

Tutorium zu „Einführung in die Computerlinguistik“

Automaten + Reguläre Ausdrücke

Automaten

$A = \{Q, \Sigma, s, F, \Delta\}$

Q Zustände

Σ Alphabet

$\{a, b, c \dots\}$ Symbole/Zeichen

s Startzustand

F Finalzustand

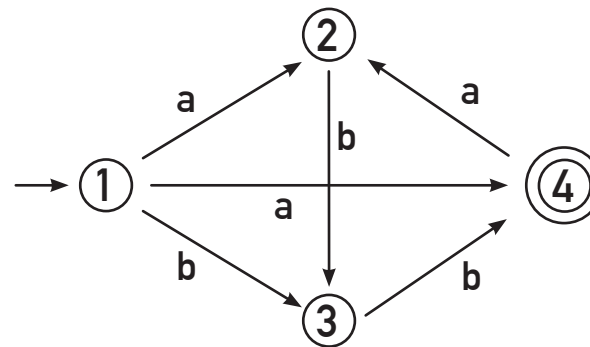
Δ Übergänge

ε leeres Wort

$w, u, v, \dots$ Wörter

$|w|$ Länge Wort w

Σ^* alle möglichen Kombinationen
der Symbole im Alphabet



$Q = \{1, 2, 3, 4\}$

$\Sigma = \{a, b\}$

$s = \{1\}$

$F = \{4\}$

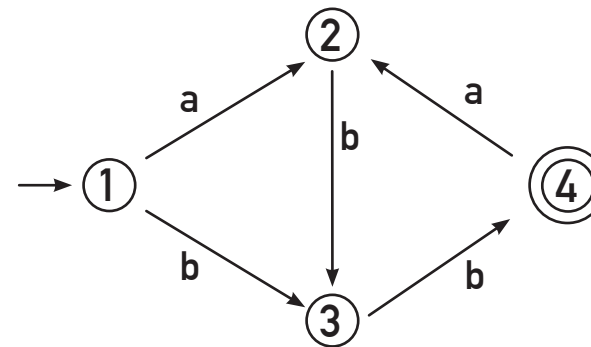
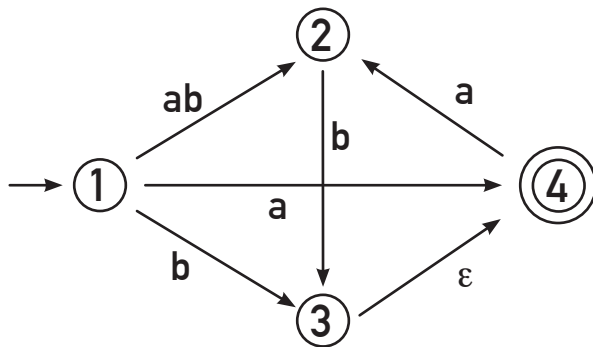
$\Delta = \{ \langle 1, a, 2 \rangle, \langle 1, b, 3 \rangle, \langle 1, a, 4 \rangle, \langle 2, b, 3 \rangle, \langle 3, b, 4 \rangle, \langle 4, a, 2 \rangle \}$

$abb, a, bb, aabb, \dots \in L(A)$

$bab, b, aaa \dots \notin L(A)$

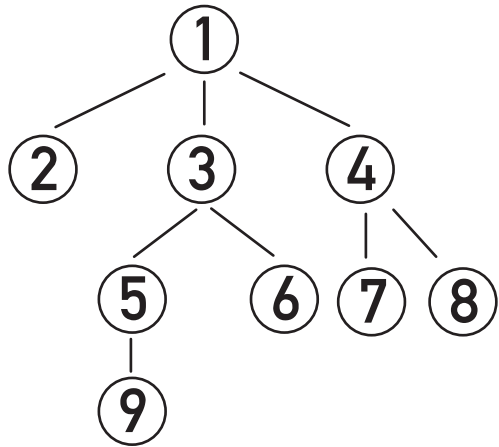
$L(A) \gg$ Sprache, die vom Automaten erkannt wird

NEA vs. DEA



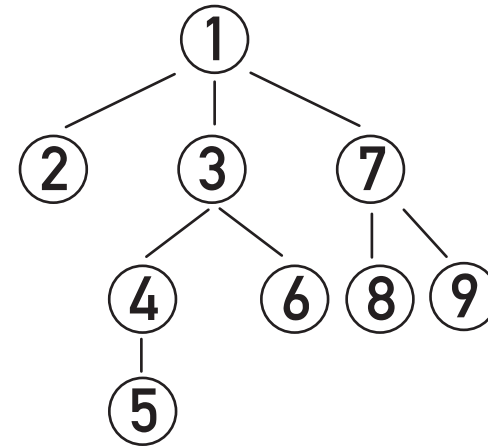
- DEAs haben eindeutige Pfade >> es darf nicht mehrere Pfeile mit dem selben Label von einem Zustand weg geben
- DEAs haben keine komplexen Labels
- DEAs haben keine ϵ -Übergänge
- jeder DEA ist auch ein NEA

Breiten- & Tiefensuche



Breitensuche

zuerst werden alle von einem Knoten erreichbaren Knoten bearbeitet;
Ebene für Ebene



Tiefensuche

einen Arm so weit wie möglich durchlaufen, dann wieder soweit nach „oben“ bis man auf einen unbearbeiteten Arm stößt

Reguläre Ausdrücke

Metazeichen

.	>>	irgendein Zeichen
[abc] [a-z]	>>	enthält eines dieser Zeichen
[^abc]	>>	enthält keines dieser Zeichen
a b	>>	enthält a oder b
*	>>	keines, oder beliebig viele Vorkommen
(ab)*	>>	keines/beliebig viele Vorkommen der Zeichenfolge ab
+	>>	min. ein Vorkommen
?	>>	ein oder kein Vorkommen
{x}	>>	genau x Vorkommen
{x,y}	>>	min. x, max. y Vorkommen
^a	>>	(Zeile) beginnt mit a
a\$	>>	(Zeile) endet mit a

Regex-Tester:
regexpal.com
regex101.com
pythex.org

Beispiele

Regex für deutsche Telefonnr.:
[0-9]{3,6} [/-\s]? [0-9]{3,}

[A-Z][a-z]*-Bert
z.B. A-Bert, Licht-Bert, Faul-Bert, Xyz-Bert, ...