データサイエンス実践B
2022/7/4 講義資料

H. Kumano

情報の基本単位

コンピュータは、一度に 0 か 1 で表される、2通りの情報しか判断できない

選択肢が2つであれば、一回で判断可能。
例：旅行に行くことにしました。国内/海外どちらにするか？

「海外にしますか？」という1つの質問に対して、「はい」か「いいえ」どちらかを選べば良い。

1：海外	0：国内

情報の基本単位

では、選択肢が4つある場合は？

北欧	北海道
地中海・エーゲ海	沖縄

一度の質問では決められない。が、こういう場合、何回か続けて質問すれば決められる。

1：海外にしますか？　はい	0：海外にしますか？　いいえ
北欧	北海道
地中海・エーゲ海	沖縄

一度回答すれば、選択肢が2つに減る。
続けての質問： 北にしますか？

→ 一つに決まる

情報の基本単位

今度は、選択肢が8つの場合はどうでしょう？

いくつの質問で、1つに絞れるでしょうか？

1つ目の質問： 8 → 4
2つ目の質問： 4 → 2
3つ目の質問： 2 → 1

1024個選択肢がある場合は？

1024 → 512 → 256 → 128 → 64 → 32 → 16 → 8 → 4 → 2 → 1

10回の質問で決められる。
（→ の数を数えればよい）

情報の基本単位

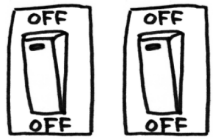
コンピュータでは、計算を
0（電気が通っていない状態）
1（電気が取っている状態）
の組み合わせで行っている。
0と1以外は判別できない。

と

1：海外	0：国内

2個の情報を表すにはスイッチが1個あればよい。

北欧	北海道
地中海・エーゲ海	沖縄



4個の情報を表すにはスイッチが2個あればよい。

情報の基本単位

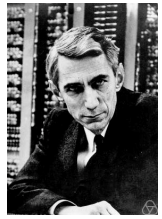
同様に、1024個の情報を表すには10個スイッチがあればよい。

$$2^1 = 2 \quad 2^2 = 4 \quad 2^{10} = 1024$$

= 情報の基本単位

2通りの情報が表せるもの
binary dig**it** (→ bit)

情報を数値化
符号化



Claude E. Shannon

計算機上では、すべての情報が（文字も画像も）符号化される

文字コード

文字コードとは、文字を表す番号のこと。
コンピューターは数字しか処理できません。よってコンピューターは、それぞれの文字の形（イメージ）に対応付けられた番号の一覧表で文字を管理しています。それが文字コードです。

文字コードはいくつか種類があり、それにより見た目が同じでも番号が異なります。

主な文字コード(1)

ASCII
もっとも基礎的な文字コード。半角英数字128文字から構成され、全ての文字を1バイトで表す。
例えば「A」はASCIIでは0x41(0xは16進数を表す)。

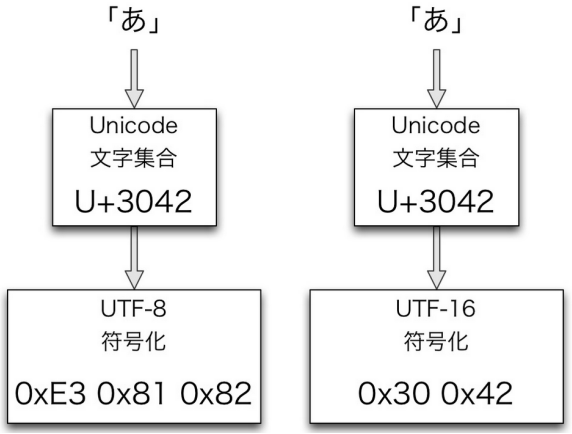
Shift_JIS
日本語を表すために多く用いられていた文字コード。
全ての文字を2バイトで表す。亜種の cp932 が Windowsで採用されていたことで広く使われていた。

主な文字コード(2)

UTF-8
現在最も広く使われている標準的な符号化形式。
UnicodeにASCIIコードを含めた規格で、全ての文字を1〜4バイトで表す。世界中の文字を扱えるため、標準的に使われるようになった。

