Reguläre Ausdrücke

Karin Haenelt

25.04.2010

Inhalt

- Einführung
- Definitionen
- Kleene-Theorem
- Schreibweisen regulärer Ausdrücke
- Eigenschaften regulärer Sprachen

Was sind reguläre Ausdrücke?

- Reguläre Ausdrücke sind
  - eine Notation zur Beschreibung der Sprachen, die mit endlichen Automaten erkannt werden können
  - wie arithmetische Ausdrücke aus Konstanten und Operatoren aufgebaut

- zumeist bekannt aus ihrer Verwendung in
  - Suchfunktionen in Betriebssystemen (grep)
  - Texteditoren
  - Textverarbeitungsprogrammen und Suchmaschinen
  - Textmusterspezifikation in Programmiersprachen

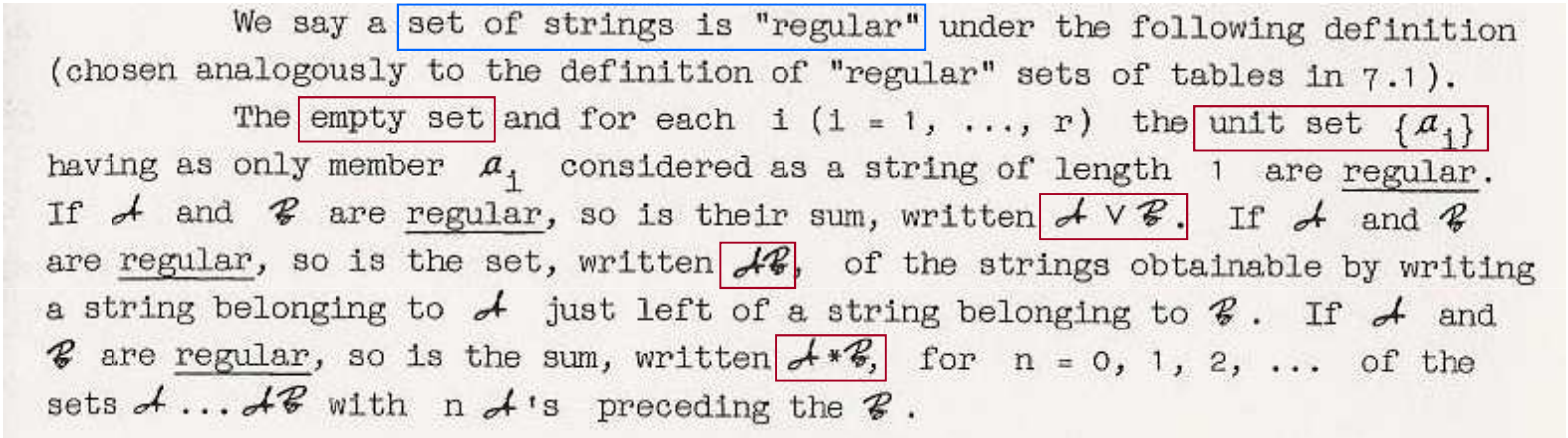
Reguläre Mengen, Reguläre Sprachen, Reguläre Ausdrücke

Stephen Kleene

- Stephen Kleene (Mathematiker)
  - untersuchte (1956), welche Mengen von Zeichenketten von endlichen Automaten akzeptiert werden können
  - entwickelte für diese Mengen eine syntaktische Charakterisierung: Komposition der Zeichenketten
    - aus Elementen und Teilmengen
    - nach bestimmten Regeln
  - führte für diese Mengen den Begriff reguläre Mengen ein

