Woche 02 Daten: Was sind Daten?

Skript

Erarbeitet von
Ann-Kathrin Selker

Lernziele

- Daten anhand von Beispielen in metrische und kategoriale Daten einsortieren
- Daten anhand von Beispielen in strukturierte und unstrukturierte Daten einsortieren
- erklären, wie Bilder im Computer dargestellt werden
- ASCII-kodierten Text in Zeichen umwandeln

Inhalt

Projekt „KI für alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Einstieg

Daten begegnen uns überall im Leben. Jeden Tag werden Wetterdaten erhoben, unsere Smartphones verwenden mobile Daten, und laut Datenschutzgrundverordnung sollen unsere persönlichen Daten geschützt werden. Doch was sind Daten überhaupt?

Metrische und kategoriale Daten

In der Informatik handelt es sich bei Daten um Informationseinheiten. Daten können dabei in metrische und kategoriale Daten unterschieden werden. Als metrische oder auch numerische Daten bezeichnen wir vereinfacht gesagt Zahlen, ob natürliche Zahlen (z. B. 1, 2, 3, 4), ganze Zahlen (-2, -1, 0, 1) oder Kommazahlen (1,2 ; 3,4567).
Im Gegensatz zu metrischen Daten kann man mit kategorialen Daten nicht rechnen. Kategoriale Daten sind entweder eine Kategorie, also z. B. die Fakultäten der HHU, oder eine Skala, zum Beispiel eine Bewertung von 1,0 bis 5,0 einer Klausur. Bei dieser Skala ist nur die Reihenfolge der Werte wichtig, nicht die Werte selber. Es ist also vollkommen egal, dass es in der Notenskala einen Sprung von 4,0 zu 5,0 gibt.
Auch die Zeichen eines Textes sind kategoriale Daten. Im Computer wird ein Zeichen in der Grundkodierung ASCII als Zahl zwischen 0 und 127 gespeichert. ASCII steht für American Standard Code for Information Interchange. Eine Kodierung ist hier eine Tabelle, die jedem Zeichen einen Zahlenwert zuordnet. Der zugeteilte Zahlenwert ist im Grunde bedeutungslos und muss nur konsistent für dasselbe Zeichen verwendet werden.

Dez	Hex	Okt	
0	0x00	000	NUL
1	0x01	001	SOH
2	0x02	002	STX
3	0x03	003	ETX
4	0x04	004	EOT
5	0x05	005	ENQ
6	0x06	006	ACK
7	0x07	007	BEL
8	0x08	010	BS
9	0x09	011	TAB
10	0x0A	012	LF
11	0x0B	013	VT
12	0x0C	014	FF
13	0x0D	015	CR
14	0x0E	016	SO
15	0x0F	017	SI
16	0x10	020	DLE
17	0x11	021	DC1
18	0x12	022	DC2
19	0x13	023	DC3
20	0x14	024	DC4
21	0x15	025	NAK
22	0x16	026	SYN
23	0x17	027	ETB
24	0x18	030	CAN
25	0x19	031	EM
26	0x1A	032	SUB
27	0x1B	033	ESC
28	0x1C	034	FS
29	0x1D	035	GS
30	0x1E	036	RS
31	0x1F	037	US

Dez	Hex	Okt	
32	0x20	040	SP
33	0x21	041	!
34	0x22	042	"
35	0x23	043	#
36	0x24	044	$
37	0x25	045	%
38	0x26	046	&
39	0x27	047	'
40	0x28	050	(
41	0x29	051	)
42	0x2A	052	*
43	0x2B	053	+
44	0x2C	054	,
45	0x2D	055	-
46	0x2E	056	.
47	0x2F	057	/
48	0x30	060	0
49	0x31	061	1
50	0x32	062	2
51	0x33	063	3
52	0x34	064	4
53	0x35	065	5
54	0x36	066	6
55	0x37	067	7
56	0x38	070	8
57	0x39	071	9
58	0x3A	072	:
59	0x3B	073	;
60	0x3C	074	<
61	0x3D	075	=
62	0x3E	076	>
63	0x3F	077	?

