0010 0000 1010 1100
AAAA BBBB BBcc cccc

BYTE 1	BYTE 2	BYTE 3
11100010	10000010	10101100
E	2	8	2	A	C

E2 82 AC

~~4~~ (5)

DO YA WANNA BUILD A SNOWMAN ?

EXERCISE 3

ENCODE THE FOLLOWING AS A UTF-8 CHARACTER. EXPRESS YOUR ANSWER IN HEXADECIMAL:

U+2603

↑ UNICODE CODE POINT

E2 98 83