Reguläre Mengen, Reguläre Sprachen, Reguläre Ausdrücke

Stephen Kleene



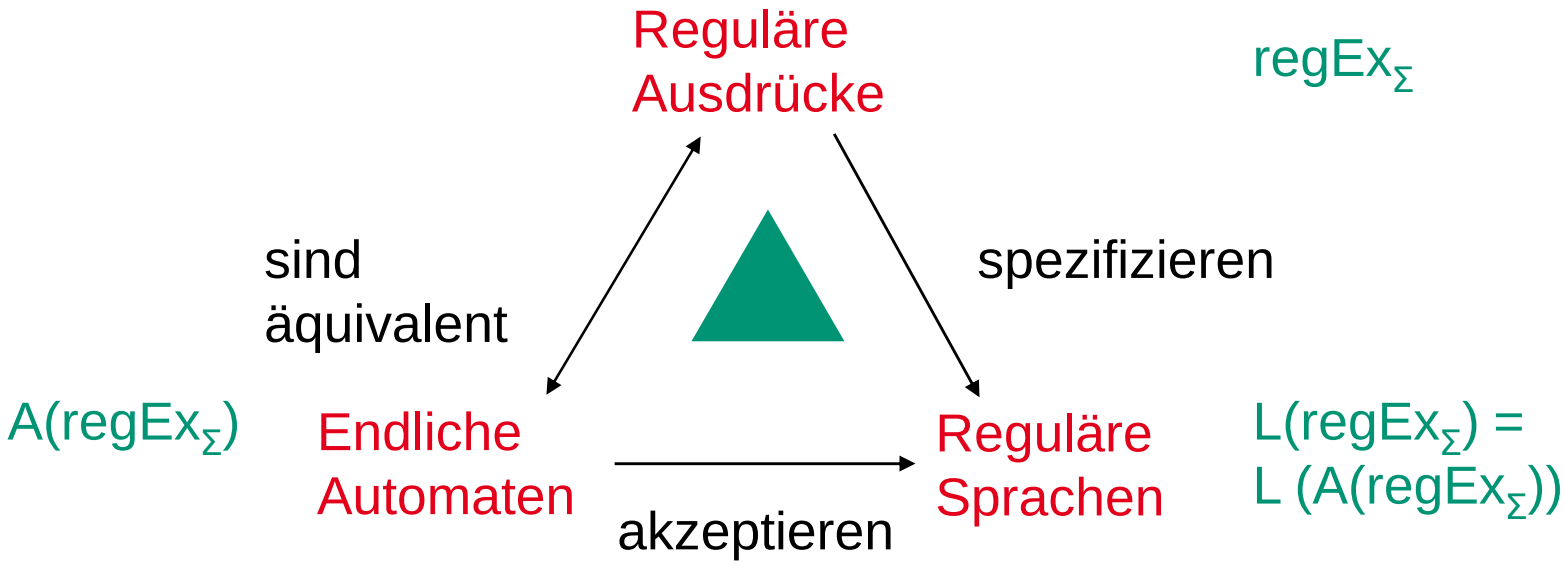
We say a set of strings is "regular" under the following definition (chosen analogously to the definition of "regular" sets of tables in 7.1).

The empty set and for each i (i = 1, ..., r) the unit set $\{a_i\}$ having as only member a_i considered as a string of length 1 are regular. If $\mathcal{A}$ and $\mathcal{B}$ are regular, so is their sum, written $\mathcal{A} \vee \mathcal{B}$. If $\mathcal{A}$ and $\mathcal{B}$ are regular, so is the set, written $\mathcal{AB}$, of the strings obtainable by writing a string belonging to $\mathcal{A}$ just left of a string belonging to $\mathcal{B}$. If $\mathcal{A}$ and $\mathcal{B}$ are regular, so is the sum, written $\mathcal{A}^*\mathcal{B}$, for n = 0, 1, 2, ... of the sets $\mathcal{A} \ldots \mathcal{AB}$ with n $\mathcal{A}$'s preceding the $\mathcal{B}$.

