Algorithmen und Datenstrukturen II
SS 2014 – Serie 4

12 (6 Punkte) DAWG

Konstruieren sie einen DAWG aus den Wörtern SENDUNG, ENDUNG, SENDEN, ENDEN. Gehen sie dazu vor genau wie in der Vorlesung beschrieben. Erstellen sie also zuerst einen Trie aus den Wörtern und zeichnen sie diesen. Verschmelzen sie dann schrittweise die Wortenden. Zeichnen sie den sich ergebenden DAWG.

Verwenden sie jeweils eine Darstellung analog zu der auf den Vorlesungsfolien, d.h. unterscheiden sie gerichtete Child- und Next-Kanten und markieren die EOW-Knoten eindeutig.

13 (15 Punkte) Textsuche

Gegeben sind der Text

$$t[\,] = \texttt{DUDELDUDIDLDADELDUDADEI}$$

und die Muster

$$q_1[\,] = \texttt{DUDA},\ q_2[\,] = \texttt{DADEL}\ .$$

(a) Wenden Sie den Algorithmus von Knuth-Morris-Pratt an, um in t nach q_1 zu suchen. Geben Sie dazu zuerst die next-Tabelle für q_1 an. Beschreiben Sie den Verlauf der Suche, indem Sie die vom Algorithmus (Folie V7-8) verglichenen Paare (i, j) mit je einem Punkt in einem Diagramm darstellen (wie auf Folie V7-9, aber für den konkreten Fall). Beschriften Sie die Achsen mit Positionsnummern und zeichnen sie so präzise, daß die angenommenen (i, j)-Paare eindeutig ablesbar sind. Wieviele Vergleiche zwischen Text und Muster werden ausgeführt? (5 Punkte)

(b) Verwenden Sie den Algorithmus von Boyer-Moore (in seiner einfachsten Form laut Vorlesung V7-11) um in t nach q_1 und q_2 zu suchen. Geben Sie für jedes dieser beiden Muster die last-Tabelle (für dessen Zeichen) an. Notieren Sie die Folge der Textpositionen i, an die das Muster "angelegt" wird. (6 Punkte)

Achtung: Der BM-Algorithmus war ursprünglich falsch auf Folie V7-11 angegeben. Richtig muss es heissen "Setze last[c]:=0 (statt last[c]:=-1) falls das Symbol c nicht im Muster vorkommt".

c) Eine Signaturfunmktion h sei gegeben als

$$h(c_1c_2c_3c_4) = \left[\sum_{i=1}^{4} \mathrm{ord}(c_i)\right] \bmod 10.$$

und ord(c) sei die Nummer von c im Alphabet, d.h. ord(A) = 1, ord(B) = 2, ...

Geben sie die Signaturfunktion h für Muster q_1 und jedes Fenster $t[i..i+3]$ der Länge 4 des Texts für $i = 1 \ldots 10$ an. Wie oft muss bei Verwendung dieser Signaturfunktion das Muster q_1 zeichenweise mit einem dieser ersten 10 Fenster des Texts t verglichen werden. (4 Punkte)

14 (9 Punkte) Editierdistanz

Gegeben sind die Zeichenketten

$U[] =$ `FLOTTE`
$V[] =$ `LOTTA`
$W[] =$ `FLOETE`

Berechnen Sie alle drei paarweisen Editierdistanzen, also zwischen U und V, zwischen V und W, zwischen W und U. Geben Sie jeweils die Matrix D_{ij} und ein optimales Alignment an. Benutzen Sie das Einheitskostenmodell.