Entropie

Ingo Blechschmidt,
Michael Hartmann

15. November 2006

Inhalt

Mögliche Definitionen

Information
(v. lat.: informare $\hat{=}$ bilden, eine Form geben):

- Muster von Materie oder Energieformen
- Beseitigung von Ungewissheit

Informationsebenen

1. Codierung
2. Syntax
3. Semantik
4. Pragmatik

Codierung

Codierung

Vorschrift, mit der Informationen zur Übertragung umgewandelt werden kann.

Beispiele:

- Sprache
- Schrift
- Brailleschrift
- Flaggenalphabet
- Morsezeichen

Syntax

Syntax

Information als bloße Struktur, ohne „Sinn“.
Inhalt oder Bedeutung der Information sind irrelevant.

Beispiele:

- Übertragung von Webseiten auf einen Computer
- Übertragung von Bildern einer Überwachungskamera
- Übertragung von Tönen beim Telefonieren

Semantik

Semantik

Interpretierte, sinnbehaftete Information

$\frac{1}{2}gh$ → Allgemeine Formel für den Flächeninhalt eines Dreiecks

S. 251/29 → Aufgabe 29 im Stochastik-Buch auf Seite 251

4 → vier Personen im Raum; vier Stunden lange Klausur (kontextabhängige Interpretation)

Pragmatik

Pragmatik

„Effektiver“ Informationsgehalt einer Information

Beispiele:

1. Situation: Person will gerade das Haus verlassen
   „Es ist kalt.“ → Informationsgewinn (warme Kleidung anziehen)
2. Situation: Person wartet frierend auf den Bus
   „Es ist kalt.“ → kein Informationsgewinn

Modellierung nach Shannon

Informationsgehalt eines Zeichens

$$I(x) := -\log_2 P(x); \qquad [1\,\text{bit}]$$

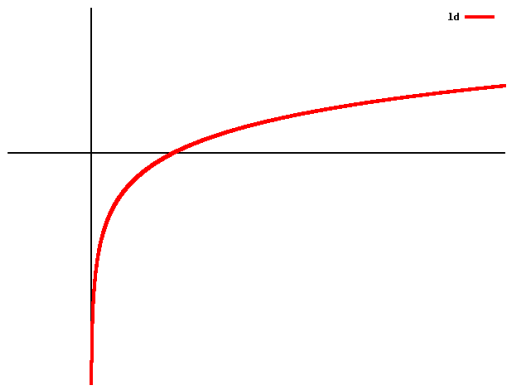
Entropie: Zu erwartender Informationsgehalt

$$E(I) := \sum_{x \in \Sigma} P(x) \cdot I(x);$$

x: ein bestimmtes Zeichen

Σ: Menge aller vorkommenden Zeichen

$P(x)$: Wahrscheinlichkeit des Auftretens von x



Eigenschaften der Entropie

- Je seltener ein Zeichen, desto höher der Informationsgehalt
- Spezialfall bei gleicher Häufigkeit aller Zeichen:

$$\underbrace{\begin{aligned} P(x_1) &= P(x_2) = \cdots = P(x_n); \\ I(x_1) &= I(x_2) = \cdots = I(x_n) = -\log_2 P(x_n); \end{aligned}}_{E(I) = -\log_2 P(x_n);}$$

  Beispiel:

$$\begin{aligned} M &= abcdabcdabcd; \\ P(x_n) &= \tfrac{1}{4}; \\ E(I) &= -\log_2 \tfrac{1}{4} = 2\,\text{bit}; \end{aligned}$$

- Maximale Entropie bei $|\Sigma| = 2^n$, $n \in \mathbb{N}$:
  $E(I) = -\log_2 \frac{1}{2^n} = n\,\text{bit};$

Beispiel: Nachricht

Definitionen

$$\begin{aligned} I(x) &:= -\log_2 P(x); \\ E(I) &:= \sum_{x \in \Sigma} P(x) \cdot I(x); \end{aligned}$$

$$\begin{aligned} M &= (i, n, f, o, r, m, a, t, i, o, n); \\ \Sigma &= \{i, n, f, o, r, m, a, t\}; \end{aligned}$$

$$\begin{aligned} I(f) = I(r) = I(m) = I(a) = I(t) &= -\log_2 \tfrac{1}{11} \approx 3{,}46\,\text{bit}; \\ I(o) = I(i) = I(n) &= -\log_2 \tfrac{2}{11} \approx 2{,}46\,\text{bit}; \end{aligned}$$

$$E(I) \approx 5 \cdot \tfrac{1}{11} \cdot 3{,}46\,\text{bit} + 3 \cdot \tfrac{2}{11} \cdot 2{,}46\,\text{bit} \approx 2{,}91\,\text{bit};$$

