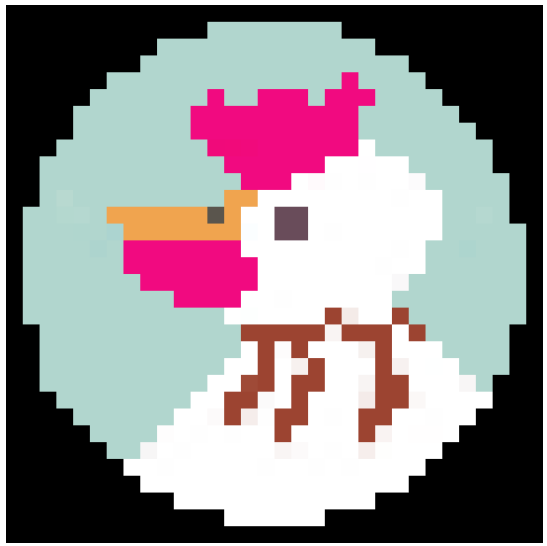
Base常用漢字encoding with

geNeraLized Base-N encoding

@yoshihiro503

NL名古屋

About yoshihiro503



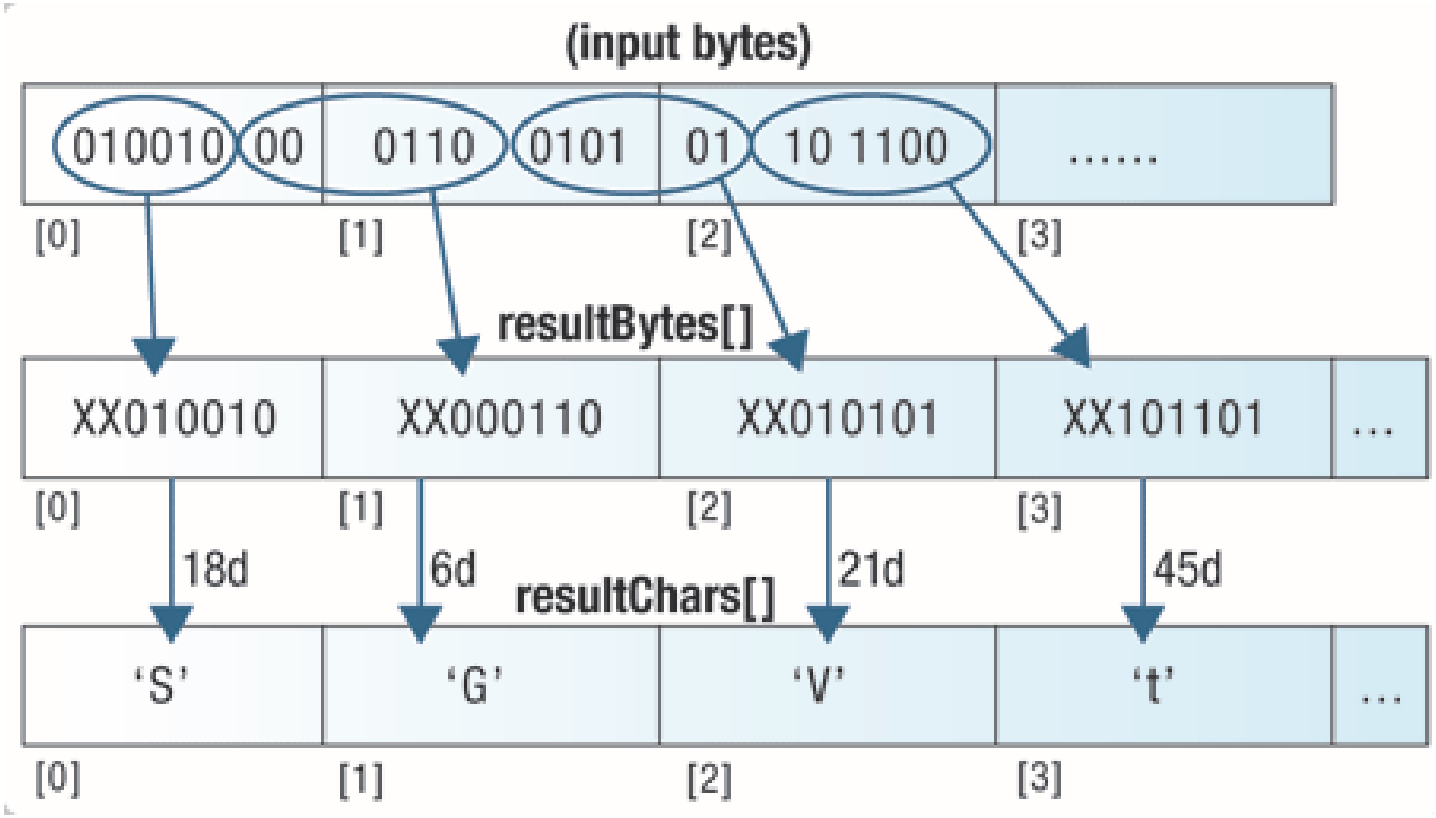
- 4月から名古屋大の客員 JK
- TopSE 講師
- (有)IT プランニング

