

ALGORITHMEN UND DATENSTRUKTUREN

ÜBUNG 9: SUCHEN & ERSETZEN

Eric Kunze

eric.kunze@mailbox.tu-dresden.de

TU Dresden, 08.01.2021

KMP-Algorithmus

Aufgabe 1

KMP-ALGORITHMUS

- Mustersuche in (großen) Texten
- Ziel: Verschiebung des Musters um mehr als eine Position bei Nichtübereinstimmung.
- Methode: Ermittlung einer Verschiebetabelle Tab[] inPhase 1
- Bedeutung des Eintrags Tab[i]=j:
 Bei Nichtübereinstimmung an Stelle i wird Position j des Musters an aktueller Vergleichsstelle angelegt.
- Suchprozess in Phase 2

j-algo: http://j-algo.binaervarianz.de/

KMP-ALGORITHMUS

Suche das Muster aaabaaaa im Text aaabaaabaaacaaabaaaa.

Position	0	1	2	3	4	5	6	7
Pattern								
Tabelle	-1	-1	-1	2	-1	-1	-1((3)

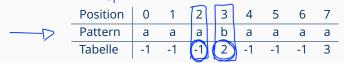
Erster Versuch:

Tabelleneintrag an Position 7 ist 3, d.h. Tab[7]=3 — Lege Position 3 des Musters an aktueller Vergleichsposition an:

Gleicher Prozess noch einmal: Missmatch an Position 7 des Musters — verschiebe Muster auf Position 3.

KMP-ALGORITHMUS (FORTSETZUNG)

Suche das Muster aaabaaaa im Text aaabaaabaaacaaabaaaa.



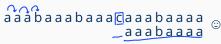
Wir legen das Muster also wieder an Position 3 an:

Tab = 0Tab = -1

Wegen Tab [3] = 2, lege Muster an Position 2 and Positio

aaabaaabaaa<mark>c</mark>aaabaaaa aa<mark>a</mark>baaaa

Wegen Tab [2] =-1, lege Muster an Position -1 an:



KMP-ALGORITHMUS — DIE ZYKLENMETHODE

Zwei Phasen:

KMP-ALGORITHMUS — DIE ZYKLENMETHODE

Zwei Phasen:

- ▶ 1. Phase: Markieren der längsten Teilwörter im Pattern, die mit einem Präfix übereinstimmen
 - ▷ ein Zyklus beginnt an einer Patternposition i falls i ≠ 0 und Pat[0] = Pat[i]
 - ein Zyklus endet an der kleisten Patternposition i+m, sodass Pat [m+1] → Pat [i+m+1]



KMP-ALGORITHMUS — DIE ZYKLENMETHODE

Zwei Phasen:

- ▶ 1. Phase: Markieren der längsten Teilwörter im Pattern, die mit einem Präfix übereinstimmen
 - ▷ ein Zyklus beginnt an einer Patternposition i falls i ≠ 0 und Pat[0] = Pat[i]
 - ▷ ein Zyklus endet an der kleisten Patternposition i+m, sodass Pat [m+1] ≠ Pat [i+m+1]
- 2. Phase: Bestimmung der Tabelleneinträge
- \rightarrow \triangleright Tab[0] = -1
 - Tabelleneinträge nach einem Zyklus:
 Länge des längsten dort endenden Zyklus



- ▶ Tabelleneinträgen in einem Zyklus: Tabelleneintrag der derzeitigen Position im <u>längsten</u> laufenden Zyklus
- ▶ verbleibende Einträge: 0

AUFGABE 1 — ZYKLENMETHODE

Teil (a) Pattern: aabaaacaab

Position	0	1	2	3	4	5	6	7	8	9
Pattern	а	а	b	а	а	а	С	а	а	b
Tabelle	-1	-1	1	-1	-1	2	2	-1	-1	1
				Fo				0	4	2
						H			1	*

AUFGABE 1 — ZYKLENMETHODE

eil (a)	Pattern: aabaaacaab									
Position	0	1	2	3	4	5	6	7	8	9
Pattern	a	а	b	a	а	а	С	a	а	b
Tabelle	-1	-1	1	-1	-1	2	2	-1	-1	1

