Constantin Lazari, Marco Wettstein 19. Oktober 2012

1. Kodieren Sie die gegebenen Zeichen mit UTF-8.

   (a) „y“

   Lösung:
   Für UTF-8 gelten die folgende Kennzahlen, sowie: Folgebyte Präfix: 10

	1 Wort	2 Wörter	3 Wörter	4 Wörter
Präfix	0	110	1110	1111 0
Freie Bits	7	11	16	21

   $y \in$ ASCII : $\mathrm{y}_{\mathrm{ASCII}} = 79_{16} = 111\,1001_2$ (7 Bits) $\rightarrow \mathrm{y}_{\mathrm{UTF\text{-}8}} = 0111\,1001_2$

   (b) „€“

   Lösung:
   Für die Bits-Wörter Tabelle siehe Lösung (a).
   „€“ gehört nicht zum ASCII Zeichensatz. Gemäss Unicode-Tabelle:
   $€_{\mathrm{Unicode}} = 20AC_{16} = 10\,0000\,1010\,1100_{16}$ (14 Bits)
   Für 14 Bits werden gemäss Tabelle drei Wörter benötigt, das erste mit Präfix „1110“, die weiteren jeweils mit Präfix „10“. Das Euro-Zeichen benötigt nur 14 statt 16 Bits, deshalb wird nach dem ersten Präfix noch „00“ eingefügt:
   $\rightarrow €_{\mathrm{UTF\text{-}8}} = 1110\,0010 : 1000\,0010 : 1010\,1100_2$

2. Geben sei der folgende Code $K = \{0010, 1000, 100, 110\}$. Dieser Code ist weder zyklisch noch präfixfrei.

   (a) Warum ist K weder zyklisch noch präfixfrei?

   Lösung:
   Er ist nicht zyklisch, weil in 110 zweimal die 1 vorkommt, in 100 hingegen nur einmal.

   Er ist nicht präfixfrei, weil 100 den gleichen Anfang hat wie 1000. 100 ist in 1000 enthalten.

   (b) Welche Codewörter in K müssen mindestens entfernt werden, damit K präfixfrei ist?

   Lösung:
   Entweder 100 oder 1000.

   (c) Welche Codewörder in K müssen mindestens hinzugefügt werden, damit K ein zyklischer Code ist?

   Lösung:
   $0100, 0001, 010, 001, 101, 011$

3. Bestimmen Sie den Hamming-Abstand für folgende Paare von Wörtern.

   (a) „1110 0110“ und „1010 0111“ (binär betrachten)

   Lösung:
   Anmerkung: Für den Hamming-Abstand spielt die Anzahl der Buchstaben im Alphabet keine Rolle.
   $1\underline{1}10\,011\underline{0}$ und $1\underline{0}10\,011\underline{1}$ → Hamming Abstand = 2

   (b) „1334“ und „1332“ (als Ziffern betrachten)

   Lösung:
   Anmerkung: Für den Hamming-Abstand spielt die Anzahl der Buchstaben im Alphabet keine Rolle.
   $133\underline{4}$ und $133\underline{2}$ → Hamming-Abstand = 1

   (c) „Abba“ und „Baba“ (als Zeichen betrachten)

   Lösung:
   Anmerkung: Für den Hamming-Abstand spielt die Anzahl der Buchstaben im Alphabet keine Rolle.
   $\underline{\text{Ab}}\text{ba}$ und $\underline{\text{Ba}}\text{ba}$ → Hamming-Abstand= 2

4. Gegeben sei ein Alphabet $Z = A, B, C, D, E, F, G$ mit:
$p(A) = 0{,}36, p(B) = 0{,}22, p(C) = 0{,}18, p(D) = 0{,}18,$
$p(E) = 0{,}03, p(F) = 0{,}02, p(G) = 0{,}01$

(a) Bestimmen Sie die Entropie $(\sum_{i=1}^{n} p_i \cdot I_i)$ eines Zeichens von Z

Lösung:

$$H_Z = \sum_{i=1}^{n} p_i \cdot I_i = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

$$= (0{,}36 \cdot \log_2 0{,}36) + (0{,}22 \cdot \log_2 0{,}22) + (0{,}18 \cdot \log_2 0{,}18) + (0{,}18 \cdot \log_2 0{,}18) + (0{,}03 \cdot \log_2 0{,}03) + (0{,}02 \cdot \log_2 0{,}02) + (0{,}01 \cdot \log_2 0{,}01) \approx 2{,}2329$$

Die Entropie ist für alle Zeichen gleich, nämlich ungefähr 2,2329

(b) Geben Sie eine mögliche Kodierung für die Zeichen von Z unter Verwendung

i. einer Gleichverteilung

Lösung:
Das Alphabet besteht aus 7 Zeichen, also werden aufgrund von $\lceil n \rceil = \log_2 |Z|$ mit $|Z| = 7$ mindestens 3 Bits benötigt. Eine mögliche Kodierung ist dann:

A = 001	B = 010	C = 011	D = 100
E = 101	F = 110	G = 111	

ii. dem Shannon-Fano Algorithmus

Lösung:

1. $G_{1A} = \{A, D\} : p = 0{,}54,$
$G_{1B} = \{B, C, E, F, G\} : p = 0{,}46$ (Erstes Bit)

2. $G_{2AA} = \{A\} : p = 0{,}36,$
$G_{2AB} = \{D\} : p = 0{,}18$
$G_{2BA} = \{B, G\} : p = 0{,}23,$
$G_{2BB} = \{C, E, F\} : p = 0{,}23$ (Zweites Bit)

