Reguläre Ausdrücke

Reguläre Ausdrücke

- sind eine Methode zur Beschreibung einfacher Sprachen
- werden z.B. eingesetzt, um die Menge möglicher Bezeichner in einer Programmiersprache zu definieren
- haben einen engen Zusammenhang mit Grammatiken (Typ 3) und endlichen Automaten (siehe nächstes Kapitel)

Grundbegriffe

- Eine endliche Menge Σ nennen wir ein **Alphabet**.
- Die Elemente von Σ nennen wir **Symbole** oder **Buchstaben**.
- Die Elemente von Σ^* nennen wir **Worte** (über Σ).
- Eine Teilmenge $L \subseteq \Sigma^*$ nennen wir eine **Sprache** (über Σ).

Definition regulärer Ausrücke

Definition 9

Sei Σ ein Alphabet. R ist ein regulärer Ausdruck über Σ, wenn R einer der folgenden Ausdrücke ist:

1. $\emptyset$,
2. a, wobei $a \in \Sigma \cup \{\epsilon\}$ ist,
3. $R_1 \mid R_2$, wobei R_1, R_2 reguläre Ausdrücke sind,
4. $R_1 \cdot R_2$, wobei R_1, R_2 reguläre Ausdrücke sind (häufig auch einfach $R_1 R_2$),
5. R_1^*, wobei R_1 ein regulärer Ausdruck ist,
6. (R_1), wobei R_1 ein regulärer Ausdruck ist.

Notationen un Abkürzungen

Notation

- Ist $A = \{a_1, \ldots, a_m\} \subseteq \Sigma$, so schreiben wir statt $(a_1 \mid a_2 \mid \cdots \mid a_m)$ auch A und nennen A ebenfalls einen regulären Ausdruck.
- Damit ist für $A \subseteq \Sigma$ der Ausdruck A^* ein regulärer Ausdruck. Insbesondere ist Σ^* ein regulärer Ausdruck.
- Ist R ein regulärer Ausdruck und $n \in \mathbb{N}$ fest, so schreiben wir statt $\underbrace{R \cdot R \cdots R}_{n\text{-mal}}$ auch R^n und nennen R^n ebenfalls einen regulären Ausdruck.

Einfache Beispiele

Beispiele

Sei $\Sigma = \{0, 1, \ldots, 9, a, b, \ldots, y, z\}$. Dann sind die folgende Ausdrücke reguläre Ausdrücke über Σ.

- $1^2\{0,1\}^*0^2 = 11(0 \mid 1)^*00$
- $(a \mid b \mid \cdots \mid z) \cdot (a \mid b \mid \cdots \mid z \mid 0 \mid 1 \mid \cdots \mid 9)^*$
- $0^{10}1^{10} \mid 0^{10}2^{10}$

Reguläre Ausdrücke und Sprachen

Definition 10

Sei Σ ein Alphabet und R ein regulärer Ausdruck über Σ. Die durch R definierte oder beschriebene Sprache $L(R)$ ist folgendermaßen definiert.

1. Ist $R = \emptyset$, so ist $L(R) = \emptyset$.
2. Ist $R = a$, $a \in \Sigma \cup \{\epsilon\}$, so ist $L(R) = \{a\}$.
3. Ist $R = R_1 \mid R_2$, wobei R_1, R_2 reguläre Ausdrücke sind, so ist $L(R) = L(R_1) \cup L(R_2)$.
4. Ist $R = R_1 R_2$, so ist $L(R) = \{uv \mid u \in L(R_1), v \in L(R_2)\}$.
5. Ist $R = R_1^*$, so $L(R) = \{u_1 \dots u_n \mid n \geq 0, u_i \in L(R_1), i = 1, \dots, n\}$.
6. Ist $R = (R_1)$, so ist $L(R) = L(R_1)$.