Beispiel: LAPLACEscher Münzwurf

Definitionen

$$\begin{aligned} I(x) &:= -\log_2 P(x); \\ E(I) &:= \sum_{x \in \Sigma} P(x) \cdot I(x); \end{aligned}$$

$\Sigma = \{\text{Kopf}, \text{Zahl}\}$;

$P(\text{Kopf}) = P(\text{Zahl}) = \frac{1}{2}$;

$I(\text{Kopf}) = I(\text{Zahl}) = -\log_2 \frac{1}{2} = 1\,\text{bit}$;

$$\begin{aligned} E(I) &= \tfrac{1}{2} I(\text{Kopf}) + \tfrac{1}{2} I(\text{Zahl}) = \\ &= \tfrac{1}{2} \cdot 1\,\text{bit} + \tfrac{1}{2} \cdot 1\,\text{bit} = 1\,\text{bit}; \end{aligned}$$

Beispiel: Gezinkter Münzwurf

Definitionen

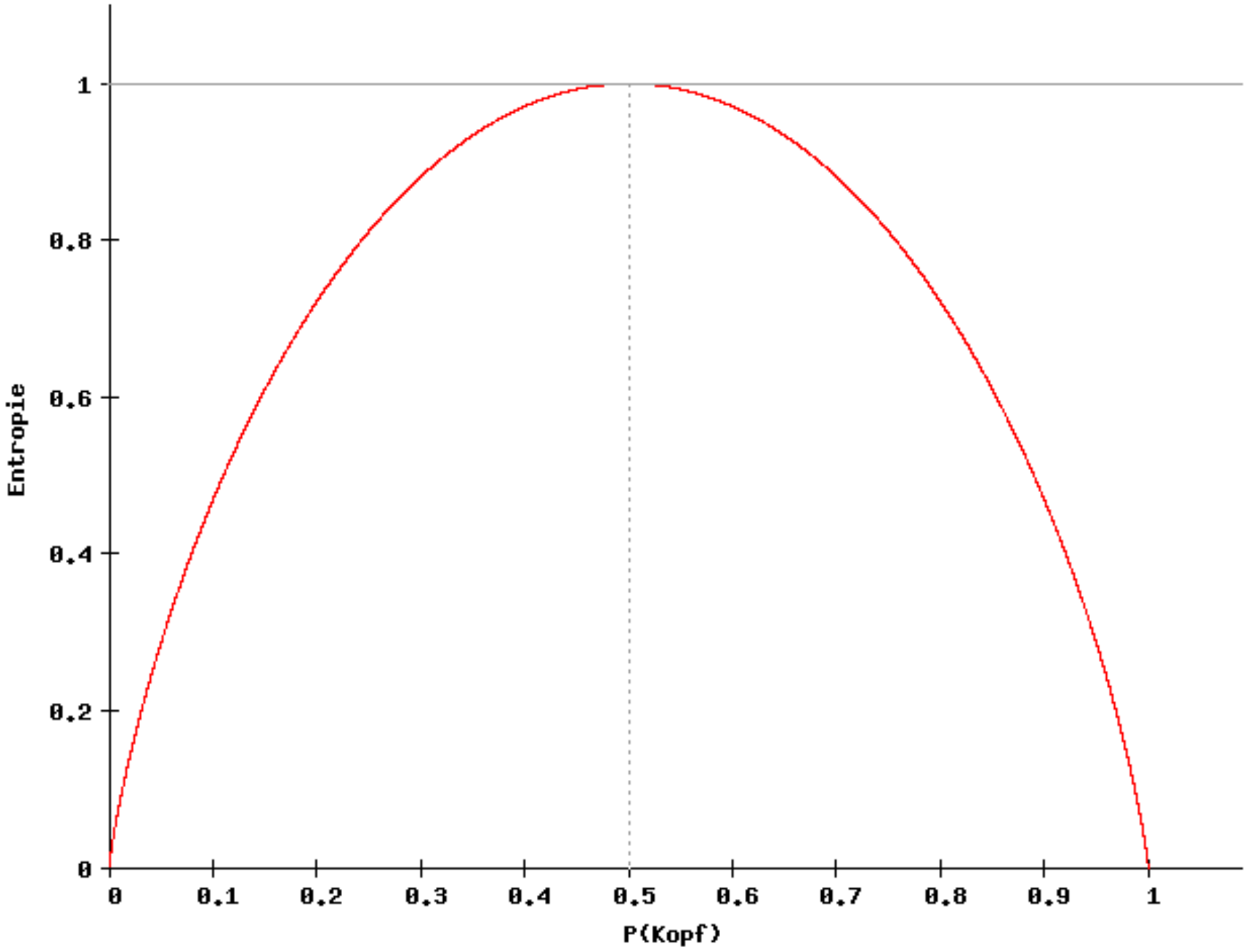
$$\begin{aligned} I(x) &:= -\log_2 P(x); \\ E(I) &:= \sum_{x \in \Sigma} P(x) \cdot I(x); \end{aligned}$$

