



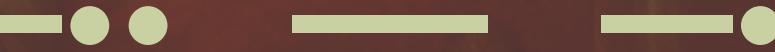


dog





dog





dog



⋮ ⋮ ⋮



dog



[64 6f 67]



80dogs.txt

Hund	hondur	köpegi	tz' i'
Ha'DIBaH	hualp	mbula	ukudla
abwo	huan	mbwa	welpe
alabai	hund	mbwagh	xolo
anjing	imba	mbwá	zwin
ashun	imbua	mbwene	āso
aso	imbwe	mbu	āšun
cane	inu	njau	šuo
cão	it	njoka	אָבָּא
cain	jindo	pa	כלב
câine	kelb	perro	كلب
câini	khuy	pes	ଫୁତ୍ତା
chien	klèb	pies	ಷೆ
chó	koer	qen	犬
cyn	koira	qeni	狗
dog	kotta	sag	ರೂಥ
emba	køter	sagol	ଜୀଏଚାଙ୍କୁ
ghaddu	kutya	sobaka	ପି
gom	kutta	suns	
hond	köpek	szczeniak	

- » Before Unicode
- » Unicode Unleashed
- » Oracle Through The Ages
- » Oracle CharacterSet

→ encode.py ascii < 80dogs.txt

✓ ascii: 53 good dogs

Hund Ha'DIBaH abwo alabai anjing ashun aso cane chien cyn dog emba
ghaddu gom hond hondur hualp huan hund imba imbua imbwe inu it
jindo kelb khuy koer koira kotta kutya kutta mbula mbwa mbwagh mbwene
njau njoka pa perro pes pies qen qeni sag sagol sobaka suns
szczeniak ukudla welpe xolo zwin

X ascii: 27 bad dogs



✓ ascii: 240 chars encoded in 240 bytes, 1.00 bytes per char



Huan

→ [48 75 61 6e] (ascii)

ascii	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x	<nul>	<soh>	<stx>	<etx>	<eot>	<enq>	<ack>	<bel>	<bs>	<tab>	<lf>	<vt>	<ff>	<cr>	<so>	<si>
1x	<dle>	<dc1>	<dc2>	<dc3>	<dc4>	<nak>	<syn>	<etb>	<can>		<sub>	<esc>	<fs>	<gs>	<rs>	<us>
2x	!	"	#	\$	%	&	'	()	*	+	,	-	.	/	
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	^	^	}	~

"WOOF"→[01010111 01001111 01001111 01000110]

"woof"→[01110111 01101111 01101111 01100110]

"HAU"→[1001000 1000001 1010101]

"hau"→[1101000 1100001 1110101]

iso-8859

Standard	Alias	Region	Year
iso-8859-1	latin1	Western Europe	1987
iso-8859-2	latin2	Central & Eastern Europe	1987
iso-8859-3	latin3	South Europe	1988
iso-8859-4	latin4	North Europe	1988
iso-8859-5	cyrillic		1988
iso-8859-6	arabic		1987
iso-8859-7	greek		1987
iso-8859-8	hebrew		1988
iso-8859-9	latin5	Turkish	1989
iso-8859-10	latin6	Nordic	1992
iso-8859-11	thai		2001
iso-8859-13	latin7	Baltics	1998
iso-8859-14	latin8	Celtic	1998
iso-8859-15	latin9	Western Europe (Improved)	1999
iso-8859-16	latin10	South-Eastern Europe	2001

```
→ encode.py ascii latin1 < 80dogs.txt
```

✓ ascii ✓ latin1: 53 good dogs

Ha'DIBaH Hund abwo alabai anjing ashun aso cane chien cyn dog emba
ghaddu gom hond hondur hualp huan hund imba imbua imbwe inu it
jindo kelb khuy koer koira kotta kutta kutya mbula mbwa mbwagh mbwene
njau njoka pa perro pes pies qen qeni sag sagol sobaka suns
szczeniak ukudla welpe xolo zwin

