情報理論レポート１(情報量)解答例

提示：2010/10/6(水)　提出：2010/10/20(水)

１．英文の情報量

26 文字からなるアルファベット $\mathcal{A}=\{a,b,\cdots,z\}$ の各記号 $\alpha\in\mathcal{A}$ の出現確率 $P(\alpha)$ は、次の表で与えられる。（http://www7.plala.or.jp/dvorakjp/hinshutu.htm 参照。）

α	e	a	t	i	o	s	n	r
$P(\alpha)$	0.1183	0.0876	0.0849	0.0740	0.0731	0.0723	0.0702	0.0645
α	h	l	d	c	u	m	p	f
$P(\alpha)$	0.0441	0.0402	0.0402	0.0331	0.0306	0.0277	0.0210	0.0206
α	g	y	w	b	v	k	j	x
$P(\alpha)$	0.0189	0.0186	0.0176	0.0171	0.0104	0.0093	0.0021	0.0018
α	q	z						
$P(\alpha)$	0.0009	0.0009						

この出現確率 $P(\alpha)$ 基づいた情報源を

$$A=\begin{Bmatrix} a & , & b & , & \cdots & , & z \\ P(a) & , & P(b) & , & \cdots & , & P(z) \end{Bmatrix}$$

とし、
アルファベット中の各記号 $\alpha\in\mathcal{A}$ が均等に現れる情報源を

$$B=\begin{Bmatrix} a & , & b & , & \cdots & , & z \\ \frac{1}{26} & , & \frac{1}{26} & , & \cdots & , & \frac{1}{26} \end{Bmatrix}$$

とする。

また、単語 w が情報源 A から生成される確率を $P_A\left(w\right)$、自己情報量を $I_A(w)$ と表し、情報源 B から生成されたときの確率を $P_B\left(w\right)$、自己情報量を $I_B(w)$ と表す。

このとき次の問いに答えよ。

(1) $w_1 = \inf ormation$、$w_2 = theory$、$w_3 = shannon$ とし、次の確率および自己情報量を有効数字 3 桁で求めよ。

$$
\begin{array}{l}
P_A\left(w_1\right), I_A\left(w_1\right), P_B\left(w_1\right), I_B\left(w_1\right), \\
P_A\left(w_2\right), I_A\left(w_2\right), P_B\left(w_2\right), I_B\left(w_2\right), \\
P_A\left(w_3\right), I_A\left(w_3\right), P_B\left(w_3\right), I_B\left(w_3\right)
\end{array}
$$

（解答）

次のように計算される。

$$
\begin{aligned}
& P_A\left(w_1\right) = \prod_{\alpha \in w_1} P_A\left(\alpha\right) \\
& = P_A\left(i\right) \times P_A\left(n\right) \times P_A\left(f\right) \times P_A\left(o\right) \times P_A\left(r\right) \\
& \quad \times P_A\left(m\right) \times P_A\left(a\right) \times P_A\left(t\right) \times P_A\left(i\right) \times P_A\left(o\right) \times P_A\left(n\right) \\
& \simeq 0.0740 \times 0.0702 \times 0.0206 \times 0.0731 \times 0.0645 \\
& \quad \times 0.0277 \times 0.0876 \times 0.0849 \times 0.0740 \times 0.0731 \times 0.0702 \\
& \simeq 3.95 \times 10^{-14}
\end{aligned}
$$

$$
\begin{aligned}
& I_A\left(w_1\right) = -\log_2 P_A\left(w_1\right) \\
& = \sum_{\alpha \in w_1} i_A\left(\alpha\right) \\
& = i_A\left(i\right) + i_A\left(n\right) + i_A\left(f\right) + i_A\left(o\right) + i_A\left(r\right) \\
& \quad + i_A\left(m\right) + i_A\left(a\right) + i_A\left(t\right) + i_A\left(i\right) + i_A\left(o\right) + i_A\left(n\right) \\
& \simeq -\log 0.0740 - \log 0.0702 - \log 0.0206 - \log 0.0731 - \log 0.0645 \\
& \quad - \log 0.0277 - \log 0.0876 - \log 0.0849 - \log 0.0740 - \log 0.0731 - \log 0.0702 \\
& \simeq 3.76 + 3.83 + 5.60 + 3.77 + 3.95 \\
& \quad + 5.17 + 3.51 + 3.56 + 3.76 + 3.77 + 3.83 \\
& \simeq 44.5 \ [bit]
\end{aligned}
$$

情報源 B では、各記号が均等に現れるので、次式が成り立つ。

$$P_B(a) = P_B(b) = \cdots = P_B(z) = \frac{1}{26} \simeq 0.0038$$

$$\begin{aligned}
&i_B(a) = i_B(b) = \cdots = i_B(z) \\
&= -\log P(\alpha) \\
&= -\log \frac{1}{26} \\
&\simeq 4.70\ [bit]
\end{aligned}$$