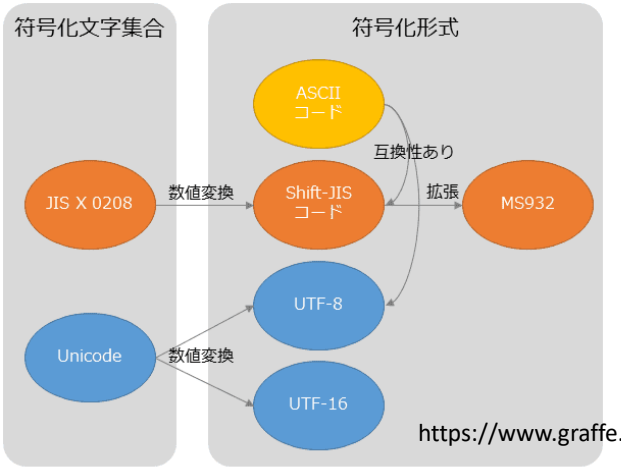
現在、Internet上で使用されている主な文字コードは、JISコード, SJIS(Shift-JIS)コード, EUC, UTF-8 の4つ。

文字集合と符号化形式



https://blog.shibayu36.org/entry/2015/09/14/102100

文字集合と符号化形式



Unicodeという「文字集合」に対し、UTF-8、UTF-16等の「符号化形式」がある。
同様に、JIS X 0208という「文字集合」に対し、ISO-2022-JP、EUC-JP、Shift_JIS等の「符号化形式」がある。

ASCIIコード表

American **S**tandard **C**ode for **I**nformation Interchange (ASCII)

ASCII = 7ビットコード
情報を表すのに7桁の2進数を用いる。
(10進数では0〜127)

0000000
から
1111111
までの128文字から成る「文字集合」

1バイト＝8ビット

文字	10進	8進	16進	文字	10進	8進	16進	文字	10進	8進	16進	文字	10進	8進	16進
(nul)	0	0000	0x00	(sp)	32	0040	0x20	@	64	0100	0x40	`	96	0140	0x60
(soh)	1	0001	0x01	!	33	0041	0x21	A	65	0101	0x41	a	97	0141	0x61
(stx)	2	0002	0x02	"	34	0042	0x22	B	66	0102	0x42	b	98	0142	0x62
(etx)	3	0003	0x03	#	35	0043	0x23	C	67	0103	0x43	c	99	0143	0x63
(eot)	4	0004	0x04	$	36	0044	0x24	D	68	0104	0x44	d	100	0144	0x64
(enq)	5	0005	0x05	%	37	0045	0x25	E	69	0105	0x45	e	101	0145	0x65
(ack)	6	0006	0x06	&	38	0046	0x26	F	70	0106	0x46	f	102	0146	0x66
(bel)	7	0007	0x07	'	39	0047	0x27	G	71	0107	0x47	g	103	0147	0x67
(bs)	8	0010	0x08	(	40	0050	0x28	H	72	0110	0x48	h	104	0150	0x68
(ht)	9	0011	0x09	)	41	0051	0x29	I	73	0111	0x49	i	105	0151	0x69
(nl)	10	0012	0x0a	*	42	0052	0x2a	J	74	0112	0x4a	j	106	0152	0x6a
(vt)	11	0013	0x0b	+	43	0053	0x2b	K	75	0113	0x4b	k	107	0153	0x6b
(np)	12	0014	0x0c	,	44	0054	0x2c	L	76	0114	0x4c	l	108	0154	0x6c
(cr)	13	0015	0x0d	-	45	0055	0x2d	M	77	0115	0x4d	m	109	0155	0x6d
(so)	14	0016	0x0e	.	46	0056	0x2e	N	78	0116	0x4e	n	110	0156	0x6e
(si)	15	0017	0x0f	/	47	0057	0x2f	O	79	0117	0x4f	o	111	0157	0x6f
(dle)	16	0020	0x10	0	48	0060	0x30	P	80	0120	0x50	p	112	0160	0x70
(dc1)	17	0021	0x11	1	49	0061	0x31	Q	81	0121	0x51	q	113	0161	0x71
(dc2)	18	0022	0x12	2	50	0062	0x32	R	82	0122	0x52	r	114	0162	0x72
(dc3)	19	0023	0x13	3	51	0063	0x33	S	83	0123	0x53	s	115	0163	0x73
(dc4)	20	0024	0x14	4	52	0064	0x34	T	84	0124	0x54	t	116	0164	0x74
(nak)	21	0025	0x15	5	53	0065	0x35	U	85	0125	0x55	u	117	0165	0x75
(syn)	22	0026	0x16	6	54	0066	0x36	V	86	0126	0x56	v	118	0166	0x76
(etb)	23	0027	0x17	7	55	0067	0x37	W	87	0127	0x57	w	119	0167	0x77
(can)	24	0030	0x18	8	56	0070	0x38	X	88	0130	0x58	x	120	0170	0x78
(em)	25	0031	0x19	9	57	0071	0x39	Y	89	0131	0x59	y	121	0171	0x79
(sub)	26	0032	0x1a	:	58	0072	0x3a	Z	90	0132	0x5a	z	122	0172	0x7a
(esc)	27	0033	0x1b	;	59	0073	0x3b	[	91	0133	0x5b	{	123	0173	0x7b
(fs)	28	0034	0x1c	<	60	0074	0x3c	\	92	0134	0x5c	|	124	0174	0x7c
(gs)	29	0035	0x1d	=	61	0075	0x3d	]	93	0135	0x5d	}	125	0175	0x7d
(rs)	30	0036	0x1e	>	62	0076	0x3e	^	94	0136	0x5e	~	126	0176	0x7e
(us)	31	0037	0x1f	?	63	0077	0x3f	_	95	0137	0x5f	(del)	127	0177	0x7f

