Information Theoretic Security

CPSC 530

Arnel Jerome Adviento

10130641

October 13, 2016

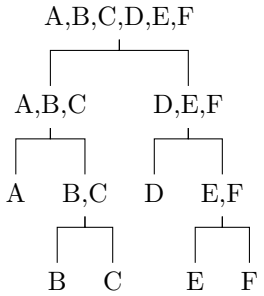
Q1

1.

see ClaudeShannon.docx

Q2

1.



(a).It will take 3 questions for Bob to know exactly what the output symbol is. This is concluded by halving the search space at each question similar to the tree above.

(b). the sequence of 3 questions to find out the output are:

- is it (A,B,C)?
- is it A or is it D
- is it B or is it E

(c) Only one is needed. since Bob only wants to know if the output is "A", he just needs to ask one question since its a specific output.

2.

$$\begin{aligned} H(X) &= \frac{1}{2}\log_2(2) + \frac{1}{4}\log_2(4) + \frac{1}{10}\log_2(10) + \frac{1}{15}\log_2(15) + \frac{1}{16}\log_2(16) + \frac{1}{48}\log_2(48) \\ &= \frac{1}{2} + \frac{1}{2} + 0.332 + 0.260 + \frac{1}{4} + 0.116 \\ &= 1.958 \end{aligned}$$

it is less than the value calculated in 1.(a) since the probability is not equally distributed between each letter

3.

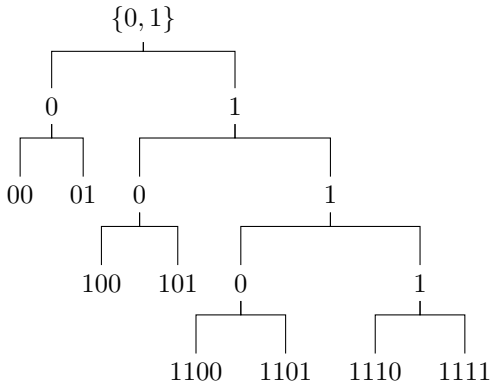
A →	0	B →	10
C →	110	D →	1110
E →	11110	F →	11111

Yes, it can relate to the "yes/no" questions in 1.(b)

- is it (does it have two or less 1's)?
- is it is it 0 or is it is there there 1's
- is it is there one 1's or is it is there four 1's

Q3

1.(a)



A →	00	B →	01
C →	100	D →	101
E →	1100	F →	1101
G →	1110	H →	1111

1.(b)

$$
\begin{aligned}
L_{exp} &= 2 * 0.25 + 2 * 0.25 + 3 * 0.15 + 3 * 0.15 + 4 * 0.05 + 4 * 0.05 + 4 * 0.05 + 4 * 0.05 \\
&= 0.5 + 0.5 + 0.45 + 0.45 + 0.2 + 0.2 + 0.2 + 0.2 \\
&= 2.7 \\
H(X) &= [0.25 \log_2(\frac{1}{0.25}) \times 2] + [0.15 \log_2(\frac{1}{0.15}) \times 2] + [0.05 \log_2(\frac{1}{0.05}) \times 4] \\
&= 1 \ + \ 0.821090 \ + \ 0.864386 \\
&= 2.685476
\end{aligned}
$$