Dez	Hex	Okt	
64	0x40	100	@
65	0x41	101	A
66	0x42	102	B
67	0x43	103	C
68	0x44	104	D
69	0x45	105	E
70	0x46	106	F
71	0x47	107	G
72	0x48	110	H
73	0x49	111	I
74	0x4A	112	J
75	0x4B	113	K
76	0x4C	114	L
77	0x4D	115	M
78	0x4E	116	N
79	0x4F	117	O
80	0x50	120	P
81	0x51	121	Q
82	0x52	122	R
83	0x53	123	S
84	0x54	124	T
85	0x55	125	U
86	0x56	126	V
87	0x57	127	W
88	0x58	130	X
89	0x59	131	Y
90	0x5A	132	Z
91	0x5B	133	[
92	0x5C	134	\
93	0x5D	135	]
94	0x5E	136	^
95	0x5F	137	_

Dez	Hex	Okt	
96	0x60	140	`
97	0x61	141	a
98	0x62	142	b
99	0x63	143	c
100	0x64	144	d
101	0x65	145	e
102	0x66	146	f
103	0x67	147	g
104	0x68	150	h
105	0x69	151	i
106	0x6A	152	j
107	0x6B	153	k
108	0x6C	154	l
109	0x6D	155	m
110	0x6E	156	n
111	0x6F	157	o
112	0x70	160	p
113	0x71	161	q
114	0x72	162	r
115	0x73	163	s
116	0x74	164	t
117	0x75	165	u
118	0x76	166	v
119	0x77	167	w
120	0x78	170	x
121	0x79	171	y
122	0x7A	172	z
123	0x7B	173	{
124	0x7C	174	|
125	0x7D	175	}
126	0x7E	176	~
127	0x7F	177	DEL

Einblendung ASCII-Tabelle (Quelle [1])

Die ASCII-Kodierung enthält die lateinischen Groß- und Kleinbuchstaben, Ziffern und Satzzeichen. Dabei wird ein kleines a durch die Zahl 97 und ein großes A durch die Zahl 65 repräsentiert. Die Ziffer 7 wird mit einer 55 dargestellt und ein Punkt hat die Zahl 46. Die ASCII-Tabelle enthält aber auch nicht darstellbare Zeichen, genannt Steuerungszeichen, wie das Tabulatorzeichen mit der 9. Die Sonderzeichen einiger Sprachen, wie zum Beispiel die deutschen Umlaute, sind nicht enthalten. Heutzutage wird die erweiterte Kodierung Unicode benutzt, die zwar mehr Speicherplatz pro Zeichen verbraucht, aber dafür auch sämtliche Sonderzeichen aller Sprachen und Emojis darstellen kann. Inzwischen enthält Unicode etwa 145.000 Zeichen. Die ASCII-Tabelle ist dabei integriert, sodass sich der Zahlenwert für z. B. ein kleines a nicht geändert hat.
Durch Aneinanderreihung einzelner Zeichen entstehen sogenannte Strings. Da die Zeichen eines Strings auch Leer- und Satzzeichen beinhalten, sind Strings daher nicht nur einzelne Wörter wie "KI", sondern auch ganze Texte wie „Herzlich willkommen in der Vorlesung ‚KI für alle´! Schön, dass ihr da seid.“

Strukturierte und unstrukturierte Daten

In den späteren Anwendungen wollen wir aber auch komplexere Daten abbilden können. Wir unterscheiden zwischen strukturierten und unstrukturierten Daten. Strukturierte Daten sind in einem festgelegten Format erstellt. Ein Beispiel dafür sind Daten in Tabellen. Tabellen bestehen aus Zeilen und Spalten, bei denen jede Zeile für ein Datenobjekt steht und jede Spalte für Eigenschaften dieses Objekts, sogenannte Features. Features können dabei verschiedene Datentypen haben.

Matrikelnummer	Name	Vorname	Studienfach
1234567	Fischer	Anna	Philosophie
1289365	Ostrowska	Waleria	Mathematik
1314532	Hietanen	Toni	Germanistik
1349823	Osterhagen	Maximilian	Computerlinguistik
1453211	Glöckner	Marina	Pharmazie
1444478	Wei	Fen	Rechtswissenschaften
1521223	Rana	Chandra	Pharmazie
1523898	Yuryeva	Nora	Kunstgeschichte
1554599	Page	Sean	Informatik
1577266	Kuster	Martin	Medizin
1612923	Schulze	Anja	Betriebswirtschaftslehre
1682789	Yamauchi	Hisako	Wirtschaftschemie
1725521	Chia	Qiang	Medizin
1921456	Preciado Mena	Pirro	Informatik