✗ ascii ✓ latin1: 9 bad dogs turned good

chó câin câine câini cão klèb köpek køter mbwá

✗ ascii ✗ latin1: 18 bad dogs

köpegi mbu tz'i' āso āšun ɬééchqá'í šuo كلب כלב אַבָּו קֹטָן ຈຳ

戌 犬 狗 犬 🐶

✓ ascii→latin1: 53→62 good dogs

✓ ascii: 240 chars encoded in 240 bytes, 1.00 bytes per char

✓ latin1: 278 chars encoded in 278 bytes, 1.00 bytes per char



cão
→[63 e3 6f] (latin1)


```
→ encode.py ascii latin2 < 80dogs.txt
```

✓ ascii ✓ latin2: 53 good dogs

Ha'DIBaH Hund abwo alabai anjing ashun aso cane chien cyn dog emba
ghaddu gom hond hondur hualp huan hund imba imbua imbwe inu it
jindo kelb khuy koer koira kotta kutta kutya mbula mbwa mbwagh mbwene
njau njoka pa perro pes pies qen qeni sag sagol sobaka suns
szczeniak ukudla welpe xolo zwin

✗ ascii ✓ latin2: 7 bad dogs turned good

chó câin câine câini köpek mbwá šuo

✗ ascii ✗ latin2: 20 bad dogs

cão klèb köpegi koter mbu tz'i' áso ášun l'ééchqá'í كلب אבו
କୁତ୍ତା ଶ୍ଵର ମୁଥ ଦୁଇ ପାତା ହଜାର ପାତା

✓ ascii→latin2: 53→60 good dogs

✓ ascii: 240 chars encoded in 240 bytes, 1.00 bytes per char

✓ latin2: 269 chars encoded in 269 bytes, 1.00 bytes per char

A painting of a yellow Labrador Retriever dressed as an Ottoman soldier. The dog wears a red fez with a black tassel, a red shaykap (skullcap) under it, and a dark blue robe with gold embroidery. A large gold medallion hangs from its neck. It holds a white tulip cup with a red tulip design in its front paws. In the background, there is a flag with a white crescent and star, a mosque with a dome and minaret, and another minaret. The sky is blue with some clouds.

köpek

→ [6b f6 70 65 6b] (latin2)

Standard	Alias	Windows Equivalent
iso-8859-1	latin1	windows-1252
iso-8859-2	latin2	windows-1250
iso-8859-5	cyrillic	windows-1251
iso-8859-6	arabic	windows-1256
iso-8859-7	greek	windows-1253
iso-8859-8	hebrew	windows-1255
iso-8859-9	latin5	windows-1254
iso-8859-11	thai	windows-874
iso-8859-13	latin7	windows-1257
iso-8859-15	latin9	windows-1252

gb2312

```
→ encode.py ascii gb2312 < 80dogs.txt
```

✓ ascii ✓ gb2312: 53 good dogs

Ha'DIBaH Hund abwo alabai anjing ashun aso cane
chien cyn dog emba ghaddu gom hond hondur
hualp huan hund imba imbua imbwe inu it
jindo kelb khuy koer koira kotta kutta kutya
mbula mbwa mbwagh mbwene njau njoka pa perro
pes pies qen qeni sag sagol sobaka suns
szczeniak ukudla welpe xolo zwin

✗ ascii ✓ gb2312: 7 bad dogs turned good

chó klèb mbwá tz'i' áso 犬 狗

✗ ascii ✗ gb2312: 20 bad dogs

câin câine câini cão köpek köpeğî kôter mbu
āšun žééchqâ'í šuo كلب كلب אַבּוֹ كُنْتَا ՚ون

RIθ Î  

✓ ascii→gb2312: 53→60 good dogs

✓ ascii: 240 chars encoded in 240 bytes, 1.00 bytes per char

✓ gb2312: 261 chars encoded in 269 bytes, 1.03 bytes per char







Hund

- [48 75 6e 64] (gb2312)
- [48 75 6e 64] (ascii)



→ [?? ?? ?? ??] (gb2312)

A painting of a dog dressed as a sailor in a gondola. The dog is wearing a straw hat with a dark band, a striped sailor's vest over a white shirt, and a red bandana. It is standing in a gondola, looking towards the camera. In the foreground, there is a pizza on a plate and a cup of coffee on a saucer. The background shows a canal in Venice with buildings and a bridge.

cane

→[63 61 6e 65] (gb2312)



→ [b9 b7 c8 ae] (gb2312)

?