The expected code length and the entropy of the source are similar

	A	B	C	D	E	F	G	H
A	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{80}$	$\frac{3}{80}$	$\frac{1}{80}$	$\frac{1}{80}$	$\frac{1}{80}$	$\frac{1}{80}$
B	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{80}$	$\frac{3}{80}$	$\frac{1}{80}$	$\frac{1}{80}$	$\frac{1}{80}$	$\frac{1}{80}$
C	$\frac{3}{80}$	$\frac{3}{80}$	$\frac{9}{400}$	$\frac{9}{400}$	$\frac{3}{400}$	$\frac{3}{400}$	$\frac{3}{400}$	$\frac{3}{400}$
D	$\frac{3}{80}$	$\frac{3}{80}$	$\frac{9}{400}$	$\frac{9}{400}$	$\frac{3}{400}$	$\frac{3}{400}$	$\frac{3}{400}$	$\frac{3}{400}$
E	$\frac{1}{80}$	$\frac{1}{80}$	$\frac{3}{400}$	$\frac{3}{400}$	$\frac{1}{400}$	$\frac{1}{400}$	$\frac{1}{400}$	$\frac{1}{400}$
F	$\frac{1}{80}$	$\frac{1}{80}$	$\frac{3}{400}$	$\frac{3}{400}$	$\frac{1}{400}$	$\frac{1}{400}$	$\frac{1}{400}$	$\frac{1}{400}$
G	$\frac{1}{80}$	$\frac{1}{80}$	$\frac{3}{400}$	$\frac{3}{400}$	$\frac{1}{400}$	$\frac{1}{400}$	$\frac{1}{400}$	$\frac{1}{400}$
H	$\frac{1}{80}$	$\frac{1}{80}$	$\frac{3}{400}$	$\frac{3}{400}$	$\frac{1}{400}$	$\frac{1}{400}$	$\frac{1}{400}$	$\frac{1}{400}$

2.(a) Design Huffman code

AA 0111	AB 1000	AC 11100	AD 11101	AE 000110	AF 000111	AG 000100	AH 000101
BA 1001	BB 1010	BC 11010	BD 11011	BE 001010	BF 001011	BG 001000	BH 001001
CA 11000	CB 11001	CC 00000	CD 00001	CE 0110010	CF 0110011	CG 0110000	CH 0110001
DA 10110	DB 10111	DC 111110	DD 111111	DE 0110110	DF 0110111	DG 0110100	DH 0110101
EA 001110	EB 001111	EC 0101010	ED 0101011	EE 111101010	EF 111101011	EG 111101000	EH 111101001
FA 001100	FB 001101	FC 0101000	FD 0101001	FE 111101110	FF 111101111	FG 111101100	FH 111101101
GA 010010	GB 010011	GC 0101110	GD 0101111	GE 111100010	GF 111100011	GG 111100000	GH 111100001
HA 010000	HB 010001	HC 0101100	HD 0101101	HE 111100110	HF 111100111	HG 111100100	HH 111100101

(b).

$$
\begin{aligned}
L_{exp} &= 4 * \frac{1}{16} * 4 + 5 * \frac{3}{80} * 8 + 5 * \frac{9}{400} * 2 + 6 * \frac{9}{400} * 2 + 6 * \frac{1}{80} * 16 + 7 * \frac{3}{400} * 16 + 9 * \frac{1}{400} * 16 \\
&= 0.0375 + 1.5 + 0.225 + 0.27 + 1.2 + 0.84 + 0.36 \\
&= 4.4325
\end{aligned}
$$

the length of the expected code is greater than the expected from S

(c).

$$\begin{aligned} H(S \times S) &= -4 * \frac{1}{16}\log_2(\frac{1}{16}) - 8 * \frac{3}{80}\log_2(\frac{3}{80}) - 4 * \frac{9}{400}\log_2(\frac{9}{400}) \\ &\quad - 16 * \frac{1}{80}\log_2(\frac{1}{80}) - 16 * \frac{3}{400}\log_2(\frac{3}{400}) - 16 * \frac{1}{400}\log_2(\frac{1}{400}) \\ &= 1 + 1.421090 + 0.492654 + 1.264386 + 0.847067 + 0.345754 \\ &= 5.370951 \end{aligned}$$

3.