よって、情報源 B からの文字列出現確率や自己情報量は文字数だけで定まる。

$$\begin{aligned}
&P_B\left(w_1\right) = \prod_{\alpha \in w_1} P_B\left(\alpha\right) \\
&= P_B\left(i\right) \times P_B\left(n\right) \times P_B\left(f\right) \times P_B\left(o\right) \times P_B\left(r\right) \\
&\quad \times P_B\left(m\right) \times P_B\left(a\right) \times P_B\left(t\right) \times P_B\left(i\right) \times P_B\left(o\right) \times P_B\left(n\right) \\
&= \frac{1}{26^{11}} \\
&\simeq 2.72 \times 10^{-16}
\end{aligned}$$

$$\begin{aligned}
&I_B\left(w_1\right) = \sum_{\alpha \in w_1} i_B\left(\alpha\right) \\
&= 11 \times i_B\left(\alpha\right) \\
&= 11 \times \log 26 \\
&\simeq 51.7\ [bit]
\end{aligned}$$

$$\begin{aligned}
&P_A\left(w_2\right) = P_A\left(t\right) \times P_A\left(h\right) \times P_A\left(e\right) \times P_A\left(o\right) \times P_A\left(r\right) \times P_A\left(y\right) \\
&\simeq 3.88 \times 10^{-8}
\end{aligned}$$

$$\begin{aligned}
&I_A\left(w_2\right) = i_A\left(t\right) + i_A\left(h\right) + i_A\left(e\right) + i_A\left(o\right) + i_A\left(r\right) + i_A\left(y\right) \\
&\simeq 3.56 + 4.50 + 3.08 + 3.77 + 3.95 + 5.75 \\
&\simeq 24.6\ [bit]
\end{aligned}$$

$$P_B\left(w_2\right) = \frac{1}{26^6} \simeq 3.88 \times 10^{-8}$$

$$I_B\left(w_2\right) = 6 \times i_B\left(\alpha\right) = 6 \times \log 26 \simeq 28.2\ [bit]$$

$$P_{A}\left(w_{3}\right) = P_{A}\left(s\right) \times P_{A}\left(h\right) \times P_{A}\left(a\right) \times P_{A}\left(n\right) \times P_{A}\left(n\right) \times P_{A}\left(o\right) \times P_{A}\left(n\right)$$
$$\simeq 7.06 \times 10^{-9}$$

$$I_{A}\left(w_{3}\right) = i_{A}\left(s\right) + i_{A}\left(h\right) + i_{A}\left(a\right) + i_{A}\left(n\right) + i_{A}\left(n\right) + i_{A}\left(o\right) + i_{A}\left(n\right)$$
$$\simeq 3.79 + 4.50 + 3.51 + 3.83 + 3.83 + 3.77 + 3.83$$
$$\simeq 27.1\,[bit]$$

$$P_{B}\left(w_{3}\right) = \frac{1}{26^{7}} \simeq 7.06 \times 10^{-9}$$

$$I_{B}\left(w_{3}\right) = 7 \times i_{B}\left(\alpha\right) = 7 \times \log 26 \simeq 32.9\,[bit]$$

これらの結果より、「記号の出現確率が不均等な場合、記号列（記号数および記号の並び方）が同じでも出現確率および自己情報量は異なる。」ことが分かる。

(2)情報源 A のエントロピー（平均情報量）$H(A)$ および情報源 B のエントロピー $H(B)$

をそれぞれ求めよ。

情報源 A に関しては次の通り。

$$
\begin{aligned}
&H(A) = -\sum_{\alpha \in A} P(\alpha) i_{A}(\alpha) \\
&= -\sum_{\beta \in B} P(\alpha) \log P(\alpha) \\
&= -\Big(0.1183\Big)\log\Big(0.1183\Big) - \Big(0.0876\Big)\log\Big(0.0876\Big) - \Big(0.0849\Big)\log\Big(0.0849\Big) - \Big(0.0740\Big)\log\Big(0.0740\Big) \\
&-\Big(0.0731\Big)\log\Big(0.0731\Big) - \Big(0.0723\Big)\log\Big(0.0723\Big) - \Big(0.0702\Big)\log\Big(0.0702\Big) - \Big(0.0645\Big)\log\Big(0.0645\Big) \\
&-\Big(0.0441\Big)\log\Big(0.0441\Big) - \Big(0.0402\Big)\log\Big(0.0402\Big) - \Big(0.0402\Big)\log\Big(0.0402\Big) - \Big(0.0331\Big)\log\Big(0.0331\Big) \\
&-\Big(0.0306\Big)\log\Big(0.0306\Big) - \Big(0.0277\Big)\log\Big(0.0277\Big) - \Big(0.0210\Big)\log\Big(0.0210\Big) - \Big(0.0206\Big)\log\Big(0.0206\Big) \\
&-\Big(0.0189\Big)\log\Big(0.0189\Big) - \Big(0.0186\Big)\log\Big(0.0186\Big) - \Big(0.0176\Big)\log\Big(0.0176\Big) - \Big(0.0171\Big)\log\Big(0.0171\Big) \\
&-\Big(0.0104\Big)\log\Big(0.0104\Big) - \Big(0.0093\Big)\log\Big(0.0093\Big) - \Big(0.0021\Big)\log\Big(0.0021\Big) - \Big(0.0018\Big)\log\Big(0.0018\Big) \\
&-\Big(0.0009\Big)\log\Big(0.0009\Big) - \Big(0.0009\Big)\log\Big(0.0009\Big) \\
&\simeq 4.19 \quad [bit \,/\, 記号]
\end{aligned}
$$

情報源 B に関しては次の通り。

$$
\begin{aligned}
&H(B) = -\sum_{\beta \in B} P(\beta) i_{B}(\beta) \\
&= -\sum_{\beta \in B} P(\beta) \log P(\beta) \\
&= -26 \times \frac{1}{26} \log \frac{1}{26} \\
&= \log 26 \\
&\simeq 4.70 \,[bit \,/\, 記号\ ]
\end{aligned}
$$

なお、情報源記号が同じ情報源に対して、エントロピーは、情報源記号の生成確率が一定の場合に最大となる。