Sprachtechnologie

6. Übung

Prof. Tobias Scheffer
Peter Haider
Uwe Dick
Paul Prasse

Sommer 2012

Ausgabe am: 04.06.12
Besprechung am: 11.06.12

Aufgabe 1 *Indexieren und Suchen*

Simulieren Sie, wie der auf Folien 9 und 10 vorgestellte Algorithmus einen invertierten Index über alle Terme des folgenden Textes aufbaut. Berücksichtigen Sie dabei, dass nur 10 Indexterme gleichzeitig in den Hauptspeicher Ihres Rechners passen. Der Text lautet:

Ich fliege mit meiner Rakete zum Mars. Ich besuche dort die auf dem Mars lebenden Marsmenschen.

- Simulieren Sie, wie Sie mit Hilfe des Index' alle Fundstellen des Wortes *Mars* suchen.
- Suchen Sie mit Hilfe des Index' nach Textstellen, an denen die Phrase *Rakete zum Mars* vorkommt.
- Konstruieren Sie einen Suffix-Baum über den oben stehenden Text auf. Verwenden Sie alle Wortanfänge als Indexpositionen. Simulieren Sie, wie der Algorithmus, den wir in der Vorlesung kennen gelernt haben, dabei vorgeht.
- Konstruieren Sie aus dem Suffix-Baum im nächsten Schritt einen Suffix-Array.
- Suchen Sie mit Hilfe des Suffix-Baumes und anschließend mit Hilfe des Suffix-Arrays nach Vorkommen des Strings *Rakete zum Mars*.

Aufgabe 2 *Knuth-Morris-Pratt*

Überprüfen Sie mit Hilfe des Suchalgorithmus von Knuth-Morris-Pratt, ob das Wort *ababac* in dem Wort *baaababacd* vorkommt. Zeigen Sie, dass der Vorlaufalgorithmus eine Laufzeit von $\mathcal{O}(m)$ und der Suchalgorithmus eine Laufzeit von $\mathcal{O}(n)$ besitzt, wobei m die Länge des Suchwortes und n die Länge des zu durchsuchenden Wortes ist.

Aufgabe 3 *Aho-Corasick-Trie*

Konstruieren Sie einen Aho-Corasick-Trie, der alle Fundstellen der Suchstrings a, $abcc$, ca, bc, aa in einer Zeichenkette findet. Identifizieren Sie anschließend mit diesem Aho-Corasick-Trie alle Fundstellen dieser Suchstrings in der Zeichenkette *aabccab*.