情報理論

情報量

英文字

アルファベット２６文字，空白，カンマ，ピリオド，
アポストロフィ　　計３０文字と想定しよう.

３０文字を用いて，

１００字の文章を書く.

生成される文章の種類

３０の１００乗通り　　$N_{100} = 30^{100}$

２００字の文章を書く.

３０の２００乗通り　　$N_{200} = 30^{200}$

文字数が２倍になれば，情報量は２倍と考えるのが妥当.

指数部の大きさを評価する　→　対数　ｌｏｇ

２を底とした対数

$\log_2(2^n) = n$

$\log_2(m^n) = n \times \log_2(m)$

$\log_2(10) = 3.3219$

$\Leftrightarrow \quad 2^{3.3219} = 10$

$2^3 \leqq 10 \leqq 2^4$

$8 \qquad\qquad 16$

底を２とした対数　$\log_2$ により情報量を測る.

N_{100} （＝ 30^{100}） の情報量

$\log_2(N_{100})$ ＝ １００ × $\log_2$（３０）

N_{200} （＝ 30^{200}） の情報量

$\log_2(N_{200})$ ＝ ２００ × $\log_2$（３０）

情報量の比較

$\log_2(N_{200})$ ＝ ２ × $\log_2(N_{100})$

ちゃんと２倍になっている.

３ビット（２進数３桁）データの情報量

２^{3}＝８通りの表現

０００　　００１　　０１０

０１１

１００　　１０１　　１１０

１１１

$log_{2}（２^{3}）$　＝　３

底を２とした対数がビット数に対応

ビット単位の情報量，ビットは情報量

log_{2}（表現の種類の数）

Ｎ 通りの表現が持つ情報量　log_{2}（Ｎ）

英文字（３０文字）の情報量

log_{2}（３０）　＝　４．９０７　ビット bits

確率と情報量

サイコロ：

６通りの表現　（Ｎ＝６）

情報量　　$\log_2(6)$ ＝ ２．５８５ bits

ある目の出る確率　　ｐ ＝ １／Ｎ ＝ １／６

逆に，Ｎ ＝ １／ｐ

確率 ｐ の事象が知れたことによる情報量は，

$\log_2(N)$ ＝ $\log_2(1/p)$

サイコロの，ある目がでたときの情報量

$\log_2(1/p) = \log_2(1/(1/6))$

$= \log_2(6)$

$= 2.585$ bits

対数の公式　$\log(1/x) = -\log(x)$

より

$\log_2(1/p) = -\log_2(p)$

確率　p　の事象の情報量

$\log_2(1/p)$，　$-\log_2(p)$

両者は等価

おみくじ：吉の方角

東（１／２），西（１／４），
南（１／８），北（１／８）

東と知れたことによる情報量
　　$-\log_2$（１／２）　＝　１ bit
西と知れたことによる情報量
　　$-\log_2$（１／４）　＝　２ bits
南と知れたことによる情報量
　　$-\log_2$（１／８）　＝　３ bits

確率が低いほど，珍しいことほど，
それが起こると情報量が多い．

情報エントロピ

全部で ｎ 個の事象

ｉ 番目の事象の起こる確率　ｐｉ
ｉ 番目の事象が起きたと知れたことによる情報量
　　－ ｌｏｇ$_2$（ｐｉ）
ｉ 番目の事象によりもたらされると期待される情報量
　　ｐｉ ×（－ ｌｏｇ$_2$（ｐｉ））

ｎ 個全部の事象により期待される情報量 Ｈ
　　（情報エントロピー）
　　ｐ1 ×（－ ｌｏｇ$_2$（ｐ1））
　　＋ ｐ2 ×（－ ｌｏｇ$_2$（ｐ2））
　　＋ ・ ・ ・ ＋ ｐｎ ×（－ ｌｏｇ$_2$（ｐｎ））

方角おみくじの情報エントロピ Ｈ

４個の事象

東（１／２），　西（１／４），

南（１／８），　北（１／８）

$$
\begin{aligned}
H &= (1/2) \times (-\log_2 (1/2)) \\
&\quad + (1/4) \times (-\log_2 (1/4)) \\
&\quad + (1/8) \times (-\log_2 (1/8)) \\
&\quad + (1/8) \times (-\log_2 (1/8)) \\
&= (1/2) + (1/2) \\
&\quad + (3/8) + (3/8) \\
&= 1.75 \text{ bits}
\end{aligned}
$$

ｎ 個の事象で，情報エントロピが最大になるとき，

$$pi = p = 1/n$$

$$H = -\log_2(p)$$

（例）

情報エントロピが最大となる方角おみくじ

（ｎ＝４）

各方角への確率が １／４

$$H = -\log_2(1/4)$$

$$= \log_2(4)$$

$$= 2$$

ｎ ＝ ８ であれば，ｐ ＝１／８ であり，Ｈ ＝ ３．

n個の事象で，ひとつの事象の確率のみが １ で，
他の事象の確率がゼロであるとき

$$p1 = 1, \quad p2 = \cdots = pn = 0$$

$$H = -1 \times \log_2(1) = -1 \times 0 = 0$$

事象間で，確率の偏りがないほど，
情報エントロピは大きくなる.

