Musterlösung: Statistische Sprachverarbeitung 2

February 12, 2008

1. Aufgabe
   - $P(\text{a}) = \frac{10080}{665475} = 0.01515$
   - $P(\text{of}) = \frac{14825}{665475} = 0.02228$
   - $P(,) = \frac{39739}{665475} = 0.05972$
   - $P(\text{and}) = \frac{17944}{665475} = 0.02696$
   - $P(") = \frac{39739}{665475} = 0.05972$
   - $P(\text{in}) = \frac{8370}{665475} = 0.01258$
   - $P(\text{to}) = \frac{16435}{665475} = 0.02470$
   - $P(\text{the}) = \frac{31784}{665475} = 0.04776$
   - $P(\text{OTHER\_WORD}) = \frac{474588}{665475} = 0.71316$
2. Aufgabe

$$
\begin{aligned}
H(X) &= -\sum_{i=1}^{n} P(x_i) * log_2(P(x_i)) \\
H(X) &= -(P(a) * log_2(P(a) + \\
&\quad P(of) * log_2(P(of) + ... + \\
&\quad P(OTHER\_WORD) * log_2(P(OTHER\_WORD))) \\
H(X) &= 1.72806
\end{aligned}
$$

3. Aufgabe
   Anzahl benötigter Bits:
   $I(X) = -log_2(P(X))$
   Gerundete Bits: Bits(WORD)

Word	$I(WORD)$	Bits(WORD)	Bits
a	-6.04481692526185	6	011111
of	-5.4882804593284	6	011110
,	-4.06575699531665	4	0010
and	-4.97749781568365	4	0000
”	-5.21281103856286	6	011100
.	-4.44339395106976	4	0011
in	-6.31301303621142	6	011011
to	-5.33954110768146	6	010111
the	-4.38801186556103	4	0001
OTHER$_W ORD$	-0.487708846892341	1	009

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} P(x_i) * I(P(x_i)) \\ H(X) &= 2.159342 \end{aligned}$$

4. Aufgabe
   Länge nach Kompression:

$$\begin{aligned} L(X) &= -\sum_{i=1}^{n} P(x_i) * Bits(x_i) \\ L(X) &= 1383524 \end{aligned}$$

5. Aufgabe
   siehe 1. Aufgabe

6. Aufgabe

$$\begin{aligned} D_{KL}(P||Q) &= \sum_{i=1}^{n} P(x_i) * log_2(\frac{P(x_i)}{Q(x_i)}) \\ &= 0.05091 \end{aligned}$$