Solutions to Assignment 1

Version 2: Reworded solution to Question 4.

Question 1

Assuming the frequency of occurrence in this story is an accurate estimate of a message's probability,* we can compute the probability of message i† by dividing the number of occurrences of message i (N_i) by the total number of messages, $N = \sum_i N_i$:

$$P_i = \frac{N_i}{N}$$

The amount information contained in message i is given by:

$$I_i = -\log_2(P_i)$$

The amount of information in the story (I) is the sum of the information in its messages:

$$I = \sum_i N_i \times I_i$$

The supplied `.csv` file gives the values of N_i so we can compute N, I_i and I using the spreadsheet `sum` and `log` functions. Here is an example of the formulas (column B is N_i, column C is I_i and line 107 computes N and I):

	A	B	C
104	the	6	=B104*-LOG(B104/B$107,2)
105	tubes	1	=B105*-LOG(B105/B$107,2)
106	gardens	1	=B106*-LOG(B106/B$107,2)
107	total	=SUM(B2:B106)	=SUM(C2:C106)

(a) If each word is a message, the story contains $N = 159$ messages (words) and $I = 1018.7$ bits of information.

(b) Similarly, if each character is a message, the story contains $N = 783$ messages (characters) and $I = 3234.4$ bits of information.

(c) If we treat each character as a message with $I_i = 8$ bits of information then the story contains $783 \times 8 = 6264$ bits of information.

*Perhaps not a good assumption for such a short sample but that's all we're given.

†The subscript i refers to the i'th unique message, not the i'th message transmitted.

Question 2

To include the effects of all factors affecting the per-user throughput we can analyze a time interval that includes transmissions from each of 10 users with one short and one long frame from each one.

The elapsed time for this sequence would be:

$$T = 10 \times (T_{\text{short}} + 8 + T_{\text{long}} + 8)\,\mu\text{s}$$

where

$$T_{\text{short or long}} = \frac{8(10 + 7 + N_\text{p} + N_\text{d})}{2 \times 10^6}\,\mu\text{s}$$

where N_p is the number of parity bytes in the message: $N_\text{p} = 12 \times \lceil \frac{64}{64} \rceil = 12$ bytes for 64-bytes messages and $N_\text{p} = 12 \times \lceil \frac{1500}{64} \rceil = 288$ bytes for 1500-byte frames and N_d is the number of data bytes in the frames (64 or 1500). The spreadsheet below calculates the throughput for one user as 164 kb/s:

data bytes/frame	Nd	64	1500	bytes
parity bytes/frame	Np	12	288	bytes
frame duration	Tshort, Tlong	372.0E-6	7.2E-3	s
duration of 20 frames	T	76.1E-3		s
data bits/user/frame		12512		
data bits/user/s		164E+3		bps

Question 3

The UTF-8 encoding table in the Unicode specification (Table 3-6) shows that each byte's value determines the allowed position of that byte in a UTF-8 encoding:

- `00` to `7F`: first byte of a 1-byte encoding
- `80` to `BF`: a continuation byte
- `C0` to `DF`: first byte of a 2-byte encoding
- `E0` to `EF`: first byte of a 3-byte encoding
- `FF`: first byte of a 4-byte encoding

For the byte sequence:

`E1 A2 84 BE E3 81 AE 45 8A D0 B7`

(a) **E1** should be followed by 2 bytes. These are **A2** and **84** which are in the required range for continuation bytes so this is a valid 3-byte UTF-8 encoding.

The next byte, **BE**, is in the continuation byte range, thus cannot begin a UTF-8 encoding and *should be skipped*.

E3 should be followed by 2 bytes. These are **81** and **AE** which are in the required range so this is a valid 3-byte UTF-8 encoding.

45 should be followed by 0 bytes. This is a valid 1-byte UTF-8 encoding.

The next byte, **8A**, is in the continuation byte range, thus cannot begin a UTF-8 encoding and *should be skipped*.

D0 should be followed by 1 byte. This is **B7** which is in the required range so this is a valid 2-byte UTF-8 encoding.

Thus **BE** and **8A** are not part of valid UTF-8 sequences and should be skipped.

(b) The sequence **E1 A2 84** has a binary representation **1110 0001 1010 0010 1000 0100** from which we can extract the bits z=**0001**, y=**10 0010**, and z=**00 0100**, and the code point **U+1884**.

The sequence **E3 81 AE** has a binary representation **1110 0011 1000 0001 1010 1110** from which we can extract the bits z=**0011**, y=**00 0001**, and x=**10 1110**, and the code point **U+306E**.

The sequence **45** has a binary representation **0100 0101** from which we can extract the bits x=**100 0101**, and the code point **U+0045**.

The sequence **D0 B7** has a binary representation **1101 0000 1011 0111** from which we can extract the bits y=**1 0000**, x=**11 0111**, and the code point **U+0437**.

(c) The names of the corresponding characters are:

- **U+1884** is the MONGOLIAN LETTER ALI GALI INVERTED UBADAMA (ᢄ).
- **U+306E** is the HIRAGANA LETTER NO (の).
- **U+0045** is the ASCII E (E).
- **U+0437** is the CYRILLIC SMALL LETTER ZE (з).

Question 4

The probability that a bit is received in error is given in the question as $p = 10^{-6}$. Since there are only two possible outcomes (error or no error), the probability that a bit is not received in error must be $1 - p \approx 1$.

Each received character has 9 bits (8 data bits and 1 parity bit).

(a) When there is a sequence of independent outcomes (e.g. coin flips) the probability of a specific sequence of outcomes is given by the product of their individual probabilities.

The probability that the first bit is in error but the other 8 bits are not in error is the product of these probabilities: $p \times (1-p) \ldots \times (1-p) = p(1-p)^8 \approx 1 \times 10^{-6}$.

(b) The probability of one of several independent outcomes is given by the sum of the probabilities of these outcomes.

If we consider each received character as an outcome, there are 9 possible outcomes that have one bit in error[‡]. Each of these has the probability computed above. The sum of their probabilities is $9p(1-p)^8 \approx 9 \times 10^{-6}$. This is the probability that one bit is in error (any one bit, but exactly one).

(c) The probability of receiving a character that has two specific bits in error is $p^2(1-p)^7$. But there are

$$C(9,2) = \frac{9!}{2!\,(9-2)!} = \frac{9 \times 8}{2} = 36$$

possible ways of having 2 errors in 9 bits where $C(n,k)$ is the number of combinations of k things taken from n. Thus the probability of any two (but exactly two) bits being in errors in a character is $36p^2(1-p)^7 \approx 36 \times 10^{-12}$.

Thus, although a single parity bit does not detect two-bit errors, these are much less likely than single-bit errors (at low bit error rates, at least).

[‡]There are 8 possible locations for a data bit error and one possible location for the parity bit error