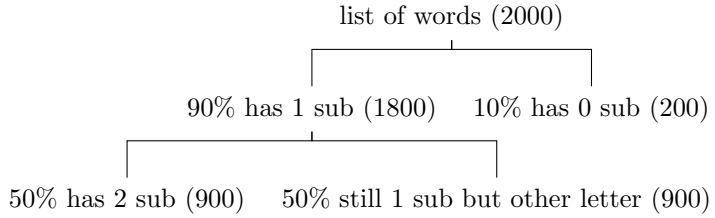
(i) see "huf_A1.txt"

(ii) see "huf_A2.txt"

when encoding A1.txt the huffman code using only the first paragraph created a smaller file size compared to the huffman code created from the whole document (see file size of "A1_encoded.txt" and "A2_encoded.txt")

Q4

1.P1



2000 + 1800 + 900 + 900 = 5600

where 2000 is the base amount of words, 1800 is the amount of words with 1 subbed in letter, 900 are ones with 2 subs and another 900 where the other letter is subbed in.

Since each of the selected words are unique the amount of words the list lowers from

2000 $\rightarrow$ 1999 ... $\rightarrow$ 1997 words.

Once calculated the changes of entropy are minor, therefore for the sake of simplicity we assume that there are 2000 words available for each of the four words selected.

$$
\begin{aligned}
totalpossiblecombination &= 5600^4 \\
&= 9.834496 \times 10^1 4 \\
H(X) &= \log_2(9.834496 \times 10^{14}) \\
&= 49.803298
\end{aligned}
$$

1.P2

$$
\begin{aligned}
permissible\ \ characters &= 26 + 26 + 10 + 4 \\
&= 66 \\
permissible\ \ password &= 66^8 \\
&= 3.600406 \times 10^{14} \\
H(X) &= \log_2(3.600406 \times 10^{14}) \\
&= 48.355152
\end{aligned}
$$

since each character is selected at random, an even distributions is assumed and the password length is 8 characters:

2. "More Bang for Your Buck" is smallest word phrase that has an entropy of 49.78 which is similar to the entropy calculated at P1 which is 49.80. The entropy for this phrase is estimated by the conclusion that each letter in the english language contains 2.62 bits of information.This in formation is found in pg 53 of [1]. Therefore 49.80/2.62 is 19 letters which is how many letters the phrase contains.

Q5

1.

$$
\begin{aligned}
p(M|T) &= p(T|M) \,/\, p(T|M) + p(T|F) \\
&= 0.30 \,/\, (0.30 + 0.20) \\
&= 0.60 \\
p(M|T) &= 60\%
\end{aligned}
$$

2. when **P is male**, since it holds more entropy than when P is female since male occurs at a lower frequency

$$
\begin{aligned}
P(male) &= -0.49 \log_2 0.49 \\
&= 0.504282 \\
P(female) &= -0.51 \log_2 0.51 \\
&= 0.495430
\end{aligned}
$$

3.

$$H(T|M) = p(T|M)\log_2(1/p(T|M))$$
$$= 0.3\log_2(1/0.3)$$
$$= 0.521090$$

$justH(T)$??

$$H(T) = p(T|M) + p(T|F)/p(T|M) + (1 - p(T|M)) + p(T|F) + (1 - p(T|F))$$
$$= 0.30 + 0.20/0.30 + 0.7 + 0.3 + 0.8$$
$$= 0.5/2$$
$$= 0.25$$
$$= 0.25\log_2(1/0.25) \qquad = 2H(M|T) = H(p(T|M)/p(T|M) + p$$
$$= H(0.30\ /\ (0.30 + 0.20))$$
$$= H(0.60)$$
$$= 0.60\log_2(1/0.6)$$
$$= 0.442179$$

We learn more information that a person is tall given that they are male, compared to that a person is male given that they are tall. This is due to the that p(T—M) occurs at a lower percentage.

