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: 12.11.2018

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 (möglichst kleine!) 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 und Praxis)

a) Überlegen Sie sich, wie das LCP-Array als Maß für die Komprimierbarkeit von Texten benutzt werden kann.

b) Implementieren Sie Ihr Kompressionsmaß und fertigen Sie verschiedene "LCP-Statistiken" für die Texte auf `http://pizzachili.dcc.uchile.cl/texts.html` an.

c) Skizzieren Sie einen Kompressionsalgorithmus, der direkt mit dem LCP-Array arbeitet und LZ77-Faktoren oder LZ-ähnliche Faktoren erzeugt.

Aufgabe 3 (Theorie)

Der $O(n \lg n)$-Suchalgorithmus aus der Vorlesung (**Korrektur im Skript fett!**) braucht RMQ-Information auf dem LCP-Array. Es werden bei der binären Suche aber nur bestimmte RMQs ausgeführt. Diese könnten auch alle in $O(n)$ Platz vorberechnet werden. Zeigen Sie, wie dies geht.

Aufgabe 4 (Theorie)

a) Zeigen Sie, wie die RMQ-Datenstruktur aus Abschnitt 5.1 im Skript in optimaler Zeit vorberechnet werden kann.

b) Entwerfen Sie eine einfache Datenstruktur linearer Größe, die RMQs in $O(\log n)$ Zeit beantworten kann.