$\Sigma = \{\text{Kopf}, \text{Zahl}\}$;

$P(\text{Kopf}) = p = 1 - P(\text{Zahl})$;

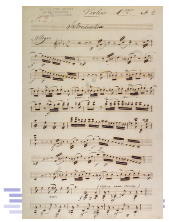
$$\begin{aligned} I(\text{Kopf}) &= -\log_2 p; \\ I(\text{Zahl}) &= -\log_2 (1-p); \end{aligned}$$

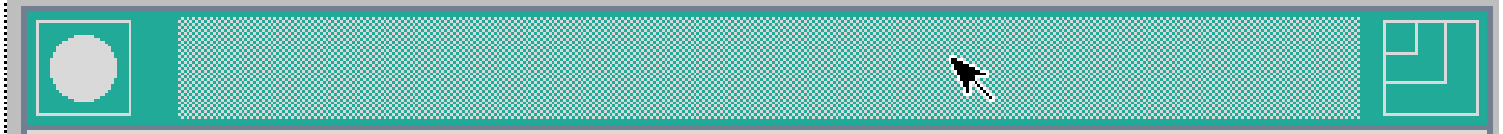
$$\begin{aligned} E(I) &= p\, I(\text{Kopf}) + (1-p)\, I(\text{Zahl}) = \\ &= p \cdot [-\log_2 p] + (1-p) \cdot [-\log_2 (1-p)]; \end{aligned}$$


Entropie
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
P(Kopf)

Beispiel: Entropie von Musikstücken

- Mozart, Bach: Klassische Komponisten – Harmonienlehre, Akkorde, Intervalle etc.
- Schönberg (20. Jhd.): Zwölftonmusik – Wiederholung bestimmter zwölf Töne auf verschiedene Arten
  $\rightarrow$ subjektiv wirr, chaotisch
- Entropie von Zwölftonstücken höher als bei Mozart und Bach!
- Kritik am Verfahren:
  (Viel zu) kleiner Korpus,
  „willkürliche“ Kodierung der Musik zu Buchstaben





Eingabe

GFGAFEFGEDEFDCDEcFcEcD=B
ccfcecdbcgfgbfefgedefdcd
CEGcDcBADGBdEdcBEAceFedcFBdfGfedecBcdBABcBAGF
ecfdgfgeagafbabgcagabgfgazzedc
dBABcAGABAGFEzzcAGABGFGAFEFTAA
bgfgafefgdcbafefgedefdcdecbc
AGFGTAAAdfedcBAGBeGFeFdeBABcAGA
dgfgcagaRbzdecARFeAe