伝達する情報の符号

符号化

文字	東	西	南	北
符号系A	０	１	０１	１１
符号系B	０	１０	１１０	１１１

送信情報　「 東西南北 」

符号系A　　０１０１１１

東西南北？　東西東北西？　南東西西西？

符号系B　　０１０１１０１１１

東西南北

一意解読可能符号系　　情報が一意に復元できる.

瞬時解読可能符号系　　順次先頭から復元できる.

一意解読可能，瞬時解読可能符号系を与える
ハフマンの符号化　（アルゴリズム）

(1)　発生確率順に事象を並べる.

(2)　小さい発生確率を持つ事象を ２ つ選び，１ つに ０，
もう １ つに １ を割り付ける.

(3)　ステップ (2) における ２ つの事象の確率の和を発生
確率とする事象を生成し，その生成した事象を，和を
取る前の事象に代わる新たな事象として加え，あらた
めて全事象を発生確率順に並べる.

(4)　事象が最後の １ つになるまで，(2)，(3) を繰り返す.

東西南北 （Ｅ，Ｗ，Ｓ，Ｎ） の符号系 （その１）

Ｅ (1/2)　　1　← 最後の事象

＃(1/2)　0

Ｗ (1/4)　　1

＊(1/4)　0

Ｓ (1/8)　1

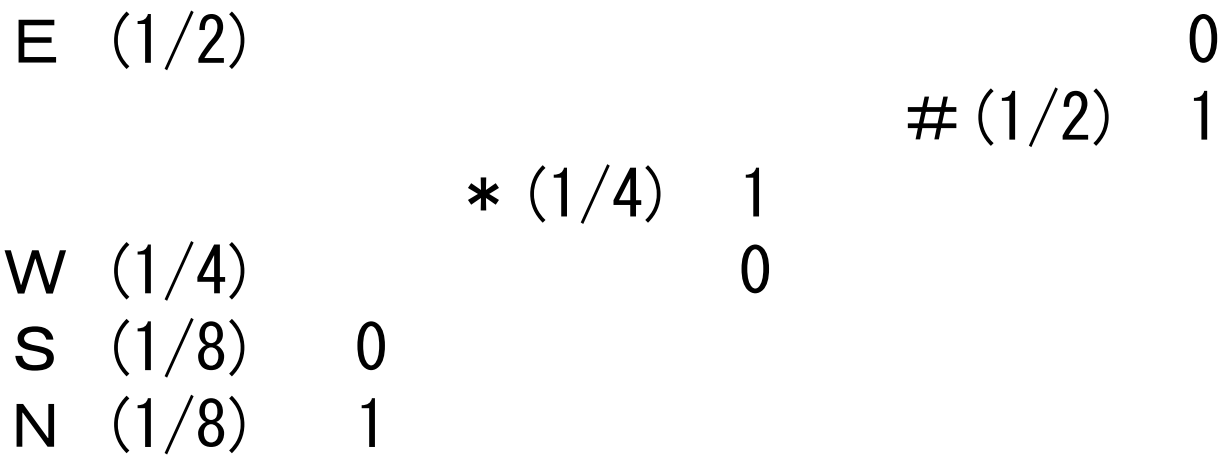
Ｎ (1/8)　0

符号系

東 (E)	１	西 (W)	０１
南 (S)	００１	北 (N)	０００

任意性：　同確率のときに，上下どちらに配置するか

東西南北 （Ｅ，Ｗ，Ｓ，Ｎ） の符号系 （その２）



符号系

東　０	西　１０
南　１１０	北　１１１

天気 （晴， 曇， 雨， 嵐， 雪， 霧） の符号化

(0.6) 1
晴れ (0.4) 0
曇 (0.35) 1
(0.25) 0
雨 (0.16) 1
(0.09) 0
(0.05) 1
嵐 (0.04) 0
雪 (0.03) 1
霧 (0.02) 0

天気の符号系

晴 ０　曇 １１　雨 １０１　嵐 １０００
雪 １００１１　霧 １００１０

情報量の比較

天気の符号系の平均情報量（平均符号長）

晴 (0.4)　１ bit,　　曇 (0.35)　２ bits,

雨 (0.16)　３ bits,　　嵐 (0.04)　４ bits,

雪 (0.03)　５ bits,　　霧 (0.02)　５ bits

$0.4 \times 1 + 0.35 \times 2 + 0.16 \times 3$

$+ 0.04 \times 4 + 0.03 \times 5 + 0.02 \times 5$

$= 1.99$ bits

天気の情報エントロピ　　１．９３２ bits

６面サイコロが持つ情報エントロピ

$-\log_2(1/6) = 2.585$ bits

ハフマンの符号化

- 確率が高いほど短いコードは短い.
- 平均符号長が最短の瞬時解読可能符号系