Character Encodings and Unicode

Exercise 1: How many bits would be required to uniquely identify 100,000 different characters? (Hint: $2^{16} = 65536$).

$2^{17} = 2^{1} \cdot 2^{16} = 2 \times 65536 > 100,000$

∴ 17 bits would be sufficient.

Exercise 2: The Chinese character for “Rice” (the grain) is 米 with Unicode value (code point) U+7C73. What is the UTF-8 encoding for this character?

Step 1: use E0 80 80 as prefixs

7C73

↓

0111 | 1100 01 | 110011

07 | 31 | 33

Step 2:

E0 + 07 = E7
80 + 31 = B1
80 + 33 = B3

Step 3

Exercise 3: Find the codepoint of the first Unicode character in the sequence of bytes: `A0 88 EB 8C 80 EC`.

└→ not an initial byte — Step 1

EB 8C 80 ↓ → Step 2

~~1110~~ 1011 ~~10~~00 1100 ~~10~~00 0000 ← Step 3

zzzzyyyy yyxxxxxx	1110zzzz	10yyyyyy	10xxxxxx

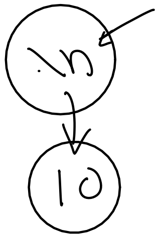
step 4

1011 | 0011 | 0000 | 0000

B | 3 | 0 | 0

Exercise 4: Four numbers are transmitted as the following CSV file:
2,_1\n
9,_3\n
How many bytes are required to transmit these four numbers formatted this way? Note that a “line feed” character is required at the end of each line and that spaces and commas also need to be transmitted.

How many bytes are required to transmit these four numbers if they are transmitted, one after another, if each is encoded as a 16-bit number? What if each number was encoded as a 32-bit number?



We need 10 characters (bytes). to transmit the 4 numbers as a text? (CSV) file.

0000.... 0010
0 0001
0 1001
0 0011
16 bits

2 bytes/number · 4 numbers = 8 bytes.

If used a 32-bit binary encoding
4 bytes/number · 4 = 16 bytes.