Komplexität und natürliche Sprache

Übung 5 (Abgabe: 05.07.2017)

Timm Lichte & Christian Wurm

1 Informationsgehalt

1. Geben sie den Shannon'schen Informationsgehalt des Wortes *Rindfleischetikettierungsgesetz* in Bits an. Als Wahrscheinlichkeiten nehmen Sie einfache die relativen Häufigkeiten der Buchstaben im Wort selbst an; nehmen Sie weiterhin an, dass die Wahrscheinlichkeiten der einzelnen Buchstaben voneinander unabhängig sind.

2. Unter denselben Annahmen: gegeben $\overline{w_1}$ = *aaaaaaabbbbbb* und $\overline{w_2}$ = *baabbabaababa*, welches Wort hat den höheren Informationsgehalt?

Zur Erinnerung:
Der Informationsgehalt eines Buchstabens x in einer Wahrscheinlichkeitsverteilung p ist $I(x) = -\log_2(p_x)$. Der Informationsgehalt eines Wortes $\overline{w}$ ist, unter der Annahme, dass alle Buchstabenwahrscheinlichkeiten im Wort voneinander unabhängig sind, $I_{\overline{w}} = I(x_1) + I(x_2) + \ldots + I(x_n) = \sum_{i=1}^{n} I(x_i)$ mit $\overline{w} = x_1 \ldots x_n$.