Base64エンコーディング

$\mathrm{hoge} \rightarrow \mathrm{aG9nZQ==}$

- バイナリを印字可能なテキストに変換する方式
- 電子MailやOAuthなどで使われている

Base64エンコーディング



Base64エンコーディング

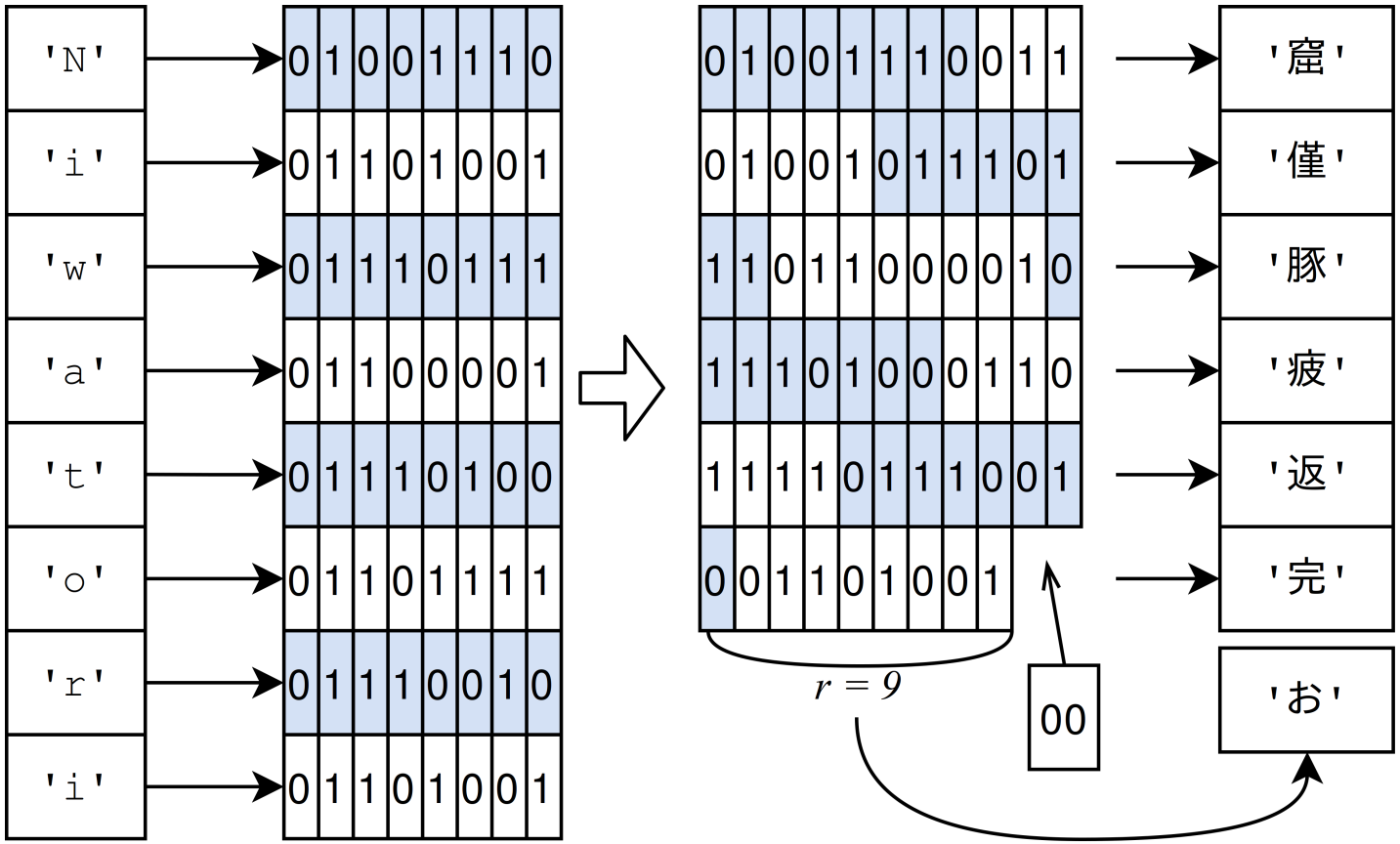
0	A	16	Q	32	g	48	w
1	B	17	R	33	h	49	x
2	C	18	S	34	I	50	y
3	D	19	T	35	j	51	z
4	E	20	U	36	k	52	0
5	F	21	V	37	l	53	1
6	G	22	W	38	m	54	2
7	H	23	X	39	n	55	3
8	I	24	Y	40	o	56	4
9	J	25	Z	41	p	57	5
10	K	26	a	42	q	58	6
11	L	27	b	43	r	59	7
12	M	28	c	44	s	60	8
13	N	29	d	45	t	61	9
14	O	30	e	46	u	62	+
15	P	31	f	47	v	63	/

