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

technische universität dortmund

Text Indexing and Information Retrieval

Übungsblatt 8

Besprechung: 3.12.2018

Aufgabe 1 (Praxis)

a) Informieren Sie sich unter `https://graphics.stanford.edu/~seander/bithacks.html` über die Möglichkeiten, in einem (32- oder 64-Bit) Wort die Anzahl der Einsen zu zählen. Implementieren Sie mit mindestens zwei dieser Ideen einfache rank-Datenstrukturen, z.B. nur mit dem Array M und Block-Größe $s = 32$ oder $s = 64$ (je nachdem, wie Sie die Einsen in einem Wort zählen). Vergleichen Sie die Laufzeit der daraus resultierenden Verfahren.

b) Implementieren Sie einen Konstruktionsalgorithmus für Wavelet-Bäume. Wenden Sie für das Zählen der Einsen dann die Lösung aus (a) an. Implementieren Sie auch den rank-Algorithmus für Wavelet Trees und testen Sie seine Laufzeit, etwa durch viele zufällige rank-Anfragen. Vergleichen Sie die Laufzeit von rank-Anfragen mit random-access-Anfragen auf den Textzeichen.

Aufgabe 2 (Theorie)

Zeigen Sie alle Datenstrukturen, die für die $O(m \log \sigma)$-Rückwärtssuche auf dem Text

$$T = \texttt{missmississippi\$}$$

benötigt werden. Führen Sie beispielhaft die Rückwärtssuche für das Pattern $P = \texttt{is}$ durch.

Aufgabe 3 (Theorie)

Sei $D[1, n]$ ein Array der Länge n mit Werten aus $[1, r]$. Zeigen Sie eine Datenstruktur der Größe $O(n \log r)$ Bits, die es ermöglicht, folgende Anfragen in $O(\log r)$ Zeit zu beantworten: für gegebene Indizes $1 \leq \ell \leq r \leq n$ und einen Wert $k \in [1, r - \ell + 1]$, finde das k-t-größte Element in $D[\ell, r]$. Hinweis: Wavelet-Bäume.

Aufgabe 4 (Theorie)

Sei S eine Menge von n Punkten auf einem $(n \times n)$-Gitter gegeben, so dass keine 2 Punkte die gleiche x-Koordinate haben. Entwerfen Sie eine Datenstruktur der Größe $O(n \log n)$ Bits, mit der Sie *4-seitige Bereichsanfragen* beantworten können. Solche Anfragen sollen für 4 der Anfrage übergebene Koordinaten $x_\mathrm{l}, x_\mathrm{r}, y_\mathrm{t}$ und y_b alle k Punkte aus S ausgeben, die in $[x_\mathrm{l}, x_\mathrm{r}] \times [y_\mathrm{b}, y_\mathrm{t}]$ liegen. Die Anfragezeit soll $O(k \log n)$ sein. Hinweis: Wavelet-Bäume.