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

technische universität dortmund

Text Indexing and Information Retrieval

Übungsblatt 5

Besprechung: 10.11.2014

Aufgabe 1 (Praxis)

Suffix Arrays für natürlichsprachliche Texte können auch wortbasiert sein: sortiere nur die Textindizes, an denen ein Wort beginnt (also z.B. nach jedem Whitespace). Implementieren Sie ein solches Verfahren (etwas auf Basis des letzten Übungsblattes, Aufgabe 2) und testen Sie die Konstruktion (etwa auf den dort angegebenen Dateien).

Aufgabe 2 (Theorie)

Entwerfen Sie einen Linearzeit-Algorithmus, der für 2 Texte T_1 und T_2 das längste Teilwort findet, das sowohl in T_1 als auch in T_2 vorkommt.

Aufgabe 3 (Theorie)

Beschreiben Sie einen Algorithmus in Pseudocode, der die S*-substrings eines Textes $T[1, n]$ mittels Radix-Sort in Linearzeit sortiert (nehmen Sie hierfür an, dass die Alphabetgröße σ höchstens n ist).

Aufgabe 4 (Theorie)

Entwerfen Sie einen Text-Index linearer Größe, der für ein Muster $P_{1\ldots m}$ ein Array $V[1, m]$ ausgibt, so dass $V[i]$ das längste Präfix von $P_{i\ldots m}$ angibt, das in T vorkommt. Die erwartete Laufzeit soll $O(m)$ sein. Hinweis: Suffixbäume mit entsprechender Zusatzinformation!