Auswertungsreihenfolge - Prioritäten der Operatoren

- Um einem regulären Ausdruck eine eindeutige Sprache zuordnen zu können, muss eine Auswertungsreihenfolge festgelegt werden.
- Dies geschieht, durch Festlegung von Prioritäten für die Operatoren $\cdot, |, ^*$ in Definition 9.

Prioritäten

1. * hat die höchste Priorität. D.h., * wird nur auf die kürzeste Folge von Symbolen auf der linken Seite von * angewandt, die einen korrekt geformten regulären Ausdruck darstellt.
2. Die nächst höhere Priorität besitzt $\cdot$.
3. Die niedrigste Priorität besitzt $|$.
4. Durch Einfügen von Klammern kann die Ausführungsreihenfolge gemäß den Klammern geändert werden.
5. Wird die Abkürzung $^n, n \in \mathbb{N}$ verwendet, so hat sie dieselbe Priorität wird $\cdot$.

Beispiele

Sei $\Sigma = \{0, 1, \dots, 9, a, b, \dots, y, z\}$ Dann sind

- $1^2\{0,1\}^*0^2 = 11(0 \mid 1)^*00$
- $(a \mid b \mid \cdots \mid z) \cdot (a \mid b \mid \cdots \mid z \mid 0 \mid 1 \mid \cdots \mid 9)^*$
- $0^{10}1^{10} \mid 0^{10}2^{10}$

reguläre Ausdrücke über Σ.

- $L\big(1^2(0 \mid 1)^*0^2\big)$ besteht aus alle $0,1$-Folgen, die mit zwei Einsen beginnen und mit zwei Nullen enden.
- $L\big((a \mid b \mid \cdots \mid z) \cdot (a \mid b \mid \cdots \mid z \mid 0 \mid 1 \mid \cdots \mid 9)^*\big)$ besteht aus allen Folgen von Buchstaben und Ziffern, die mit einem Buchstaben beginnen.
- $L\big(0^{10}1^{10} \mid 0^{10}2^{10}\big) = \big\{0^{10}1^{10}, 0^{10}2^{10}\big\}$.

Regulärer Sprachen

Definition 11

Sei Σ eine Alphabet und $L \subseteq \Sigma^*$ eine Sprache. L heißt regulär, wenn es einen regulären Ausdruck R mit $L = L(R)$ gibt.

Satz 12

Sei Σ ein Alphabet und $L \subseteq \Sigma^*$ eine reguläre Sprache. Dann ist

$$L^+ := \{w \in \Sigma^* | \exists n \in \mathbb{N}, w_1, \ldots, w_n \in L : w = w_1 \cdots w_n\}$$

ebenfalls eine reguläre Sprache.

Korollar 13

$\Sigma^+ := \Sigma^* \setminus \{\epsilon\}$ *ist eine reguläre Sprache.*

Beispiele regulärer Sprachen (1)

Lemma 14

Die Sprache $L_1 = \{0,1\}^* \setminus \{11\}$ *ist regulär.*

Beweis.

$L_1 = L(R_1)$, wobei

$$R_1 = 0\{0,1\}^*|10\{0,1\}^*|111\{0,1\}^*|110\{0,1\}^*|1|\epsilon.$$

$\square$

Beispiele regulärer Sprachen (2)

Lemma 15

Die Sprache

$$L_2 = \{w \in \{0,1\}^* | w \textit{ enthält nicht die Teilfolge } 101\}$$

ist regulär.

Beweis.

$L_2 = L(R_2)$, wobei

$$R_2 = ((0^*1^*)100)^*\, 0^*1^*(10|\epsilon).$$

□

Die leere Menge und das leere Wort

Folgende Begriffe und Notationen müssen unterschieden werden:

1. das leere Wort ϵ
2. der reguläre Ausdruck ϵ, der die Menge $\{\epsilon\}$ definiert
3. $\{\epsilon\}$, die Menge, die nur das leere Wort enthält
4. $\emptyset$, die leere Menge.