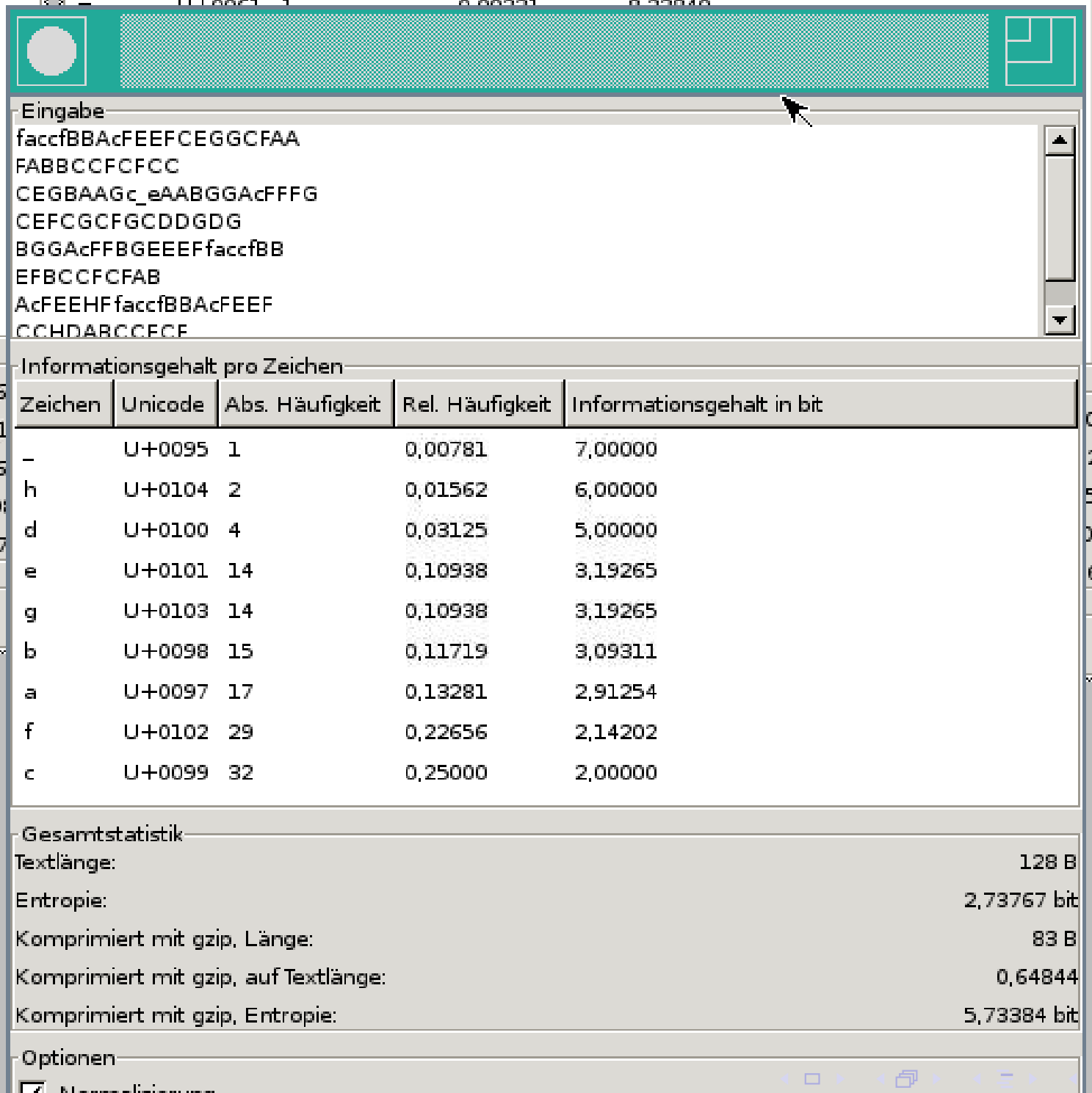
Informationsgehalt pro Zeichen

Zeichen	Unicode	Abs. Häufigkeit	Rel. Häufigkeit	Informationsgehalt in bit
=	U+0061	1	0,00331	8,23840
t	U+0116	2	0,00662	7,23840
z	U+0122	6	0,01987	5,65344
d	U+0100	35	0,11589	3,10912
b	U+0098	37	0,12252	3,02895
c	U+0099	37	0,12252	3,02895
a	U+0097	41	0,13576	2,88085
f	U+0102	45	0,14901	2,74655
e	U+0101	49	0,16225	2,62369

Gesamtstatistik

Textlänge:	302 B
Entropie:	2,94182 bit
Komprimiert mit gzip, Länge:	152 B
Komprimiert mit gzip, auf Textlänge:	0,50331
Komprimiert mit gzip, Entropie:	6,57356 bit

Optionen



Eingabe

faccfBBAcFEEFCEGGCFAA
FABBCCFCFCC
CEGBAAGc_eAABGGAcFFFG
CEFCGCFGCDDGDG
BGGAcFFBGEEEFfaccfBB
EFBCCFCFAB
AcFEEHFfaccfBBAcFEEF
CCHDABCCFCF