KMP-ALGORITHMUS — DIE ZWEI-FINGER-METHODE

Die Methode beruht auf der Gleichung

$$\operatorname{Tab}[\mathtt{i}] = \max_{\longleftarrow} \{-1\} \cup \left\{ m \middle| \begin{array}{ccc} 0 \leq m \leq i-1 \\ b_0 \dots b_{m-i} = b_{i-m} \dots b_{i-1} \\ b_m \neq b_j \end{array} \right\} \qquad (\star)$$

Daraus ergibt sich nach Initialisierung von $\underline{\text{Tab}}[0] = -1$ für jeden folgenden Eintrag $\underline{\text{Tab}}[i]$ folgendes $\underline{\text{Verfahren:}}$



- ► linker Finger: wähle m < i in absteigender Reihenfolge (also i 1, i 2, ...), sodass Pat [i] \neq Pat [m] ℓ
- ▶ Parallelverschiebung beider Finger bis zum linken Rand: wenn Pat[0...m-1] = Pat[i-m...i-1], dann fülle Tab[i] = m
- wenn keine passende Position m gefunden werden kann, dann fülle Tab[i] = -1.

AUFGABE 1 — ZWEI-FINGER-METHODE

Teil (a) Pattern: aabaaacaab

Position	0	1	2	3	4	5	6	7	8	9
Pattern	а	а	b	а	а	а	С	а	а	b
Tabelle	-1 1 = e	-1 -1 -1 -1 -1	1	-1 1	-1	2	2	-1	-1	1

AUFGABE 1 — ZWEI-FINGER-METHODE

16	eii (a)	Pattern: aabaaacaab									
	Position	0	1	2	3	4	5	6	7	8	9
	Pattern	a	a	b	a	a	a	С	a	a	b
	Tabelle	-1	-1	1	-1	-1	2	2	-1	-1	1

AUFGABE 1 — TEIL (B)

Teil (b)

Position	0	1	2	3	4	5
Pattern	C	b	C	C	b	a
Tabelle	-1	0	-1	1	0	2
			-	.(

Teil (b)

Position	0	1	2	3	4	5
Pattern	С	b	С	С	b	а
Tabelle	-1	0	-1	1	0	2

- Pat[0...1] = Pat[3...4] wegen Tab[5] = 2 (Zyklenmethode), d.h. Pat[3] = Pat[0] = c und Pat[4] = Pat[1] = b
- wegen Tab[3] = 1 ist Pat[2] = Pat[0] = c (Zyklenmethode)
- oder: wegen Tab[3] = 1 ist Pat[1] # Pat[3] und
 Pat[2] = Pat[0] = c (Parallelverschiebung in der Zwei-Finger-Methode bzw. Gleichung (*))

Levenshtein-Distanz

Aufgabe 2

LEVENSHTEIN-DISTANZ

Kosten zur Überführung eines Wortes $w = w_1 \dots w_n$ in ein Wort $v = v_1 \dots v_k$; schreibe $d(w_1 \dots w_j, v_1 \dots v_i) = d(j, i)$.

LEVENSHTEIN-DISTANZ

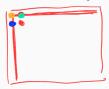
Kosten zur Überführung eines Wortes $w = w_1 \dots w_n$ in ein Wort $v = v_1 \dots v_k$; schreibe $d(w_1 \dots w_j, v_1 \dots v_i) = d(j, i)$.

$$d(0,i) = i \longrightarrow \text{Eintrage in oler 0-ten 2eile}$$

$$d(j,0) = j \longrightarrow \text{Spalte}$$

$$d(j,i) = \min \{ \underline{d(j,i-1)} + 1, \underline{d(j-1,i)} + 1, \underline{d(j-1,i-1)} + \delta_{j,i} \}$$

für alle $1 \le j \le n$ und alle $1 \le i \le k$ wobei



$$\delta_{j,i} = \begin{cases} 1 & \text{wenn } w_j \neq v_i \\ 0 & \text{sonst} \end{cases}$$

LEVENSHTEIN-DISTANZ

Kosten zur Überführung eines Wortes $w = w_1 \dots w_n$ in ein Wort $v = v_1 \dots v_k$; schreibe $d(w_1 \dots w_j, v_1 \dots v_i) = d(j, i)$.

