Buchstaben, Zeichen und Text

Informatik Grundlagen
Kantonsschule am Burggraben

Ivo Blöchliger

Alles Bits und Bytes

- Der Computer speichert nur Bits und Bytes
- Im Prinzip alles als natürliche Zahl interpretierbar.
  - z.B. ist ein Video eine einzige, riesige Zahl.
- Codierung:
  - Abmachung, wie die Bytes zu interpetieren sind.

Buchstaben

- Idee: Buchstaben (und andere Zeichen) nummerieren.
- Standard ASCII (7 Bits!)
- Was ergeben folgende Bytes, wenn man diese als ASCII-Zeichen interpretiert?

  0x4f 0x6b 0x21 0x0a 0x36 0x2e 0x30

- Suchen Sie dazu eine ASCII-Tabelle im Internet

0x4f 0x6b 0x21 0x0a 0x36 0x2e 0x30

0x4f 0x6b 0x21 0x0a 0x36 0x2e 0x30

ASCII-Zeichentabelle, hexadezimale Nummerierung															
	…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*

Was ist mit ä, ö, ù?

- ASCII-Code nur lateinische Buchstaben ohne Akzente
- Nur 7 Bits, also nochmals 128 Zeichen mit achtem Bit!
- Früher (und Windows noch heute):
  - Viele mögliche Zeichentabelle, je nach Sprachregion
  - Bei uns gebräuchlich: Latin-1 (ISO8859-1), CP-1252
- Heute (bis auf Windows fast ausschliesslich):

  Unicode: Jedem Zeichnen eine Nummer

  Momentan über 100’000 Zeichen spezifiziert.

Unicode

- Weltweit jedem Zeichen seine Nummer
- Emojis, Akzente, Schriftrichtung, Ligaturen
  - Siehe Unicode-Video
- Codierung meist UTF-8:
  - Standard-ASCII mit 1 Byte
  - Andere Zeichen mit 2 oder mehr Bytes
    - Erstes Byte gibt auch an, aus wie vielen Bytes das Zeichen besteht.
  - Speichereffizient für Sprachen mit lateinischem Alphabet.

Text-Dateien (nicht Word!)

- Folge von Buchstaben (Bytes)
  - Und Steuerzeichen wie Zeilenumbrüche ‘\n’, Tabulatoren ‘\t’,…
- Universal lesbar und veränderbar, wenn in ASCII
  - Unicode in UTF-8 codiert heute ebenfalls.
- Keine Stil-Information (Schriftart, Grösse, Farbe, etc.)
- Beispiel: Python-Code, CSV-Dateien (Tabellarische Daten)

Und mit Stil?

- Variante 1
  Proprietäres Format, nur mit einem Programm lesbar
- Variante 2
  Markup: z.B. HTML, XML

  <h1>**Titel**</h1>Rot!

- Office-Dokumente heute: zip-Archiv von XML-Dateien

Und in Python?

Und in Python?

```
>>> ord("A")
65
>>> chr(97)
'a'
>>> s = chr(0x1f600)
>>> s
'😀'
>>> [bin(x) for x in s.encode("utf-8")]
['0b11110000', '0b10011111', '0b10011000', '0b10000000']
```