Informationsgehalt pro Zeichen

Zeichen	Unicode	Abs. Häufigkeit	Rel. Häufigkeit	Informationsgehalt in bit
_	U+0095	1	0,00781	7,00000
h	U+0104	2	0,01562	6,00000
d	U+0100	4	0,03125	5,00000
e	U+0101	14	0,10938	3,19265
g	U+0103	14	0,10938	3,19265
b	U+0098	15	0,11719	3,09311
a	U+0097	17	0,13281	2,91254
f	U+0102	29	0,22656	2,14202
c	U+0099	32	0,25000	2,00000

Gesamtstatistik

Textlänge:	128 B
Entropie:	2,73767 bit
Komprimiert mit gzip, Länge:	83 B
Komprimiert mit gzip, auf Textlänge:	0,64844
Komprimiert mit gzip, Entropie:	5,73384 bit

Optionen

☑ Normalisierung

root's X desktop (m19s28:1)

Eingabe

aaebbdadfdfhgcecafbdbbcagghd

eedageecedfffadh bbecfhgbc hcgaccachfahhedh

Informationsgehalt pro Zeichen

Zeichen	Unicode	Abs. Häufigkeit	Rel. Häufigkeit	Informationsgehalt in bit
g	U+0103	6	0,08696	3,52356
b	U+0098	8	0,11594	3,10852
f	U+0102	8	0,11594	3,10852
d	U+0100	9	0,13043	2,93860
e	U+0101	9	0,13043	2,93860
h	U+0104	9	0,13043	2,93860
a	U+0097	10	0,14493	2,78660
c	U+0099	10	0,14493	2,78660

Gesamtstatistik

Textlänge:	69 B
Entropie:	2,98481 bit
Komprimiert mit gzip, Länge:	62 B
Komprimiert mit gzip, auf Textlänge:	0,89855
Komprimiert mit gzip, Entropie:	5,28937 bit

Optionen

☑ Normalisierung

Shannon–Fano-Kodierung

Shannon–Fano-Kodierung

Entropiekodierung („Kompressionsverfahren“)

- Darstellung *häufiger* Zeichen durch *kurze* Bitfolgen;
  Darstellung *seltener* Zeichen durch *lange* Bitfolgen
- Eindeutigkeit der Bitfolgen („Präfixfreiheit“)

Problembeispiel: $A \mapsto 10 \quad B \mapsto 01 \quad C \mapsto 0$

$$\left.\begin{array}{l} ABC \mapsto 10010 \\ ACA \mapsto 10010 \end{array}\right\} \text{nicht eindeutig}$$

