Proseminar String Matching

PD Dr. habil. Hanjo Täubig

Lehrstuhl für Theoretische Informatik
(Prof. Dr. Susanne Albers)
Institut für Informatik
Technische Universität München

Wintersemester 2016/17

Übersicht

Übersicht

Formalitäten

- Modul: IN0013
- Bereich:
  Proseminar im Bereich Informatik III (Theoretische Informatik)
- Semesterwochenstunden:
  2 SWS Proseminar
- ECTS: 4 Credits
- Seminarzeiten:

  Dienstag 14:15 – 15:45 Uhr (MI 03.11.018)
- Web:
  http://www14.in.tum.de/lehre/2016WS/proseminar/

Dozent

- PD Dr. habil. Hanjo Täubig
  (Lehrstuhl für Effiziente Algorithmen / Prof. Mayr)
  (Lehrstuhl für Theoretische Informatik / Prof. Albers)
- eMail:
  taeubig@in.tum.de
- Web:
  http://www14.in.tum.de/personen/taeubig/
- Telefon: 089 / 289-17740
- Raum: MI 03.09.039

- Sprechstunde: vorzugsweise 13-14 Uhr
  (bzw. nach Vereinbarung)

Ziele des Proseminars

- Literaturrecherche
  - Bibliothek: Buch ausleihen / Online-Zugriff
  - Web: Google Scholar, Microsoft Academic Search, etc.
- Vortrag
  - Erstellung der Slides mit LaTeX (beamer package), PowerPoint oder mit einem ähnlichen Programm nach Wahl
  - Halten des Vortrags (ca. 45 min)
- Mitarbeit / Diskussion
- Ausarbeitung
  - Erstellung mit LaTeX
  - ca. 6 Seiten Text (plus Abbildungen, plus Literaturverzeichnis)

(Bewertung)

Übersicht

Alphabet

Definition

Ein Alphabet ist eine endliche Menge von Symbolen.

Wir bezeichnen es meistens mit Σ.

Beispiel

$\Sigma = \{a, b, c, \ldots, z\}$, $\Sigma = \{0, 1\}$, $\Sigma = \{A, C, G, T\}$

Wörter

Definition

Wörter über Σ sind endliche Folgen von Symbolen aus Σ.

Wir notieren Wörter meist in der Form
$w = w_0 \cdots w_{n-1}$ oder $w = w_1 \cdots w_n$.

Beispiel

$\Sigma = \{a, b\}$, dann ist $w = abba$ ein Wort über Σ.

Wortlänge, Wortmengen

Definition

Die Länge eines Wortes w wird mit $|w|$ bezeichnet und entspricht der Anzahl der Symbole in w.

Das Wort der Länge 0 heißt leeres Wort und wird mit ε bezeichnet.

Definition

Die Menge aller Wörter über Σ wird mit Σ^* bezeichnet.

Die Menge aller Wörter der Länge größer gleich 1 über Σ wird mit $\Sigma^+ := \Sigma^* \setminus \{\varepsilon\}$ bezeichnet.

Die Menge aller Wörter über Σ der Länge k wird mit $\Sigma^k \subseteq \Sigma^*$ bezeichnet.

Präfix, Suffix, Teilwort

Definition

Im Folgenden bezeichne $[a : b] = \{n \in \mathbb{Z} \mid a \leq n \wedge n \leq b\}$ für $a, b \in \mathbb{Z}$.
Sei $w = w_1 \cdots w_n$ ein Wort der Länge n über Σ, dann heißt

- w' Präfix von w, wenn $w' = w_1 \cdots w_\ell$ mit $\ell \in [0 : n]$
- w' Suffix von w, wenn $w' = w_\ell \cdots w_n$ mit $\ell \in [1 : n + 1]$
- w' Teilwort von w, wenn $w' = w_i \cdots w_j$ mit $i, j \in [1 : n]$

(Für $w' = w_i \cdots w_j$ mit $i > j$ soll gelten $w' = \epsilon$.)

Das leere Wort ε ist also Präfix, Suffix und Teilwort eines jeden Wortes über Σ.

Textsuche

Problem:

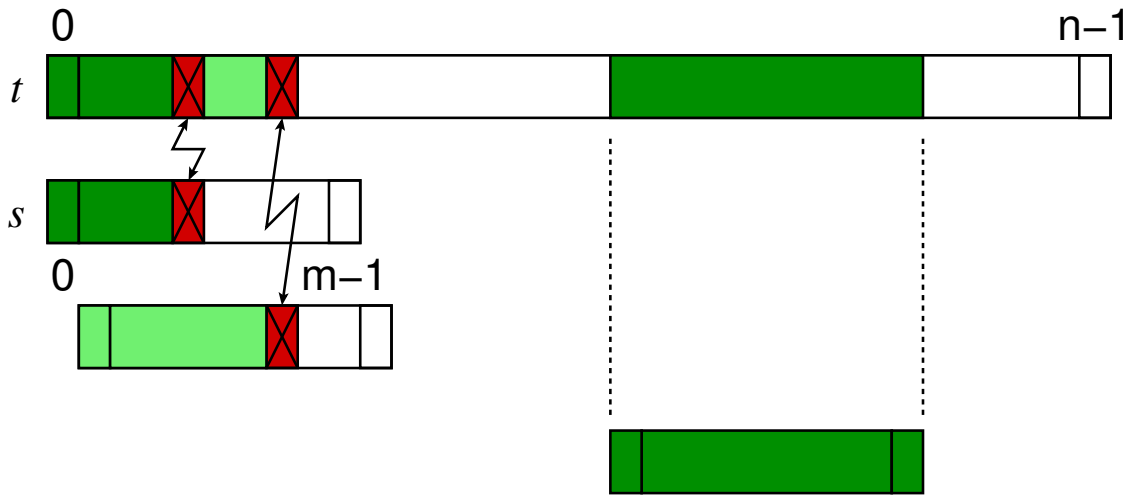
Gegeben: Text $t \in \Sigma^*$; $|t| = n$;
Suchwort $s \in \Sigma^*$; $|s| = m \leq n$

Gesucht: $\exists\ i \in [0 : n - m]$ mit $t_i \cdots t_{i+m-1} = s$?
(bzw. alle solchen Positionen i)

Naiver Algorithmus

- Suchwort s wird Buchstabe für Buchstabe mit dem Text t verglichen
- Stimmen zwei Buchstaben nicht überein ($\rightarrow$ Mismatch), so wird s um eine Position „nach rechts“ verschoben und der Vergleich von s mit t beginnt von neuem.
- Vorgang wird wiederholt, bis s in t gefunden wird oder bis klar ist, dass s in t nicht enthalten ist.

Naiver Algorithmus



Naiver Algorithmus: Implementation

Algorithmus 1 : bool Naiv (char $t[]$, int n, char $s[]$, int m)

```
int i := 0, j := 0;
while (i ≤ n − m) do
    while (t[i + j] = s[j]) do
        j++;
        if (j = m) then
            return TRUE;
    i++;
    j := 0;
return FALSE;
```

Match und Mismatch

Definition

Stimmen zwei Zeichen bei einem Vergleich überein, so nennt man dies ein Match, ansonsten ein Mismatch.

Analyse des naiven Algorithmus

- zähle Vergleiche von Zeichen,
- äußere Schleife wird $(n - m + 1)$-mal durchlaufen,
- die innere Schleife wird maximal m-mal durchlaufen.
- maximale Anzahl von Vergleichen: $(n - m + 1)m$,
- Laufzeit: $\mathcal{O}(nm)$