→[. . b9 b7 . .] (gb2312)





肮发

→ [b0 b9 b7 a2] (gb2312)

Unicode

Codepoint

$U+0000 \rightarrow U+10ff ff$

Plane		Name
0	U+0000 → U+ffff	Basic Multilingual Plane
1	U+10000 → U+1ffff	Supplementary Multilingual Plane
2	U+20000 → U+2ffff	Supplementary Ideographic Plane
3	U+30000 → U+3ffff	Tertiary Ideographic Plane
4 → 13	U+40000 → U+dffff	Reserved
14	U+e0000 → U+effff	Suppl. Special-Purpose Plane
15	U+f0000 → U+fffff	Private Use Area A
16	U+100000 → U+10ffff	Private Use Area B

Range	Block
U+0000→U+007f	ASCII
U+0080→U+00ff	Latin-1 Supplement
U+0100→U+024f	Latin Extended
U+0250→U+02ff	Phonetic Symbols
U+0300→U+05ff	Greek, Cyrillic, Hebrew
U+0600→U+0fff	Arabic & Indian
U+1000→U+17ff	S & SE Asian
U+1800→U+24ff	E Asian
U+2000→U+27ff	Symbols, Punctuation
U+2800→U+28ff	Braille, Shapes
U+2e80→U+9fff	Chinese
U+ac00→U+d7af	Korean
U+d800→U+dbff	High Surrogates
U+dc00→U+dkff	Low Surrogates
U+e000→U+f8ff	Private Use

Unicode Encodings

» UTF-32

» UTF-16

» UTF-8

Unicode Encodings

» UTF-32

» UTF-16

» UTF-8

```
→ encode.py utf-32be < 80dogs.txt
```

✓ utf-32be: 80 good dogs

✓ utf-32be: 338 chars encoded in 1352 bytes, 4.00 bytes per char

犬 = U+72ac

- [00 00 72 ac] (utf-32be)
- [ac 72 00 00] (utf-32le)
- [00 00 fe ff 00 00 72 ac] (utf-32be + BOM)
- [ff fe 00 00 ac 72 00 00] (utf-32le + BOM)