(Kleene, 1956)

Reguläre Mengen, Reguläre Sprachen, Reguläre Ausdrücke

Äquivalenzen



nach einer Darstellung von Martin Kay, zitiert in Jurafsky/Martin (2000)

Inhalt

- Einführung
- Definitionen
- Kleene-Theorem
- Schreibweisen regulärer Ausdrücke
- Eigenschaften regulärer Sprachen

Definitionen (1)

- Symbol a
  - ein Symbol ist ein unzerlegbares Grundzeichen
  - unzerlegbar bezüglich der Art der Betrachtung, um die es gerade geht
  - Beispiele:
    - a, b, c
    - 谢
    - dete, adje, nomn
- Alphabet Σ
  - ein Alphabet ist eine endliche nichtleere Menge von Symbolen

Definitionen (2)

- Wort w
  - ein Wort ist eine endliche Folge von Symbolen aus einem Alphabet.
  - Wir schließen das leere Wort (Wort, das kein Zeichen enthält) in die Definition ein und bezeichnen es mit ε. ■

Definitionen (3)

- Sprache L
  - Eine Sprache ist die Menge aller endlichen Folgen w von Symbolen aus Σ.
  - es gilt
    - ∅ die leere Sprache ist eine Sprache
    - {ε} die Menge, die nur ein leeres Wort enthält, ist eine Sprache
    - Σ* die Universalsprache, die aus der Menge aller endlichen Folge von Symbolen aus einem Alphabet besteht, ist eine Sprache

Definitionen (4)
Reguläre Mengen

Sei Σ ein Alphabet. Die regulären Mengen über Σ werden wie folgt induktiv definiert

1. $\varnothing$ ist eine reguläre Menge über Σ.
2. $\{\varepsilon\}$ ist eine reguläre Menge über Σ.
3. Für alle $a \in \Sigma$ ist $\{a\}$ eine reguläre Menge über Σ.
4. Seien P und Q reguläre Mengen über Σ. Dann sind auch
   - $P \cup Q$ (Vereinigung von P und Q)
   - $P \cdot Q := \{pq \mid p \in P, q \in Q\}$ (Konkatenation von P und Q) und
   - $P^* := \cup_{n \geq 0} P^n$, wobei $P^0 := \{\varepsilon\}$, $P^n := P \cdot P^{n-1}$ für $n > 0$ (Kleenesche Hülle von P)

   reguläre Mengen. ■

Definitionen (5)
Reguläre Sprachen

- Eine reguläre Sprache ist eine reguläre Menge von Zeichenfolgen ■

Definitionen (6)
Reguläre Ausdrücke

- Beschreibt man reguläre Mengen und reguläre Sprachen mit den Ausdrucksmitteln der angegebenen Definitionen, so erhält man Ausdrücke, die gebildet werden aus
  - Operatoren (hier: “ |“ , “ ·“ , “ *“)
  - Grundeinheiten (hier: Zeichen aus einem Alphabet, leeres Zeichen, leere Menge).
- Diese Ausdrücke heißen reguläre Ausdrücke.

Definitionen (7)
Reguläre Ausdrücke

Sei $\Sigma = \{a_1, \ldots, a_n\}$ ein Alphabet. Ein *regulärer Ausdruck* über Σ st eine Sequenz von Symbolen, die durch wiederholte Anwendung der folgenden Regeln gebildet wird:

Reguläre Ausdrücke sind			bezeichnen
1)	$\varnothing$		die reguläre Menge $\varnothing$
2)	ε		die reguläre Menge {ε}
3)	a	für $\forall a \in \Sigma$	die reguläre Menge {a}

4) wenn *p* und *q* reguläre Ausdrücke sind, die die Sprachen *P* und *Q* bezeichnen, dann ist

4 a)	$p + q$	auch:	die *Vereinigung*	die reguläre Menge $P \cup Q$
		$p \mid q$	(auch *Disjunktion* oder *Addition*)	
4 b)	$p \cdot q$	auch:	die Konkatenation	
		pq	(auch Multiplikation)	die reguläre Menge $P \cdot Q$
4 c)	p^*		die *Kleenesche Hülle*	die reguläre Menge P* ■

