Algorithmische Grundlagen und Vermittlung der Informatik
Lehrstuhl 11 (Algorithm Engineering)
Prof. Dr. Johannes Fischer

technische universität dortmund

Text Indexing and Information Retrieval

Übungsblatt 4

Besprechung: 3.11. 2014

Aufgabe 1 (Praxis)

Implementieren Sie ein oder mehrere Kompressionsverfahren für die Postings Lists aus Übungsblatt 2, Aufgabe 1. Welche Platzersparnis erreichen Sie damit für Ihren Invertierten Index?

Aufgabe 2 (Theorie + Praxis)

a) Ein naives Suffix-Array Konstruktionsverfahren wäre es, einfach einen beliebigen Standard-Sortieralgorithmus (z.B. aus der Java- oder C++-Bibliothek) zu nehmen. Welche theoretische Laufzeit erreichet man hiermit im besten Fall? Implementieren Sie dieses Verfahren und messen Sie die Laufzeit anhand von Dateien auf `http://pizzachili.dcc.uchile.cl/texts.html`.

b) Ebenso kann man das LCP-Array H auf naive Art und Weise erstellen, indem man die Formel $H[i] = \max\{h \geq 0 : T[A[i], ..., A[i]+h-1] = T[A[i-1], ..., A[i-1]+h-1]\}$ für alle Werte $1 < i \leq n$ anwendet. Welche theoretische Laufzeit ergibt dies? Implementieren Sie das Verfahren ebenfalls und testen Sie die Laufzeit.

c) Ebenso ließe sich der Suffixbaum naiv konstruieren, indem man einfach der Reihe nach die Suffixe $T^1, T^2, \ldots, T^n$ in den bereits bestehenden Trie einfügt (statt in der Reihenfolge $T^{A[1]}, T^{A[2]}, \ldots, T^{A[n]}$ wie in der VL). Welche theoretische Laufzeit erreicht man hiermit?

Aufgabe 3 (Theorie)

Zeigen oder widerlegen Sie: wenn im LCP-Array der Wert ℓ (an einer beliebigen Stelle) auftritt, dann tritt auch der Wert $\ell - 1$ (an einer beliebigen anderen Stelle) auf.

Aufgabe 4 (Theorie)

Entwerfen Sie einen Linearzeitalgorithmus, der für einen Text T das *kürzeste* Teilwort findet, das nur einmal in T vorkommt. Hinweis: Suffixbäume!