Teilwortsuche mit dem Algorithmus von Rabin und Karp, Fingerabdrücke

a = 5531865469145954396567143945974398746326191124612942912998753987349717 5...

b = 1415926

$a = a_1 a_2 \ldots\ldots a_n \qquad 0 \le a_i, b_i \le B-1$ Bytes

$b = b_1 b_2 \ldots b_m \qquad B = 256$

In unserem Beispiel $B = 10$ (der Einfachheit halber)

- GRUNDIDEE: Interpretiere die Zeichen a_i, b_i als Ziffern einer Zahl zur Basis B.

$$b \to \hat{b} = b_1 B^{m-1} + b_2 B^{m-2} + \ldots + b_{m-1} B + b_m$$

$$a_i a_{i+1} \ldots a_{i+m-1} \to \hat{a}_i = a_i B^{m-1} + a_{i+1} B^{m-2} + \ldots + a_{i+m-1} \cdot 1$$

$$a_{i+1} \ldots\ldots a_{i+m} \to \hat{a}_{i+1} = a_{i+1} B^{m-1} + \ldots + a_{i+m-1} \cdot B + a_{i+m}$$

($\hat{a}_i \times B$)

$$\hat{a}_{i+1} = (\hat{a}_i - a_i B^{m-1}) \times B + a_{i+m}$$

$a_i a_{i+1} \ldots a_{i+m-1} = b \iff \hat{a}_i = \hat{b} \not\Leftarrow \Rightarrow \tilde{a}_i = \tilde{b}$

als Zeichenketten — als Zahlen

Das Teilwortproblem kann in $O(m+n)$ Rechenschritten gelöst werden.

Berechne $\hat{b}, \hat{a}_1 .. O(m)$ (Hornerschema!) $\quad \hat{a}_i \to \hat{a}_{i+1}$ $O(1)$ $n \times$ mal

• IDEE: Wir rechnen modulo Q.

Q wird so gewählt, dass die Rechnungen mit Maschinenarithmetik durchgeführt werden können.

Bsp. long 64 Bit. $B = 256$ $\quad Q \leq 2^{56}$

$\tilde{b} := \hat{b} \bmod Q \qquad \tilde{a}_i := \hat{a}_i \bmod Q$

$$\boxed{\tilde{a}_{i+1} = \left(\tilde{a}_i - a_i B^{m-1}\right) \times B + a_{i+m}} \pmod{Q}$$

$\hat{C} := B^{m-1} \bmod Q$

8 | 56 bits (= 64)

56 | 8 (= 64)

• Wahl von Q: $Q = 2^{56}$ SCHLECHT!

Q ... eine zufällige Primzahl mit $2^{55} \leq Q \leq 2^{56}$

SATZ: Die Anzahl der Primzahlen $\leq N$ ist $\sim \frac{N}{\ln N}$.

$\ln 2^{56} \approx 40$. Es kommen ca. $\frac{2^{55}}{40}$ Primzahlen Q in Frage.

$$\Pr[\text{Fehlalarm an Stelle } i] = \Pr[\hat{a}_i \neq \hat{b}, \text{ aber } \underbrace{\tilde{a}_i = \tilde{b}}_{\hat{a}_i \equiv \hat{b} \pmod{Q} \Leftrightarrow Q \mid \hat{a}_i - \hat{b}}]$$

$$= \Pr\left[Q \mid \underbrace{\hat{a}_i - \hat{b}}_{8m \text{ bits}} \neq 0\right] \leq \frac{m/6}{2^{55}/40} \leq \frac{7m}{2^{55}} \leq \frac{m}{10^{15}}$$

$\leq 2^{8m}$

$\hat{a}_i - \hat{b}$ hat $\leq \frac{\log 2^{8m}}{\log 2^{55}}$ verschiedene Primfaktoren $> 2^{55}$

$= \frac{8m}{55} \leq \frac{m}{6}$

- Es gibt effiziente Primzahltests. (randomisiert.)

Probiere wiederholt eine Zufallszahl im Bereich $2^{55} < Q < 2^{56}$,
bis Q eine Primzahl ist.

Nach 40 Versuchen (im Mittel) hat man Erfolg.

```
abcacbacbabcabcab
bcacbacabcbacabac      bac
accnacancnacdnana      caa
cadbadcbacdacabda      ncn
acandadadcabcdabc
adbcdbcdbdacdadna
aaandnaadcancdann
```

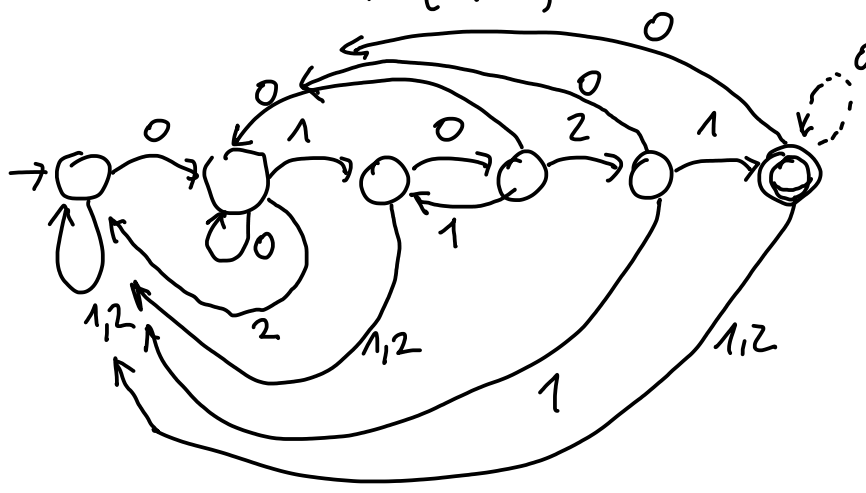
zweidimensionale Mustersuche

Teilwortsuche mit endlichen Automaten

b kommt in a vor $\Leftrightarrow$ $a \in \underbrace{\Sigma^* b \Sigma^*}_{\text{regulär}}$

$\Rightarrow$ Es gibt einen deterministischen endlichen Automaten, der diese Sprache akzeptiert.

$b = 01021$ $\quad \Sigma = \{0,1,2\}$



$\Sigma^* b$ = Sprache

Speicher $|b| \cdot |\Sigma|$

für beliebige reguläre Ausdrücke.