Hopcroft/Ullmann 1988:29

Inhalt

- Einführung
- Definitionen
- Kleene-Theorem
- Schreibweisen regulärer Ausdrücke
- Eigenschaften regulärer Sprachen

Theorem von Kleene

- Kleene-Theorem: Eine Sprache ist genau dann regulär, wenn sie durch einen endlichen Automaten erkannt werden kann

- Kleenes Formulierung (1956):

  - Synthesetheorem (Kleene, 1956, Theorem 3):
    Jede reguläre Sprache kann in einem endlichen Automaten dargestellt werden

  - Analysetheorem (Kleene, 1956, Theorem 5):
    Jede in einem endlichen Automaten darstellbare Sprache ist regulär

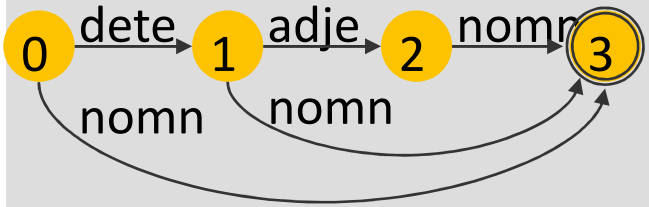
Reguläre Ausdrücke, Reguläre Sprachen, endliche Automaten

Äquivalenzen

e = (dete,adje,nomn) |
(dete,nomn) |
(nomn)

e = (dete|ε) (adje|ε) nomn

e = dete? adje? nomn



spezifizieren

akzeptieren

L(e)= { (dete,adje,nomn),
(dete,nomn),
(nomn) }

L(a)= { (dete,adje,nomn),
(dete,nomn),
(nomn) }

Inhalt

- Einführung
- Definitionen
- Kleene-Theorem
- Schreibweisen regulärer Ausdrücke
- Eigenschaften regulärer Sprachen

Auswertungsreihenfolge der Operatoren regulärer Ausdrücke

- Klammern sparen beim Schreiben:
- Wie für die algebraischen Operatoren ’+’ und ’·’ gibt es auch für die Operatoren regulärer Ausdrücke eine festgelegte Auswertungsreihenfolge.
  Für den Ausdruck *4+5·2* gilt, dass zuerst die Multiplikation und dann die Addition durchzuführen ist. Für die Operatoren für reguläre Ausdrücke gilt folgende Reihenfolge:

1. Kleenesche Hülle (Symbol: *)
2. Multiplikation *bzw.* Verkettung (Symbol ’·’ oder ohne explizites Symbol).
3. Addition *bzw.* Vereinigung (Symbol ’+’ oder ’|’).

Auswertungsreihenfolge der Operatoren regulärer Ausdrücke
Beispiel

- Der Ausdruck de([mns]|ssen) ist äquivalent zu
  - de([mns]|"ssen") oder
  - "de"([mns]|"ssen")

Basisoperatoren und abgeleitete Operatoren

Basis

elementare Ausdrücke

Ø	leere Menge
ε	leeres Zeichen
a	elementares Symbol

zusammengesetzte Ausdrücke

r|s	r oder s
rs	r verkettet mit s
r*	r beliebig oft (auch: null Mal)mit sich selbst verkettet

abgeleitete Operatoren

r+	$rr*$
r?	$(r \mid \varepsilon)$
r{2,4}	$rr \mid rrr \mid rrrr$
[rst]	$(r \mid s \mid t)$

Spezifikation von Zeichenmengen in Programmiersprachen

Beispiele

[A-Z] A, Z und alle Buchstaben die in der Kodierung (z.B. ASCII) zwischen A und Z liegen

[^A-E] alle Zeichen außer A, E und den Zeichen, die in der Kodierung (z.B. ASCII) zwischen A und E liegen