3. $G_{3BAA} = \{B\} : p = 0{,}22, G_{2BAB} = \{G\} : p = 0{,}01$
$G_{3BBA} = \{C\} : p = 0{,}18, G_{3BBB} = \{E, F\} : p = 0{,}05$ (Drittes Bit)

4. $G_{4BBBA} = \{E\} : p = 0{,}03, G_{4BBBB} = \{F\} : p = 0{,}02$ (Viertes Bit)

Daraus folgt, als eine mögliche Kodierung:

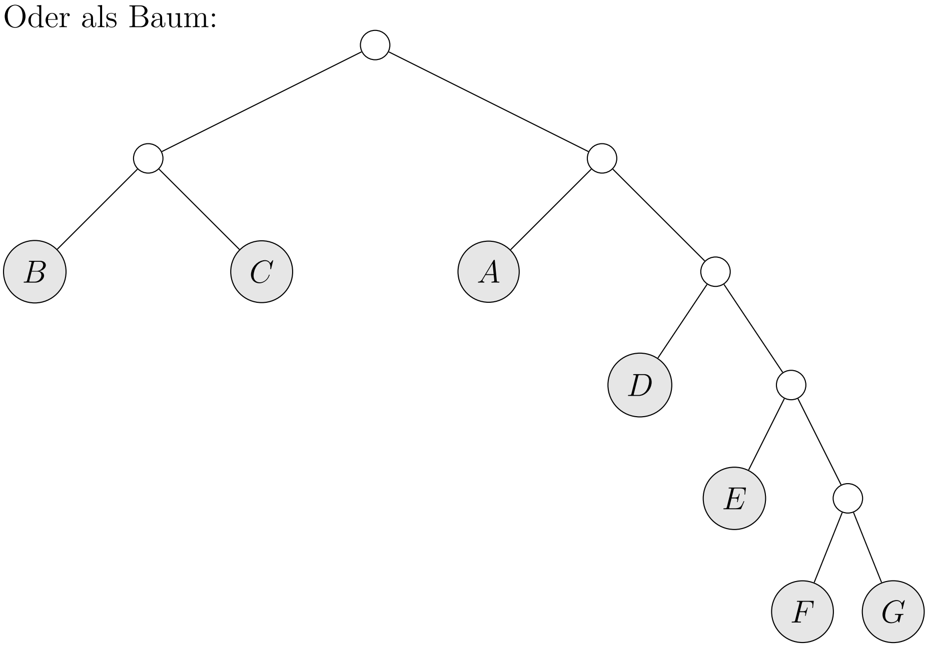
A = 00	B = 100	C = 110	D = 01
E = 1110	F = 1111	G = 101	

iii. dem Huffman Algorithmus an

Lösung:

1. $G_1 = \{F, G\} : p = 0{,}03$
2. $G_2 = \{E\{F, G\}\} : p = 0{,}06$
3. $G_3 = \{D, \{E\{F, G\}\}\} : p = 0{,}24$
4. $G_4 = \{B, C\} : p = 0{,}40$
5. $G_5 = \{A, \{D, \{E\{F, G\}\}\}\} : p = 0{,}60$
6. $G_6 = \{\{B, C\}, \{A, \{D, \{E\{F, G\}\}\}\}\} : p = 1{,}00$

Oder als Baum:



Daraus folgt, als eine mögliche Kodierung:

A = 10	B = 00	C = 01	D = 111
E = 1101	F = 11110	G = 11111	

5. Gegeben sei das Alphabet $Z = \{A, B, C\}$, mit $p(A) = 0{,}4$, $p(B) = 0{,}1$ und $p(C) = 0.5$

   (a) Geben Sie eine mögliche Codierung für das Wort „AAC“ mit der arithmetischen Kodierung an

   Lösung:

   1. Startintervall sei $I_0 = [0, 1)$
   2. Aufgrund der Häufigkeiten gelten folgende Teilintervalle:
      $I_a = [0, 0{,}4)$ $\quad I_b = [0{,}4, 0{,}5)$ $\quad I_c = [0{,}5, 1)$
   3. Erstes Zeichen: „A“ liegt im Interval $I_a = [0, 0{,}4)$
   4. I_a nach Häufigkeiten unterteilt ergibt:
      $I_a = [0, 0{,}16)$ $\quad I_b = [0{,}16, 0{,}2)$ $\quad I_c = [0{,}2, 0{,}4)$
   5. Zweites Zeichen: „A“ liegt im Interval $I_a = [0, 0{,}16)$
   6. I_a nach Häufigkeiten unterteilt ergibt:
      $I_a = [0, 0{,}064)$ $\quad I_b = [0{,}064, 0{,}08)$ $\quad I_c = [0{,}08, 0{,}16)$
   7. Drittes Zeichen: „C“ liegt im Intervall $I_c = [0{,}08, 0{,}16]$
   8. Kürzesten Wert aus $I_c = [0{,}08, 0{,}16]$ wählen: 0,1

   Eine mögliche Kodierung für „AAC“ ist 0,1.

   (b) Wieso gibt es mehrere / viele Möglichkeiten für eine Codierung von „AAC“ mit der arithmetischen Kodierung?

   Lösung:
   Bei der arithmetischen Kodierung werden Intervalle bestimmt um ein Wort zu kodieren. Das kodierte Wort liegt dann in einem berechneten Intervall. Da ein Intervall immer unendlich viele Werte enthält, gibt es unendlich viele mögliche Kodierungen.