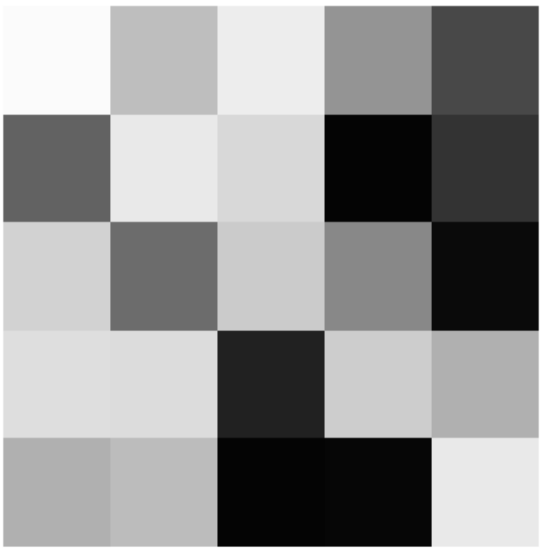
Einblendung Tabelle Studierendendaten

In dieser Tabelle beinhaltet jede Zeile die Daten eines Studierenden, in der ersten Zeile z. B. von Anna Fischer mit Matrikelnummer 1234567 und Studienfach Philosophie. Hier werden Matrikelnummern als natürliche Zahlen dargestellt, Namen als Strings, und das Feature „Studienfach“ beinhaltet Kategorien. Dieses Schema ist für alle Studierenden gleich, d. h. der erste Wert ist immer eine Matrikelnummer etc.

Unter unstrukturierten Daten verstehen wir Daten, die keine normalisierte, feste Struktur aufweisen. Darunter fallen Bilder, Texte und ähnliche Daten. Damit diese verarbeitet werden können, muss ihnen erst noch eine Art Struktur gegeben werden. Bei Bildern ist dies zum Beispiel durch Raster möglich. Dabei wird ein Bild in Zeilen und Spalten aufgeteilt. Jedes dadurch entstandene Kästchen wird als Pixel bezeichnet und je nach Anwendung durch einen oder mehrere Zahlenwerte repräsentiert. Bei sogenannten Grayscale-Bildern (also Graustufenbildern) wird ein Pixel durch eine Zahl zwischen 0 und 255 dargestellt, die die Intensität des Lichts angibt. Dabei steht 0 für kein Licht, also schwarz, und 255 für nur Licht, also weiß. Die Zahlen dazwischen sind Abstufungen von grau. Da es insgesamt 256 mögliche Pixelwerte gibt, können auch nur 256 verschiedene Graustufen angezeigt werden.

Einblendung Schmetterlingsbild (Quelle [2])

Das ursprüngliche Schmetterlingsbild besitzt über 300.000 Pixel. Wenn wir jetzt in der Kopfgegend des Schmetterlings in das Bild herein zoomen, erhalten wir diesen Teilabschnitt mit 5 Zeilen und 5 Spalten und damit insgesamt 25 Pixeln,



250	176	231	131	56
78	226	206	4	37
199	90	191	117	10
214	213	26	195	161
160	174	5	6	226

Einblendung gezoomtes Bild + Tabelle mit Pixelwerten

bei dem man klar das Raster erkennen kann. Zum Beispiel hat das Pixel in der linken oberen Ecke den Wert 250, das daneben den Wert 176 usw.
Um Farbbilder darzustellen, reichen Lichtintensitäten nicht aus. Bei RGB-Bildern wird jedes Pixel durch drei Zahlenwerte zwischen 0 und 255 dargestellt: Dabei steht der erste Wert für rot (R), der zweite für grün (G), und der dritte für blau (B). Diese drei Farben bezeichnen wir auch als Grundfarben. Je niedriger der Zahlenwert für eine Grundfarbe, desto weniger Anteile hat diese an der resultierenden Farbe.

R	G	B	Farbe
0	0	255	
255	128	0	
0	0	0	
255	255	255	

Zum Beispiel hat reines Blau den Wert (0 ,0, 255), reines Orange den Wert (255, 128, 0), schwarz den Wert (0, 0, 0) und weiß den Wert (255, 255, 255). Da jede Kombination von Rot-, Grün- und Blauwerten eine neue Farbe angibt, können insgesamt 256 * 256 * 256 verschiedene Farben angezeigt werden, also mehr als 16 Millionen.

Abschluss

In diesem Video habt ihr gelernt, was wir unter metrischen und kategorialen Daten verstehen, was wir unter unstrukturierten und strukturierten Daten verstehen, wie Textkodierung funktioniert und wie Bilder im Computer dargestellt werden können.

Quellen

Quelle [1] Bild „ASCII“ von Hador, Hervorhebungen hinzugefügt.
https://de.wikiversity.org/wiki/Datei:ASCII.pdf
Lizenz CC BY-SA 3.0

Quelle [2] pixabay.com

Weiterführendes Material

http://christianherta.de/lehre/infoDarstellung.pdf

Disclaimer