

# Text Indexing and Information Retrieval

## Übungsblatt 3

Besprechung: 27.10. 2014

### Aufgabe 1 (Theorie)

Sei  $\Sigma$  ein Alphabet und  $h$  eine Hashfunktion für Strings der Länge  $\ell$  über  $\Sigma$ . Man nennt  $h$  eine „rolling hash function“, falls man  $h(s_2s_3 \dots s_\ell + 1)$  aus  $h(s_1s_2 \dots s_\ell)$  in  $O(1)$  Zeit berechnen kann. Zeigen Sie, dass  $h(t_1t_2 \dots t_\ell) = \left( \sum_{i=1}^{\ell} a^{\ell-i} \cdot t_i \right) \bmod p$  für eine Primzahl  $p > |\Sigma|$  und  $a \in [1, p)$  eine „rolling hash function“ ist.

### Aufgabe 2 (Praxis)

Schreiben Sie ein Programm/Skript, welches für eine gegebene Maximalzahl  $n$  für alle Zahlen  $x$  mit  $1 \leq x \leq n$  die Bitlängen der Codierungen  $(x)_1$ ,  $(x)_2$ ,  $(x)_\gamma$ ,  $(x)_\delta$ ,  $(x)_3$ ,  $(x)_\phi$  und  $(x)_{\text{Gol}(b)}$  (Parameter  $b$ ) berechnet und tabellarisch ausgibt. Plotten Sie diese Ergebnisse in einen Graphen, z.B. mit gnuplot.

### Aufgabe 3 (Theorie)

Finden Sie Formeln für die Bit-Längen der Kodierungen  $(x)_\phi$  und  $(x)_{G(b)}$ .

### Aufgabe 4 (Theorie)

- Definieren Sie einen Quartärcode  $(x)_4$  für die Repräsentation einer Zahl  $x$ , so dass der Code präfixfrei ist.
- Vergleichen Sie Ihren Code mit dem Ternärcode. Welchen würden Sie bevorzugen?

c) Definiere  $E(b, x) = b(\lfloor \log_b x \rfloor + 1)$ . Zeigen Sie, dass  $E(\cdot, x)$  bei  $b = e$  ein Minimum aufweist. Hinweis: Zeigen Sie, dass  $x/\ln x$  bei  $x = e$  ein Minimum aufweist. Was folgern Sie aus dieser Tatsache für den Vergleich von Ternär- und Quartärcode?