

Text Indexing and Information Retrieval

Übungsblatt 1

Besprechung: 13.10.2014

Aufgabe 1 (Theorie)

Erstellen Sie einen Trie für die Wörter {aardvark, affenpinscher, ant, baboon, bat, bonobo, goat, goose, gorilla, grasshopper, zebra}. Zeigen Sie zumindest für die Wurzel die verschiedenen Varianten, wie die ausgehenden Kanten repräsentiert werden können.

Aufgabe 2 (Praxis)

Implementieren Sie einen Trie, sodass damit Strings lexikographisch sortiert werden können. Verwenden Sie dabei eine einfache Repräsentation der Trie-Knoten aus der Vorlesung.

Testen Sie die Laufzeit mit dem auf der Webseite zur Verfügung gestellten Strings und vergleichen Sie sie mit `gnu sort`.

Aufgabe 3 (Theorie)

Analysieren Sie die Laufzeit für das Sortieren von k Strings der Gesamtlänge n mit Hilfe von Tries. Nehmen Sie an, dass die ausgehenden Kanten an internen Knoten mit

- a) balancierten binären Suchbäumen
- b) Hashtabellen

implementiert sind. Die Alphabetgröße ist σ . Kennen Sie *theoretisch schnellere* Sortierverfahren für Strings?

Aufgabe 4 (Theorie)

Tries beinhalten „automatisch“ eine Art der Kompression: gemeinsame Präfixe werden nur einmal gespeichert. Gibt es auch eine Möglichkeit, gemeinsame *Suffixe* zu komprimieren?