Muster für nicht-reguläre Sprachen in Programmiersprachen

- einige Programmiersprachen (z.B. Perl) erlauben auch Muster, die über reguläre Ausdrücke hinausgehen
  (z.B. Zusatzspeicher: Nummerierung der Muster und Kopieren der dem Muster entsprechenden Ausdrücke)
- Beispiel in Perl : (.*)\1
  - beliebige Zeichenfolge, verkettet mit der Zeichenfolge, die durch die erste Klammer beschrieben wird
  - Zusatzspeicher zum Merken von Klammerinhalten erforderlich (endliche Automaten haben aber kein Gedächtnis)
- Muster mit unbegrenzter Zahl von Rückverweisen beschreiben nicht einmal kontextfreie Sprachen, Erkennung NP-vollständig

Inhalt

- Einführung
- Definitionen
- Kleene-Theorem
- Schreibweisen regulärer Ausdrücke
- Eigenschaften regulärer Sprachen

Abgeschlossenheit

- Eine Menge ist abgeschlossen ist unter einer bestimmten Verknüpfung, wenn die Anwendung dieser Verknüpfung auf beliebige Elemente der Menge wieder ein Element der Menge ergibt.
- Beispiel (N,+): Jede Addition natürlicher Zahlen ergibt wieder eine natürliche Zahl.

Abgeschlossenheitseigenschaften regulärer Sprachen

- Die wichtigsten Angeschlossenheitseigenschaften regulärer Sprachen:
  - Die Vereinigung zweier regulärer Sprachen ist regulär
  - Der Durchschnitt zweier regulärer Sprachen ist regulär
  - Das Komplement zweier regulärer Sprachen ist regulär
  - Die Differenz zweier regulärer Sprachen ist regulär
  - Die Spiegelung einer regulären Sprache ist regulär
  - Die Hülle (Sternoperator) einer regulären Sprache ist regulär
  - Die Verkettung von regulären Sprachen ist regulär
  - Ein Homomorphismus (Ersetzung von Symbolen durch Zeichenreihen) einer regulären Sprache ist regulär
  - Der inverse Homomorphismus einer regulären Sprache ist regulär

Hopcroft/Motwani/Ullman, 2002: 141

Entscheidbarkeit

- Es gibt algorithmische Verfahren,
  bei deren Anwendung
  auf ein beliebiges Element x und
  für beliebige reguläre Sprachen L_1 und L_2 sich nach endlich vielen Schritten ergibt
  - $L_1 = \varnothing$ Leerheit
  - $x \in L_1$ Zugehörigkeit
  - $L_1 = L_2$ Äquivalenz
  - L_1 ist endlich Endlichkeit
  - $L_1 \cap L_2 = \varnothing$ Schnitt

Kleene Algebra

- Menge von Axiomen, die die Ableitung aller Gleichungen zwischen regulären Ausdrücken ermöglicht
- liefert die Regeln, mit denen man die Äquivalenz von Ausdrücken auf algebraischer Basis angeben kann
- nützlich für
  - Vereinfachung von Ausdrücken
  - Beweise
  - Entwicklung von Algorithmen

Vielen Dank

Für das Aufspüren von Fehlern in früheren Versionen und für Verbesserungsvorschläge danke ich

Eva Mujdricza

Literatur

- Hopcroft, John E. und Jeffrey D. Ullman (1988). *Einführung in die Automatentheorie, formale Sprachen und Komplexitätstheorie.* Bonn u. a.: Addison-Wesley, 1988 (engl. Original Introduction to automata theory, languages and computation).
- Hopcroft, John E., Rajeev Motwani und Jeffrey D. Ullman (2002). *Einführung in die Automatentheorie, Formale Sprachen und Komplexität*. Pearson Studium engl. Original: *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley. www-db.stanford.edu/~ullman/ialc.html
- Kleene, Stephen Cole (1956). Representations of Events in Nerve Sets and Finite Automata, In: C. E. Shannon and J. McCarthy, Hgg., *Automata Studies*, S. 3-42, Princeton, NJ, 1956. Princeton University Press.
- Jurafsky, Daniel und James H. Martin (2000): *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. New Jersey: Prentice Hall.*