

ASCII, Unicode, UTF-8

Informatik

Ein Computer speichert Zahlen. Um mit Zeichen umgehen zu können, wird über eine Zeichensatztafel (Codepage) jedem Zeichen eine Zahl zugeordnet. Der **ASCII-Code** (American Standard Code for Information Interchange) sieht in seiner ursprünglichen Version 7 Bits zur Codierung von Zeichen vor. Damit lassen sich $2^7 = 128$ Zeichen darstellen.

ASCII-Zeichentabelle, **hexadezimale** Nummerierung

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1...	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2...	SP	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

ISO-8859-1 ist eine Erweiterung des ASCII-Codes auf 8 bit und reicht für die meisten westeuropäischen Sprachen aus. Es fehlt aber das Eurozeichen und einige französische Zeichen.

ISO/IEC 8859-1

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	nicht belegt															
1...																
2...	SP	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
8...	nicht belegt															
9...																
A...	NBSP	ı	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	SHY	®	¯
B...	°	±	²	³	´	µ	¶	·	,	¹	º	»	¼	½	¾	¿
C...	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D...	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E...	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F...	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Unicode ist ein internationaler Standard, der jedem Schriftzeichen aller bekannter Sprachen eine eindeutige Zahl zuordnet (Code Point).

A	→	65
a	→	97
ß	→	223
€	→	8364

UTF-8 (1992) ist die am weitesten verbreitete Kodierung für Unicode-Zeichen.

Unicode-Zeichen größer als 127 werden in der UTF-8-Kodierung zu Byteketten der Länge zwei bis vier kodiert.

Unicode-Bereich (hexadezimal)	UTF-8-Kodierung (binär)	Bemerkungen	Möglichkeiten (theoretisch)	
0000 0000 – 0000 007F	0xxxxxxx	In diesem Bereich (128 Zeichen) entspricht UTF-8 genau dem ASCII-Code: Das höchste Bit ist 0, die restliche 7-Bit-Kombination ist das ASCII-Zeichen.	2^7	128
0000 0080 – 0000 07FF	110xxxxx 10xxxxxx	Das erste Byte beginnt immer mit 11, die folgenden Bytes mit 10. Die xxxxx stehen für die Bits des Unicode-Zeichenwerts. Dabei wird das niederwertigste Bit des Zeichenwerts auf das rechte x im letzten Byte abgebildet, die höherwertigen Bits fortschreitend <i>von rechts nach links</i> . Die Anzahl der Einsen vor der ersten 0 im ersten Byte ist gleich der Gesamtzahl der Bytes für das Zeichen. (In Klammern jeweils die theoretisch maximal möglichen.)	$2^{11} - 2^7$ (2^{11})	1920 (2048)
0000 0800 – 0000 FFFF	1110xxxx 10xxxxxx 10xxxxxx		$2^{16} - 2^{11}$ (2^{16})	63.488 (65.536)
0001 0000 – 0010 FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx		2^{20} (2^{21})	1.048.576 (2.097.152)

Beispiel: Die UTF-8 Codierung von €.

Code Point (dezimal) = 8364

Code Point (hex) = 20AC (zur Codierung werden 3 Bytes benötigt)

Code Point (binär) = 0010 0000 1010 1100

Aufteilung der bits in die 3 Bytes :

1110xxxx 10xxxxxx 10xxxxxx

11100010 10000010 10101100

E2 82 AC ist die Codierung des Eurozeichens.

In Python ist ein String-Objekt eine Folge von Zeichen in Unicode. Die Funktionen `chr` und `ord` wandeln den Unicode-Codepoint in das Zeichen um und umgekehrt.

Die Methode `encode('utf8')` gibt die UTF-8 Codierung des Zeichens als Bytefolge zurück.

```
>>> '\u20ac'
'€'
>>> chr(8364)
'€'
>>> chr(8364).encode('utf8')
b'\xe2\x82\xac'
>>>
```

Ein Texteditor soll folgenden Speicherbereich darstellen:

Address	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
00000000	11101111	10111011	10111111	11100010	10011001	10010100	00001010	11100010	10011001	10010101	00001010	11100010	10011001	10010110	00001010	11100010
00000010	10011001	10010111	00001010	11100010	10011001	10011000	00001010	11100010	10011001	10011001	00001010	11100010	10011001	10011010	00001010	11100010
00000020	10011001	10011011	00001010	11100010	10011001	10011100	00001010	11100010	10011001	10011101	00001010	11100010	10011001	10011110	00001010	11100010
00000030	10011001	10011111	00001010													

Ein Texteditor soll folgenden Speicherbereich darstellen:

Address	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
00000000	11101111	10111011	10111111	11100010	10011001	10010100	00001010	11100010	10011001	10010101	00001010	11100010	10011001	10010110	00001010	11100010
00000010	10011001	10010111	00001010	11100010	10011001	10011000	00001010	11100010	10011001	10011001	00001010	11100010	10011001	10011010	00001010	11100010
00000020	10011001	10011011	00001010	11100010	10011001	10011100	00001010	11100010	10011001	10011101	00001010	11100010	10011001	10011110	00001010	11100010
00000030	10011001	10011111	00001010													

Derselbe Speicherbereich hexadezimal:

Address	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
00000000	ef	bb	bf	e2	99	94	0a	e2	99	95	0a	e2	99	96	0a	e2
00000010	99	97	0a	e2	99	98	0a	e2	99	99	0a	e2	99	9a	0a	e2
00000020	99	9b	0a	e2	99	9c	0a	e2	99	9d	0a	e2	99	9e	0a	e2
00000030	99	9f	0a													

Die Bytefolge EF BB BF heißt **Byte Order Mark (BOM)** und gibt den Editor einen Hinweis darauf, dass eine UTF-8 Kodierung vorliegt.

0A ist die ASCII (und UTF-8) Codierung für den Zeilenvorschub.

Ein Texteditor soll folgenden Speicherbereich darstellen:

Address	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
00000000	11101111	10111011	10111111	11100010	10011001	10010100	00001010	11100010	10011001	10010101	00001010	11100010	10011001	10010110	00001010	11100010
00000010	10011001	10010111	00001010	11100010	10011001	10011000	00001010	11100010	10011001	10011001	00001010	11100010	10011001	10011010	00001010	11100010
00000020	10011001	10011011	00001010	11100010	10011001	10011100	00001010	11100010	10011001	10011101	00001010	11100010	10011001	10011110	00001010	11100010
00000030	10011001	10011111	00001010													

Derselbe Speicherbereich hexadezimal:

Address	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
00000000	ef	bb	bf	e2	99	94	0a	e2	99	95	0a	e2	99	96	0a	e2
00000010	99	97	0a	e2	99	98	0a	e2	99	99	0a	e2	99	9a	0a	e2
00000020	99	9b	0a	e2	99	9c	0a	e2	99	9d	0a	e2	99	9e	0a	e2
00000030	99	9f	0a													

Die Bytefolge EF BB BF heißt **Byte Order Mark (BOM)** und gibt den Editor einen Hinweis darauf, dass eine UTF-8 Kodierung vorliegt.

0A ist die ASCII (und UTF-8) Codierung für den Zeilenvorschub.

E2 99 94 ist die UTF-8 Codierung des hexadezimalen Codepoints 2654

Darstellung der Bit-Folge mit dem Font Lucida Sans Unicode.