Base64エンコーディング

- $64 (= 2^6)$ 種類の文字で表現する
- バイト数に対して およそ $\frac{8}{6}$ 倍の文字数
- Twitterや掲示板などでバイナリを表現したい時便利
- Tweetは最大140文字, つまり105バイトまでしか表現できない

Generalized Base-N encoding

$N = 2^{11}$ のとき



コード

```
Definition encode_ (r: nat) bs :=
  bs |> concat_map bits_of_byte
     |> fun bits ⇒ bits++repeat false(C.char_size-r)
     |> ntake C.char_size
     |> map C.of_bits
.
```

https://github.com/yoshihiro503/gbasen

証明

Theorem decode_encode : forall bs,
  GBaseN.decode (GBaseN.encode bs) = bs.
Proof.
  証明略

Qed.

https://github.com/yoshihiro503/gbasen

Base 常用漢字 encoding

- 常用漢字は2136($> 2^{11}$)字[1]
- 常用漢字をターゲットにすればBase2048 encodingができる
- バイト数に対し、文字数は約 $\frac{8}{11}$ に縮まる

nlnagoya → 際降津妬遍層お

[1](常用漢字表 平成22年内閣告示第2号 より)

Base 常用漢字 encoding

- unicodeは U+0000 - U+10FFFF までで (2^{20}) 字以上ある
- unicode文字をターゲットにすればBase1048576 encodingができる
- バイト数に対し、文字数は約 $\frac{8}{20}$

$\mathrm{nlnagoya} \rightarrow \mathrm{U{+}06e6c6\ U{+}0e6167}$
$\mathrm{U{+}06f796\ U{+}010000\ U{+}000004}$

次回ProofCafeは 5/21

- http://proofcafe.org/wiki/
- 東山線 伏見駅 から徒歩約1分
- 株式会社オンザロード様 会議室