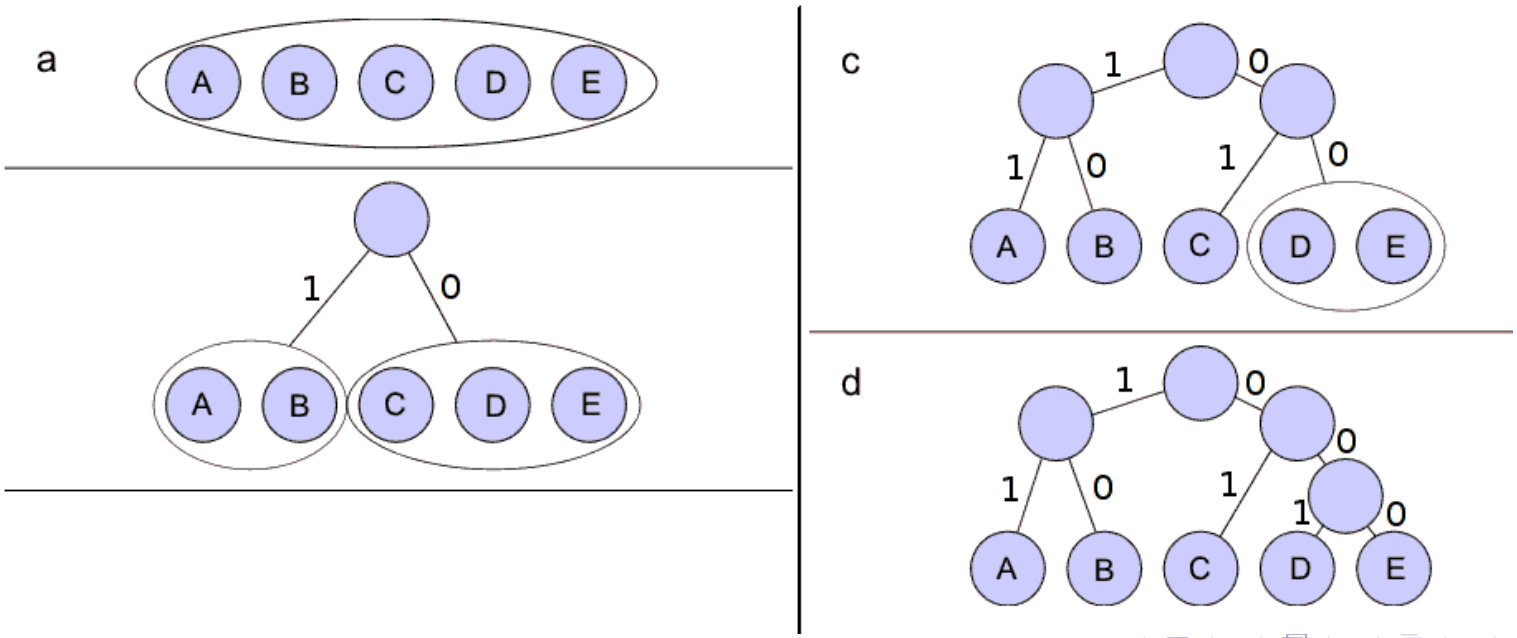
Algorithmus

1. Sortierung der Zeichen nach rel. Häufigkeit
2. Einteilung der Zeichen in zwei Gruppen, sodass Summen der Häufigkeiten etwa gleich
3. So lange fortfahren, bis Entsprechung jedes Zeichens durch einen Pfad im Baum

Beispiel

Text (39 Zeichen):
ABADDCCAABABEDAECBDDDAAAABAAAABBCAECECE

Zeichen	A	B	C	D	E
Abs. Häufigkeit	15	7	6	6	5



Beispiel

- Original
  (117 bit; Entropie $\approx$ 0,82 bit):
  00000100001101101001000000000
  10000011000110001000100010110
  11011000000000000001000000000
  00000100101000010001010010100

Zeichen	A	B	C	D	E
Abs. Häufigkeit	15	7	6	6	5
Benötigte Bits	3	3	3	3	3

Bit	0	1
Abs. Häufigkeit	87	30

$A \mapsto 000$
$B \mapsto 001$
$C \mapsto 010$
$D \mapsto 011$
$E \mapsto 100$

- Komprimiert
  (89 bit; Entropie $\approx$ 0,99 bit (!!)):
  11101100100101011111 10
  11100000011100001100 01
  00100111111111101111 11
  11101001110000100001000

Zeichen	A	B	C	D	E
Abs. Häufigkeit	15	7	6	6	5
Benötigte Bits	2	2	2	3	3

Bit	0	1
Abs. Häufigkeit	40	49

$A \mapsto 11$
$B \mapsto 10$
$C \mapsto 01$
$D \mapsto 001$
$E \mapsto 000$

Anwendung des Entropiekonzepts

- Maß für die untere Schranke verlustfreier Kompression
- Maß für die Zufälligkeit von Information (Zufallsgeneratoren)
- Maß für „Chaos“
  (hohe Entropie $\hat{=}$ hohe Überraschung)
- Charakteristika für Autoren
  (Entropienvergleiche $\rightarrow$ Schlüsse auf Autoren)

Nachteile

- Keine Beachtung der Reihenfolge:
  Entropien von
  00001111 und
  01101000 gleich! –
  (mögliche) Lösung: Algorithmische Information –
  Länge des kürzesten Beschreibung in einer bestimmten Sprache
- Analyse nur auf syntaktischer Ebene –
  (mögliche) Lösung: Memetik:
  Übertragung der Prinzipien der Evolutionstheorien auf Gedanken
  ($\rightarrow$ Mutation, Viren)

Bildnachweis

- http://www.thewirelessreport.com/media/2006/06/library-books.jpg
- http://www.borismatas.de/Schichten.jpg
- http://www.library.upenn.edu/exhibits/rbm/music/2-6a.jpg
- http://www.zappelfillip.de/wordpress/postimage/jahr2006/braillegoogle.gif
- http://www.catch-artists.de/telefon.jpg
- http://www.pixelplexus.co.za/blog/pics/MIA01.jpg
- http://de.wikipedia.org/wiki/Bild:ShannonCodeAlg.png