$$d(0, i) = i$$

$$d(j, 0) = j$$

$$d(j, i) = \min \{d(j, i - 1) + 1, d(j - 1, i) + 1, d(j - 1, i - 1) + \delta_{j,i}\}$$

für alle $1 \le j \le n$ und alle $1 \le i \le k$ wobei

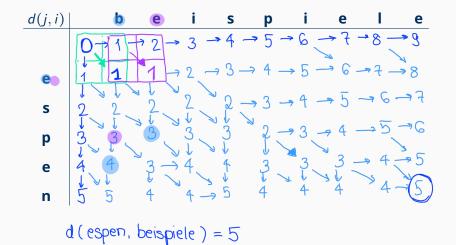
$$\delta_{j,i} = \begin{cases} 1 & \text{wenn } w_j \neq v_i \\ 0 & \text{sonst} \end{cases}$$

Anschaulich: Überlagerung durch Pattern → Pfeile zeigen "Ursprung" des Minimums an

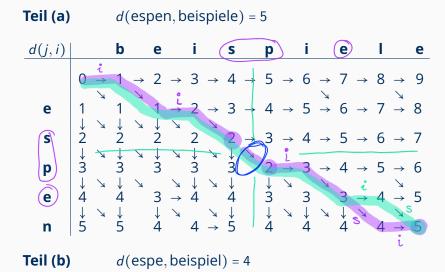
$$w_j \neq v_i$$
: $+1$ $?$ $w_j = v_i$:

AUFGABE 2

Teil (a)



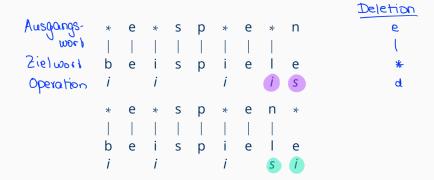
Teil (a) d(espen, beispiele) = 5



AUFGABE 2

Teil (c) Alignments mit minimaler Levenshtein-Distanz:

Teil (c) Alignments mit minimaler Levenshtein-Distanz:



Teil (d) 2 Alignments = 2 Backtraces

mit Lösungen

Weitere Aufgaben aus der

Aufgabensammlung

AUFGABE 7.1.13 (AGS)

- (a) Bestimmen Sie die mit Hilfe des KMP-Algorithmus berechnete Verschiebetabelle für das Pattern abbabbaa.
- (b) Mit Hilfe des KMP-Algorithmus ist unten stehende Verschiebetabelle berechnet worden. Die mit einem "?" markierten Einträge sind unbekannt. Vervollständigen Sie das aus den Symbolen a, b und c bestehende Pattern.

Position	0	1	2	3	4	5
Pattern	b					С
Tabelle	-1	?	?	0	?	3

AUFGABE 7.1.13 (AGS)

Teil (a) Pattern: abbabbaa

Position	0	1	2	3	4	5	6	7
Pattern	a	b	b	a	b	b	а	a
Tabelle								

Teil (a	1)	Pattern: abbabbaa									
Po	sition	0	1	2	3	4	5	6	7		
Pa	ittern	а	b	b	a	b	b	а	a		
Ta	belle	-1	0	0	-1	0	0	-1	4		
Teil (b	o)										
Po	sition	0		1	2	3	3	4	5		
Pa	ttern	b							С		
Tal	oelle	-1		?	?	С)	?	3		

	π (α)	Tatte	iii, ab	Dabb	aa					
	Position	0	1	2	3	4	5	6	7	
Ī	Pattern	a	b	b	a	b	b	a	а	
	Tabelle	-1	0	0	-1	0	0	-1	4	

Dattern: abbabbaa

Teil (b)

Tail (a)

Position	0	1	2	3	4	5
Pattern	b	a	b	a	b	С
Tabelle	-1	?	?	0	?	3

- Pat[0 ... 2] = Pat[2 ... 4] wegen Tab[5] = 3 (Zyklenmethode), d.h. Pat[2] = Pat[0] = Pat[4] = b
- ▶ wegen Tab[3] = 0 ist Pat[3] ≠ Pat[0] = b und wegen Tab[5] = 3
 ist Pat[3] ≠ Pat[5] = c (Zwei-Finger-Methode bzw. Gleichung (*))
 ⇒ Pat[3] = Pat[1] = a

AUFGABE 7.2.1 (AGS)

Gegeben seien die Wörter w = Dinstas und v = Distanz.

