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: 5.11.2018

Aufgabe 1 (Praxis)

Suffix Arrays für natürlichsprachliche Texte können auch wortbasiert sein: sortiere nur die Textpositionen, an denen ein Wort beginnt (also z.B. nach jedem Whitespace und Satzzeichen). Implementieren Sie ein solches Verfahren (z.B. mit Hilfe Ihrer bisherigen Implementierungen oder der Implementierung von sais aus dem letzten Aufgabenblatt) und vergleichen Sie Platz- und Zeitbedarf mit "herkömmlichen" Suffix Arrays. Testen Sie *verschiedene* Textarten von `http://pizzachili.dcc.uchile.cl/texts.html`.

Aufgabe 2 (Praxis)

Im Skript steht in Abschnitt 3.7.1 eine praktische Verbesserung des $O(n)$-LCP-Algorithmus (Reduktion der Cache-Misses). Implementieren Sie die beiden $O(n)$-LCP-Algorithmen und vergleichen Sie die praktische Laufzeit anhand der Texte von Pizza&Chili (s. Aufgabe 1).

Aufgabe 3 (Theorie)

a) Zeigen oder widerlegen Sie: wenn im LCP-Array an einer Stelle i ein Wert $H[i] = \ell \geq 1$ auftritt, dann gibt es auch eine Stelle j mit $H[j] = \ell - 1$.

b) Sei H das LCP-Array von $T\$$, und H^R das LCP-Array des reversen Textes $T^R := T[n]T[n-1]\ldots T[1]\$$. Zeigen Sie, dass H^R eine Permutation von H ist, d.h.: $\exists\ i \in [1,n]$ s.d. $H[i] = \ell \iff \exists\ i' \in [1,n]$ s.d. $H^R[i'] = \ell$.

Aufgabe 4 (Theorie)

Bei der LZ77-Definition im Skript (Def. 10) kann das vorherige Vorkommen ja in den aktuellen Faktor hineinragen. Manchmal ist jedoch genau das nicht gewünscht.

a) Überlegen Sie sich eine Definition einer so modifizierten LZ77-Zerlegung.

b) Passen Sie den $O(n)$-Zeit-Algorithmus aus der Vorlesung entsprechend an.