Datenkompression

Sommersemester 2019 | **Übungsblatt 1** | Prof. Dr. Enno Ohlebusch

Institut für Theoretische Informatik | Uwe Baier

Ausgegeben am 30.04.2019 | Besprechung am 07.05.2019

Aufgabe 1.1

Wir betrachten die folgenden Mengen von Codewörtern:

$$M_1 := \{\texttt{01}, \texttt{10}, \texttt{000}, \texttt{011}, \texttt{100}, \texttt{111}\}, \qquad M_2 := \{\texttt{0}, \texttt{10}, \texttt{11}, \texttt{101}\},$$
$$M_3 := \{\texttt{00}, \texttt{01}, \texttt{11}, \texttt{100}, \texttt{101}\}, \qquad M_4 := \{\texttt{00}, \texttt{11}, \texttt{010}, \texttt{0100}\}.$$

(a) Welche Mengen erfüllen die Kraft-McMillan-Ungleichung?

(b) Welche Mengen sind Bildbereich eines eindeutig dekodierbaren Codes, und warum?

(c) Geben Sie für jede Menge M_i einen Präfixcode an, welcher dieselben Codewortlängen wie M_i hat. Für welche Mengen ist dies nicht möglich?

Aufgabe 1.2

Auf dem Alphabet $\{d, i, l, m, n, r, u\}$ sei eine Wahrscheinlichkeitsfunktion P definiert:

x	d	i	l	m	n	r	u
$P(x)$	1/24	1/24	1/6	7/24	1/12	1/24	1/3

(a) Bestimmen Sie einen Präfixcode nach dem Shannon-Verfahren.

(b) Bestimmen Sie einen Präfixcode nach dem Shannon-Fano-Verfahren (Variante mit Teilmengenbetrachtung, siehe Vorlesungsfolien H. Fernau, Vorlesung 2, Folie 15).

(c) Bestimmen Sie einen Präfixcode nach dem Huffman-Verfahren.

(d) Berechnen Sie die erwarteten Codewortlängen der Codes aus den Teilaufgaben (a) bis (c).

Aufgabe 1.3

Gegeben sei ein Text S der Länge n über einem Alphabet Σ der Größe $|\Sigma| = \sigma$. Weiter sei die Häufigkeit des Buchstabens mit Rang i gegeben durch c_i, es gilt $\sum_{i=1}^{\sigma} \mathsf{c}_i = n$. In der Vorlesung haben Sie den Begriff der Shannon-Entropie kennen gelernt:

$$H(S) \quad = \quad \sum_{i=1}^{\sigma} \frac{\mathsf{c}_i}{n} \log_2\left(\frac{n}{\mathsf{c}_i}\right) \quad \overset{P(i):=\frac{\mathsf{c}_i}{n}}{=} \quad \sum_{i=1}^{\sigma} P(i) \log_2\left(\frac{1}{P(i)}\right)$$

(a) Gegeben sei die Häufigkeitsverteilung $\mathsf{c}_1, \ldots, \mathsf{c}_\sigma$ wie am Anfang der Aufgabe beschrieben. Wie viele verschiedene Texte können mithilfe dieser Buchstabenhäufigkeiten erzeugt werden? Geben Sie eine allgemeine Formel in Abhängigkeit von n und $\mathsf{c}_1, \ldots, \mathsf{c}_\sigma$ an. Zeigen Sie, dass diese zu $\frac{n!}{\mathsf{c}_1! \cdot \mathsf{c}_2! \cdot \ldots \cdot \mathsf{c}_\sigma!}$ äquivalent ist.

(b) Leiten Sie mithilfe der Teilaufgabe (a), der Stirlingschen Abschätzung $\log_2(n!) \approx n \log_2(n)$ und folgendem Theorem die Shannon-Entropie her:

> **Theorem**
> Sei A eine Menge von Texten. Treten die Elemente in A mit gleicher Wahrscheinlichkeit auf, so benötigt ein beliebiger, eindeutig dekodierbarer Code für A (jedem Element wird eine Binärsequenz zugeordnet) im Mittel mindestens $\log_2(|A|)$ Bits um ein einzelnes Element aus A zu kodieren.