- (a) Berechnen Sie die Levenshtein-Distanz d(w, v) zwischen w und v. Geben Sie die Berechnungsmatrix vollständig an.
- (b) Geben Sie alle Alignments mit minimaler Levenshtein-Distanz zwischen w und v an.

AUFGABE 7.2.1 (AGS)

d(j,i)	D	i	S	t	a	n	Z
D							
i							
n							
S							
t							
a							
S							

d(j,i)		D	i	S	t	a	n z
	0 →	1 →	2 →	3 →	4 →	5 →	6 → 7
D	1	0 →	1 →	2 →	3 →	4 →	5 → 6
i	2	1	0 → ↓ ¾			3 →	
n	3	<u></u>	1	$\begin{array}{ccc} 1 & \stackrel{\vee}{\rightarrow} \\ & \stackrel{\vee}{\rightarrow} \end{array}$	2 → \(\frac{1}{2}\)		3 → 4 ↓ ↘
S	4	↓ 3	2	1 →		3 →	4 4
t	5	4	↓ 3	ż	1 →	2 →	3 → 4
a	6	↓ 5 Ⅰ	4 ↓ <i>√</i>	↓ 3 ↓	↓ ↓ 2 		2 → 3
S	Ž	ě	š Î	$\check{4}$	↓ 3	↓ \ 2	$2 \rightarrow 3$

d(j,i)		D	i	S	t	a	n z
	0 →	1 →	2 →	3 →	4 →	5 →	6 → 7
D	1	0 →	1 →	2 →	3 →	4 →	5 → 6
i	2	1	0 →	1 →	2 →	3 →	4 → 5
n	3	↓ 2	1 \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \	1 → \(\(\)	2 → \(\frac{\frac{1}{3}}{3}\)	3 ,	3 → 4
S	4	↓ 3	2	1 →		3 →	
t	5	4	↓ 3 □	2	1 →	2 →	3 → 4
a	6	↓ 5 	4 ↓ <i>√</i>	→ 3 →	2		2 → 3
S	Ž	ě	š 1	4	↓ 3	↓ ↓ 2	2 → 3

d(j,i)		D	i	S	t	a	n z
	0 →	1 →	2 →	3 →	4 →	5 →	6 → 7
D	1	0 →	1 →	2 →	3 →	4 →	5 → 6
i	2	1	0 →	1 →		3 →	4 → 5
n	3	<u>†</u>	1 \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \	$\begin{array}{ccc} 1 & \stackrel{\vee}{\rightarrow} \\ & \stackrel{\vee}{\rightarrow} \end{array}$	2 → \(\frac{\frac{1}{2}}{2}\)	3 🖁	3 → 4
S	4	3	2			3 →	↓
t	5	4	3	2		2 →	3 → 4
a	6	5	4 ↓ ↓	↓ 3 ↓	↓ 2 		2 → 3
S	Ž	ě	š 1	4	↓ 3	¹ / ₂	2 - 3

AUFGABE 7.2.1 (AGS)

Alignments mit minimaler Levenshtein-Distanz:

Alignments mit minimaler Levenshtein-Distanz:

AUFGABE 7.2.2 (AGS)

- (a) Berechnen Sie die Levenshtein-Distanz d(bürste, schürze). Geben Sie die Berechnungsmatrix vollständig an. Wieviele Backtraces enthält die Berechnungsmatrix?
- (b) Geben Sie zwei Alignments mit minimaler Levenshtein-Distanz zwischen den Wörtern bürst und sch an.

AUFGABE 7.2.2 (AGS) — TEIL (A)

