



Informationstheorie & Textcodierung



Teilnahme auf ALMA registrieren





BESPRECHUNG ÜBUNG 1



Informationstheorie

Diskrete gedächtnislose Quelle

→ Betrachten jedes Zeichen unabhängig von seinen Nachbarn

- Quelle Q : ANANAS
- Zeichenvorrat $X : \{A, N, S\}$
- Auftrittswahrscheinlichkeiten p : $p_A = \frac{3}{6}, p_N = \frac{2}{6}, p_S = \frac{1}{6}$
- Informationsgehalt $I_i = -\log_2 p_i$: $I_A = -\log_2 p_A$
- Entscheidungsgehalt $H_0 = \log_2 |X|$
- Entropie $H(Q) = -\sum p_i \log_2 p_i$
- Durschn. Wortlänge $\bar{L} = \sum p_i \cdot |c(x_i)|$ hier $\forall x_i: |c(x_i)| = 1$
- Redundanz $R_Q = H_0 - H(Q)$ oder $R = \bar{L} - H(Q)$



Entropiekodierungen

- ❖ Datensemantik wird ignoriert
- ❖ Verlustfreie Komprimierung
- **Laufängenkodierung** (RLE) ist nur gewinnbringend, wenn es viele gleiche Zeichen hintereinander in einem Wort gibt
- **Huffman-Kodierung** erzeugt einen optimalen Code mit möglichst kleiner mittlerer Codewortlänge
- **Burrows-Wheeler-Transformation** sortiert die Zeichen einer Quelle, so dass mit RLE komprimiert werden kann