4.

$$H(F|T) = H(p(T|F)\ /\ p(T|M) + p(T|F))$$
$$= H(0.20\ /\ (0.30 + 0.20))$$
$$= -0.4\log_2(0.4)$$
$$= 0.528771$$
$$H(M|T) = H(p(T|M)\ /\ p(T|M) + p(T|F))$$
$$= H(0.30\ /\ (0.30 + 0.20))$$
$$= -0.6\log_2(0.6)$$
$$= 0.442179$$

Q7

1.

$$
\begin{aligned}
H(X_1) &= -0.3\log_2(0.30) - 0.3\log_2(0.30) - 0.4\log_2(0.40) \\
&= 0.521090 + 0.521090 + 0.528771 \\
&= 1.570951 \\
H(X_2) &= -0.35\log_2(0.35) - 0.3\log_2(0.30) - 0.4\log_2(0.35) \\
&= 0.530101 + 0.521090 + 0.530101 \\
&= 1.581292 \\
H(X_1X_2) &= -0.2\ log_2(0.2) - 0.1\ log_2(0.1) - 3 \times 0.21\ log_2(0.15) - 0.25\ log_2(0.25) \\
&= 0.464386 + 0.332193 + 1.23163 + 0.5 \\
&= 2.528209 \\
H(X_2|X_1) &= \sum p_{(x_1x_2)} - log_2(p(X_2|X_1)) \\
&= -p_{x_1x_2}log_2(p(X_2 = a|X_1 = a)) - p_{x_1x_2}log_2(p(X_2 = a|X_1 = b)) - p_{x_1x_2}log_2(p(X_2 = a|X_1 = c)) \\
&\quad - p_{x_1x_2}log_2(p(X_2 = b|X_1 = a)) - p_{x_1x_2}log_2(p(X_2 = b|X_1 = b)) - p_{x_1x_2}log_2(p(X_2 = b|X_1 = c)) \\
&\quad - p_{x_1x_2}log_2(p(X_2 = c|X_1 = a)) - p_{x_1x_2}log_2(p(X_2 = c|X_1 = b)) - p_{x_1x_2}log_2(p(X_2 = c|X_1 = c)) \\
&= -\frac{2}{3}\ log_2(\frac{2}{3}) - \frac{1}{2}\ log_2(\frac{1}{2}) - 0\ log_2(0) \\
&\quad - 0\ log_2(0) - \frac{1}{2}\ log_2(\frac{1}{2}) - \frac{3}{8}\ log_2(\frac{3}{8}) \\
&\quad - \frac{1}{3}\ log_2(\frac{1}{3}) - 0\ log_2(0) - \frac{5}{8}\ log_2(\frac{5}{8}) \\
&= -\frac{2}{3}\ log_2(\frac{2}{3}) - \frac{1}{2}\ log_2(\frac{1}{2}) - \frac{1}{2}\ log_2(\frac{1}{2}) - \frac{3}{8}\ log_2(\frac{3}{8}) - \frac{1}{3}\ log_2(\frac{1}{3}) - \frac{5}{8}\ log_2(\frac{5}{8}) \\
&= 0.389975 + 0.5 + 0.5 + 0.530639 + 0.528320 + 0.423795 \\
&= 2.872729
\end{aligned}
$$