Unicode Encodings

» UTF-32

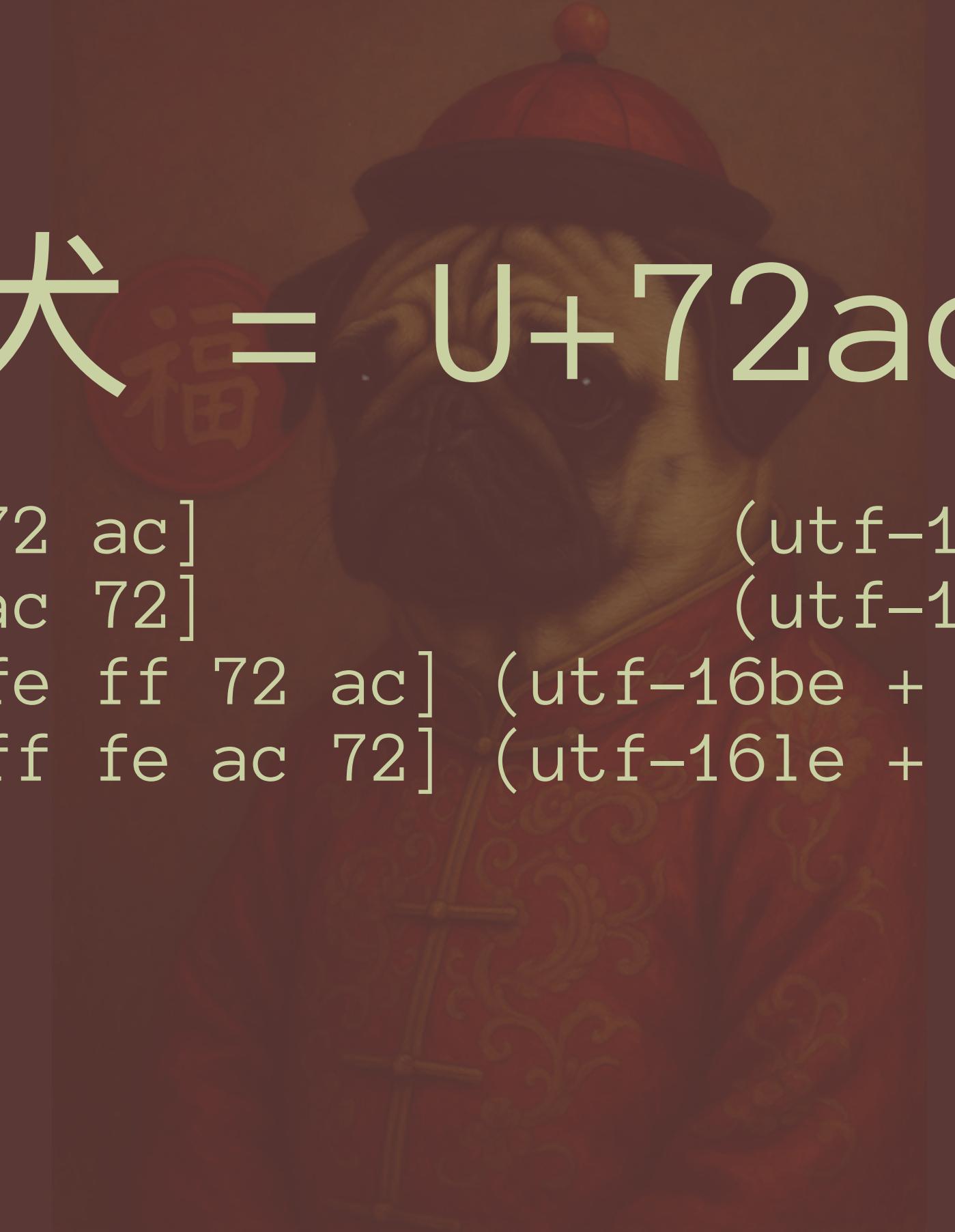
» UTF-16

» UTF-8

```
→ encode.py utf-16be < 80dogs.txt
```

✓ utf-16be: 80 good dogs

utf-16be: 338 chars encoded in 682 bytes, 2.02 bytes per char



犬 = U+72ac

- [72 ac] (utf-16be)
- [ac 72] (utf-16le)
- [fe ff 72 ac] (utf-16be + BOM)
- [ff fe ac 72] (utf-16le + BOM)