各言語圏の文字コード

地域	1960 年代	1970 年代	1980 年代	1990 年代～
ラテン文字圏	ASCII/ISO 646	欧州言語への拡張(6937, 8859)		
キリル文字圏	GOST 13052	露語以外の諸語への拡張		
アラビア文字圏			ASMO 449	
ヘブライ文字圏			ECMA 121	
日本	JIS C 6220	JIS C 6226		
中国			GB 2312	少数民族文字
韓国			KS C 5601	KS X 1005
タイ			TIS 620	
インド			ISSCII 83	IS 13194
ベトナム				TCVN 5412
スリランカ				SLS 1134
国際符号化文字集合				ISO/IEC 10646

数字の表現（進数）

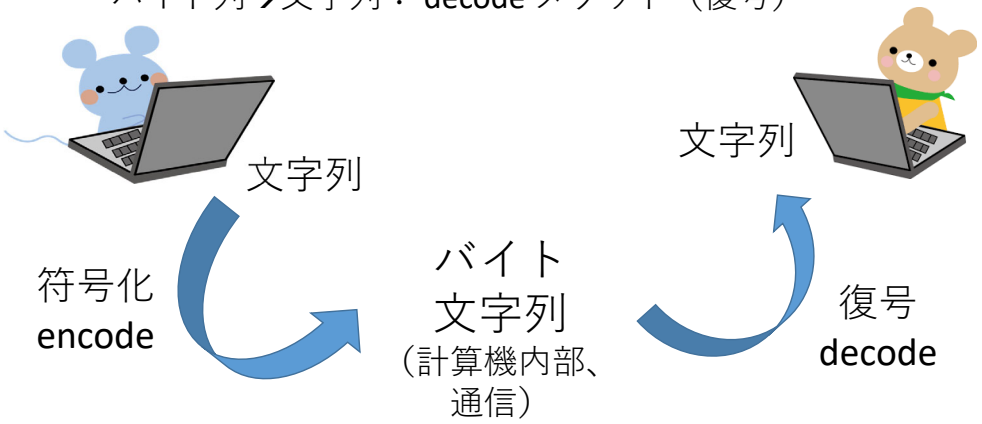
2進数	10進数	16進数
0001	1	1
0010	2	2
0001 0000	16	10
0110 0100	100	64
0001 0000 0000	256	100
1 0000 0000 0000 0000	65536	10000

16進数世界の文字
0,1,2,3,4,5,6,7,8,9,a,b,c,d,e,f

文字列とバイト列の変換

文字列とバイト列との変換用関数

文字列→バイト列： encode メソッド（符号化）
バイト列→文字列： decode メソッド（復号）



文字集合の例

日本語

文字集合	含まれる文字
JIS X 0211	JIS制御コード
JIS X 0201	JISローマ字 (ASCII)
JIS X 0201	JISカナ（半角カナ）
JIS X 0208	JIS漢字
JIS X 0212	JIS補助漢字

Unicode
ユニコード・コンソーシアムによって制定された、世界中の文字を表現しようとする世界統一文字集合規格。