$$
\begin{aligned}
H(X_1|X_2) &= \sum p_{(x_1x_2)} - log_2(p(X_1|X_2)) \\
&= -p_{x_1x_2}log_2(p(X_1 = a|X_2 = a)) - p_{x_1x_2}log_2(p(X_1 = a|X_2 = b)) - p_{x_1x_2}log_2(p(X_1 = a|X_2 = c)) \\
&\quad - p_{x_1x_2}log_2(p(X_1 = b|X_2 = a)) - p_{x_1x_2}log_2(p(X_1 = b|X_2 = b)) - p_{x_1x_2}log_2(p(X_1 = b|X_2 = c)) \\
&\quad - p_{x_1x_2}log_2(p(X_1 = c|X_2 = a)) - p_{x_1x_2}log_2(p(X_1 = c|X_2 = b)) - p_{x_1x_2}log_2(p(X_1 = c|X_2 = c))
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{4}{7}log_2(\frac{4}{7}) - 0log_2(0) - \frac{2}{7}log_2(\frac{2}{7}) \\
&\quad - \frac{3}{7}log_2(\frac{3}{7}) - \frac{1}{2}log_2(\frac{1}{2}) - 0log_2(0) \\
&\quad - 0log_2(0) - \frac{1}{2}log_2(\frac{1}{2}) - \frac{5}{7}log_2(\frac{5}{7}) \\
&= -\frac{4}{7}log_2(\frac{4}{7}) - \frac{2}{7}log_2(\frac{2}{7}) - \frac{3}{7}log_2(\frac{3}{7}) - \frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) - \frac{5}{7}log_2(\frac{5}{7}) \\
&= 0.461345 + 0.516387 + 0.523882 + 0.5 + 0.5 + 0.346733 \\
&= 2.848347
\end{aligned}
$$

2.

$$
\begin{aligned}
p(aaa) &= 0.3 + \frac{2}{3} + \frac{2}{3} = \frac{2}{15} \\
p(aac) &= 0.3 + \frac{2}{3} + \frac{1}{3} = \frac{1}{15} \\
p(acb) &= 0.3 + \frac{2}{3} + 0.375 = \frac{3}{80} \\
p(acc) &= 0.3 + \frac{2}{3} + 0.625 = \frac{1}{16}
\end{aligned}
$$

$$
\begin{aligned}
p(baa) &= 0.3 + \frac{1}{2} + \frac{2}{3} = \frac{1}{10} \\
p(bac) &= 0.3 + \frac{1}{2} + \frac{1}{3} = \frac{1}{20} \\
p(bba) &= 0.3 + \frac{1}{2} + \frac{1}{2} = \frac{3}{40} \\
p(bbb) &= 0.3 + \frac{1}{2} + \frac{1}{2} = \frac{3}{40}
\end{aligned}
$$

$$
\begin{aligned}
p(cba) &= 0.4 + 0.375 + \frac{1}{2} = \frac{3}{40} \\
p(cbb) &= 0.4 + 0.375 + \frac{1}{2} = \frac{3}{40} \\
p(ccb) &= 0.4 + 0.625 + 0.375 = \frac{3}{32} \\
p(ccc) &= 0.4 + 0.625 + 0.625 = \frac{5}{32}
\end{aligned}
$$

$$
\begin{aligned}
H(S_1, S_2, S_3) &= -\frac{2}{15}\, log_2(\frac{2}{15}) - \frac{1}{15}\, log_2(\frac{1}{15}) - \frac{3}{80}\, log_2(\frac{3}{80}) - \frac{1}{16}\, log_2(\frac{1}{16}) \\
&\quad - \frac{1}{10}\, log_2(\frac{1}{10}) - \frac{1}{20}\, log_2(\frac{1}{20}) - \frac{3}{40}\, log_2(\frac{3}{40}) - \frac{3}{40}\, log_2(\frac{3}{40}) \\
&\quad - \frac{3}{40}\, log_2(\frac{3}{40}) - \frac{3}{40}\, log_2(\frac{3}{40}) - \frac{3}{32}\, log_2(\frac{3}{32}) - \frac{5}{32}\, log_2(\frac{5}{32}) \\
&= 0.387585 + 0.260459 + 0.177636 + 0.25 + 0.332192 + 0.216096 + (4 * 0.280272) + 0.320159 + 0.418448 \\
&= 3.483663
\end{aligned}
$$

References

[1] C.E. Shannon, Prediction and Entropy of Printed English, The Bell System Technical Journal, January1951.

[2] Claude Shannon Demonstrates Machine Learning (03/16/2010) [Video] retrieved from: http://techchannel.att.com/play-video.cfm/2010/3/16/In-Their-Own-Words-Claude-Shannon-Demonstrates-Machine-Learning