d(j,i)	9	c	h	ü	r	Z	е
b							
ü							
r							
S							
t							
е							

_(d(j,i)			S		C		h		ü		r		Z		е
		0	\rightarrow	1	\rightarrow	2	\rightarrow	3	\rightarrow	4	\rightarrow	5	\rightarrow	6	\rightarrow	7
		↓	×		×		\checkmark		\checkmark		×	_	×		\checkmark	_
)	1 ↓	×	1	\rightarrow	2	\rightarrow	3	\rightarrow	4	\rightarrow	5	\rightarrow	6	\rightarrow	7
į	ü	2	А	ž	7	2		3		3	\rightarrow	4	\rightarrow	5	\rightarrow	6
		→ 3	¥	→ 3	A	→	7	2	\rightarrow	↓ 4	A	3		4		5
		J ↓	¥	5	×	→	¥	ے ↓	\rightarrow	4		→	\rightarrow	4	\rightarrow	5
9	5	4		_	\rightarrow	4		4		4		4		4		5
1	t	↓ 5		↓ 4	A	4	\rightarrow	↓ 5	A	↓ 5	A	↓ 5	A	↓ 5	A	5
	•	↓ 6		.	¥	.	7		7	\downarrow	7	↓	\checkmark	\downarrow	\checkmark	
•	9	6		5		5		5	\rightarrow	6		6		6		5

 $d(b\ddot{u}rste, sch\ddot{u}rze) = 5$

d(j,i)		S	С	h	ü	r	Z	е
	0 →	1 →	2 →	3 →	4 →	5 →	6 →	7
b	1 \	1 →	2 →	3 →	4 →		6 →	7
ü	↓ ¾ 2	↓ \ 2	2 →	3	3 →	4 →	5 →	6
r	3	↓ \ 3	↓	3 →	↓ <u>↓</u> 4	3 →	4 →	5
S	↓ \ 4	3 →	↓ \ 4	↓ \ 4		↓ <u>↓</u> 4	\checkmark	5
t	5	↓ \ 1	4 →	↓ ¾			↓ ↘	5
e	↓ 6	↓ \ 5	↓ ¼ 5	5 5 →	↓ \ 6		↓ 🔌	5

 $d(b\ddot{u}rste, sch\ddot{u}rze) = 5$

d(j,i)		S	С	h	ü	r	z e
	0 →	1 →	2 →	3 →	4 →	5 →	6 → 7
b	1 ×	\(\frac{\frac}\frac{\frac{\frac{\frac{\frac{\frac{\frac{\fin}}}}}}{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac}\frac{\frac{\frac{\frac}\fint}}}}{\frac{\frac{\frac{\frac{\frac{\frac{\frac}\fint}}}}{\firac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac}}}}}}}{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac}}}}}}{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac}}}}}}}{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac}}}}}}}{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac{\frac}}}}}}{\frac{\fra	2 →	3 <i>→</i>	×	_	6 → 7
ü	↓ \ 2	↓ ↓ 2	\searrow	3			5 → 6
r	↓ 3	→ → → 3		3 →	↓ ↓		4 → 5
-	↑	7	\downarrow \searrow	↓ ↓		↓ ∡	\checkmark
S	4 <u>↓</u>	3 →	7	4 ↓ ↓			4 → 5 ↓ \
t	5 ↓	4 ↓ ↓ 5	4 → ↓ ¼	5	5 ↓ \	5 ↓ ∖	55↓✓
е	6	5	5	5 →	6	6	6 5

 $d(b\ddot{u}rste, sch\ddot{u}rze) = 5$

d(j,i)		S	С	h	ü	r	Z	е
	0 →	1 →	2 →	3 →	4 →	• 5 →	. 6 →	7
b	1	1 →	2 →	3 <i>→</i>			6 →	7
ü	[↓] ¾	↓ ↓ 2	2 →	3		• 4 <i>→</i>	5 →	6
r	3	3	3 >	3 <i>→</i>	↓ <u>↓</u> 4	3 →	· 4 →	5
s	4	3 →	-	↓	4	4	4 →	5
t	5	4	4 →	↓ \ 5	5	5	5	5
е	6	↓ \ 5	↓	5 →	↓ \ 6	6	6	5

 $d(b \ddot{u} rste, sch \ddot{u} rze) = 5$ Anzahl der Backtraces = 3 * 2 = 6

AUFGABE 7.2.2 (AGS) — TEIL (B)

d(j,i)		S	С	h
	0 -	→ 1 –	→ 2 →	3
b	↓ <u> </u>	1 -	_	3
ü	2	2	2 →	3
r	3	3	3	3
S	4	3 -	4 ↓ \ → 4	↓ 4
t	↓ 5	↓ 4	4 →	↓ 5

AUFGABE 7.2.2 (AGS) — TEIL (B)

```
      b ü r s t
      b ü r s t

      | | | | | | | | | | | |

      s c h * * * * * s c h

      s s s d d
      d d s s s
```