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: 21.11.2016

Aufgabe 1 (Praxis)

Implementieren Sie den in der Vorlesung besprochenen (einfachen) $O(n \lg \sigma)$-Zeit-Algorithmus für LZ78 (d.h. den LZ78-Trie aufbauen, in dem die ausgehenden Kinder geeignet verwaltet werden). Finden Sie ebenfalls eine geeignete Kodierung der LZ78-Faktoren und schreiben Sie den so komprimierten Text in eine Textdatei.

Testen Sie Ihren Algorithmus auf *verschiedenen* Textarten von `http://pizzachili.dcc.uchile.cl/texts.html` und vergleichen Sie die Kompressionsrate mit üblichen Kompressionstools wie gzip, bzip2, etc.

Aufgabe 2 (Theorie)

Führen Sie die in der Vorlesung besprochenen Linearzeit-Algorithmen für LZ77 und LZ78 auf dem Text `inulmundumulmherum$` aus.

Aufgabe 3 (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 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!