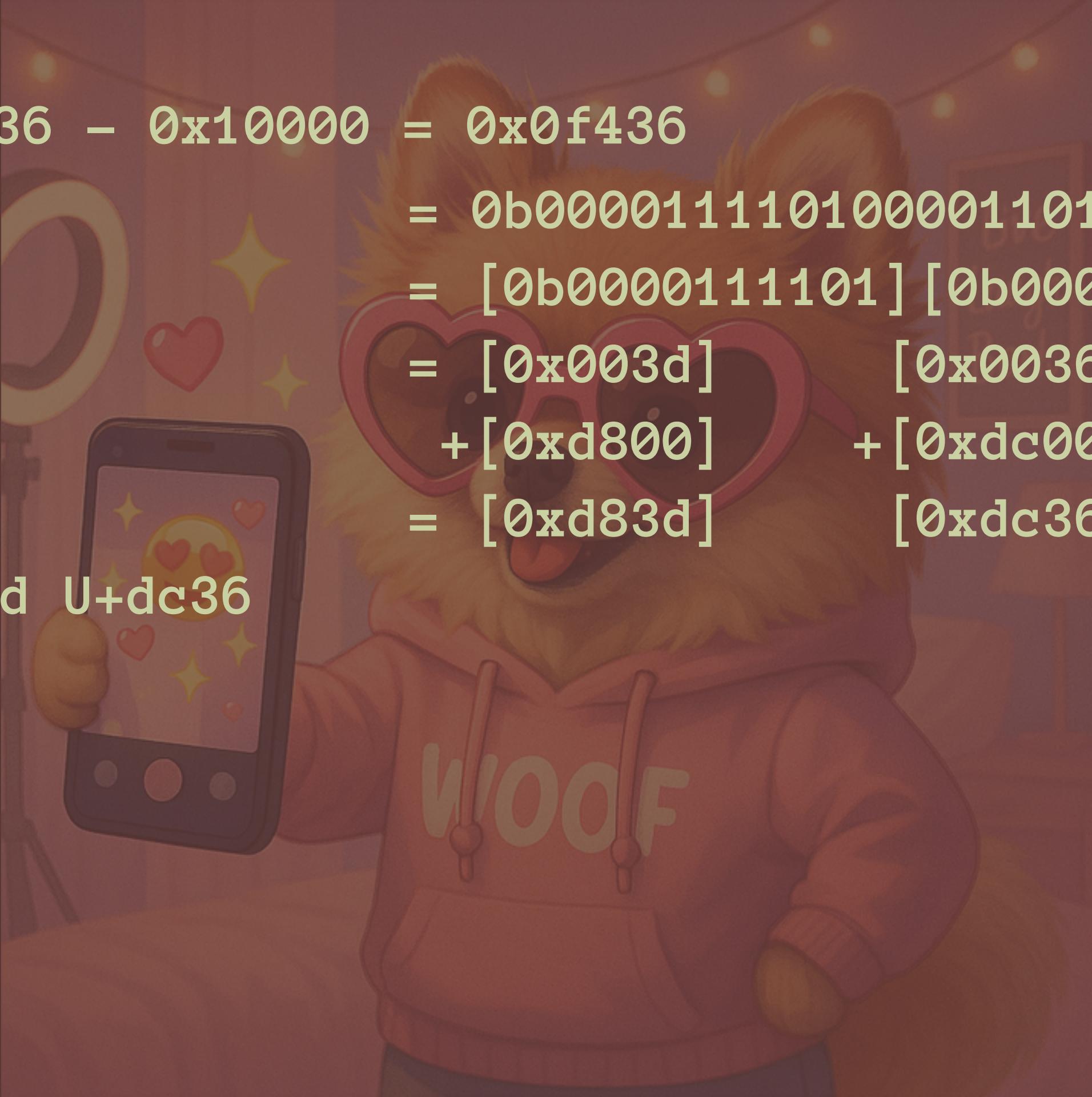




= U+1f436

→ [?? ??] (utf-16)

Range	Block
U+0000→U+007f	Latin Basic
U+0080→U+00ff	Latin Supplement
U+0100→U+024f	Latin Extended
U+0250→U+02ff	Phonetics
U+0300→U+05ff	Greek, Cyrillic, Hebrew
U+0600→U+0Fff	Arabic, Indian
U+1000→U+17ff	S & SE Asian
U+1800→U+24ff	E Asian
U+2000→U+27ff	Symbols
U+2800→U+28ff	Braille, Shapes
U+e80→U+9fff	Chinese
U+ac00→U+d7af	Korean
U+d800→U+dbff	High Surrogates
U+dc00→U+dfef	Low Surrogates
U+e000→U+f8ff	Private Use



🐶 = U+1f436 - 0x10000 = 0xf436
= 0b0000111010000110110
= [0b000011101] [0b0000110110]
= [0x003d] [0x0036]
+ [0xd800] + [0xdc00]
= [0xd83d] [0xdc36]

→ U+d83d U+dc36



= U+1f436

→ U+d83d U+dc36
→ [d8 3d] [dc 36] (utf-16)

Unicode Encoding

UTF-8

```
→ encode.py utf-8 < 80dogs.txt
```

✓ utf-8: 80 good dogs

✓ utf-8: 338 chars encoded in 413 bytes, 1.22 bytes per char

U+0000→U+007F

→[0xxxxxxx]

U+0080→U+07FF

→[110xxxxx 10xxxxxx]

U+0800→U+FFFF

→[1110xxxx 10xxxxxx 10xxxxxx]

U+10000→U+10FFFF

→[11110xxx 10xxxxxx 10xxxxxx 10xxxxxx]

U+0141 ('Ł')

= 00101000001

= 00101 000001

110xxxxx 10xxxxxx

→ [11000101 10000001]

= [c5 81]

[0xxxxxxx]

[110xxxxx 10xxxxxx]

[1110xxxx 10xxxxxx 10xxxxxx]

[11110xxx 10xxxxxx 10xxxxxx 10xxxxxx]

[0xxxxxxx] ASCII

[10xxxxxx] Continuation

[110xxxxx] Lead of 2 byte sequence

[1110xxxx] Lead of 3 byte sequence

[11110xxx] Lead of 4 byte sequence

[...10100011 10100011 11000011 10100011 01010101...]

A painting of a shaggy brown dog wearing a black top hat with a red feather and a dark blue vest with red embroidery. The dog is standing behind a wooden table. On the table is a plate of pierogi and a large mug of beer. In the background, a Polish flag flies from a pole, and a small church is visible in a green field under a blue sky with white clouds.

pies

→ [70 69 65 73] (ascii)
→ [70 69 65 73] (utf-8)



câine

→[63 c3 a2 69 6e 65] (utf-8)





→ [f0 9f 90 b6]

Encoding	Good Dogs	Bytes per Char
ascii	53	1.00
latin1	62	1.00
latin2	60	1.00
gb2312	60	1.03
utf-8	80	1.22
utf-16	80	2.02
utf-32	80	4.00

```
create database piesek  
characterset we8iso8859p1  
national character set al16utf16
```

...

```
create table dogs ( name nchar (10),  
                    breed char (10) );
```

```
insert into dogs (name, breed)  
values (n'犬', 'Shiba Inu');
```

we8iso8859p1
al16utf16

we 8 iso8859p1
al 16 utf16

Database Version

6

Character Set

we8iso8859p1

National Character Set

7

we8iso8859p2,
zhs16cgb231280

8i

utf8

utf8

9i

a132utf8

a116utf16

UTF8 (\neq UTF-8)

U+0000→U+007F

[0xxxxxxx]

U+0080→U+07FF

[110xxxxx 10xxxxxx]

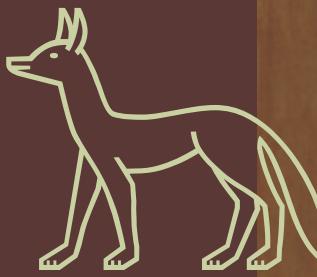
U+0800→U+FFFF

[1110xxxx 10xxxxxx 10xxxxxx]

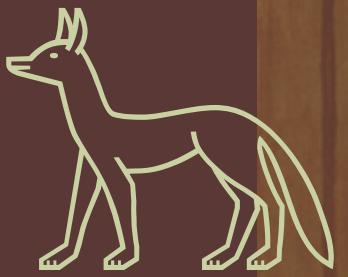
U+10000→U+10FFFF

[11110***-10*****-10*****-10*****]





= U+130e5



= U+130e5

→ U+d83c U+dce5
→ [ed a0 bc ed b3 a5]

What haven't covered

- » Windows Encodings
- » Byte/Char Semantics
- » Database Migration Assistant for Unicode (DMU)