

Beschreibung regulärer Sprachen mit regulären Ausdrücken

Reguläre Sprachen können außer durch eine Grammatik auch durch sogenannten **reguläre Ausdrücke** beschrieben werden.

Grundlegende reguläre Operatoren

Reguläre Ausdrücke formuliert man mit regulären Operatoren. Die wichtigsten regulären Operatoren sind die folgenden:

- a^* - "a-Stern": a beliebig oft (auch 0 mal) hintereinander
- ab - Verkettung: ein a, dann ein b
- $a|b$ - Oder: entweder ein a oder ein b
- Mit runden Klammern () können Teilausdrücke gruppiert werden

Beispiele

(1) $\alpha_1 = c^*h^*$ beschreibt alle Zeichenketten, die zunächst aus beliebig vielen Buchstaben c bestehen und anschließend aus beliebig vielen Buchstaben h. ¹⁾ $L_1 = \{\epsilon, c, h, ccch, ccchhh, hhhh, cc, \dots\}$.

(2) $\alpha_2 = c^*|h^*$ beschreibt die alle Zeichenketten, die entweder nur aus c's oder nur aus h's bestehen. $L_2 = \{\epsilon, c, h, cc, hh, ccc, hhh, \dots\}$.

(3) $\alpha_3 = (c|h)^*$ beschreibt die alle Zeichenketten, die aus beliebig vielen c's und h's bestehen. $L_3 = \{\epsilon, c, h, hc, ch, ccc, hchh, \dots\}$.



(A1)

Welche Sprachen werden durch die folgenden regulären Ausdrücke beschrieben? Liste eine Beispielmengende von akzeptierten Wörtern auf, aus der hervorgeht, welche Regeln gelten.

(A) $\alpha_4 = 01^*0$ $L = \{00, 010, 0110, 01110, 011110, 0111110, \dots\}$

(B) $\alpha_5 = (01)^*$ $L = \{\epsilon, 01, 0101, 010101, 01010101, 0101010101, \dots\}$

(C) $\alpha_6 = 0|(10^*)^*$ $L = \{\epsilon, 0, 1, 10, 100, 1000, 11, 101, 10100, 100100, \dots\}$

(D) $\alpha_7 = (1(110)^*)^*$ $L = \{\epsilon, 1, 1110, 1110110, 11, 11101110, \dots\}$

(E) $\alpha_8 = 1(1^*|10^*)$ $L = \{1, 11, 111, 1111, 11111, 110, 1100, 11000, \dots\}$

(F) $\alpha_9 = 0|1^*|10^*$ $L = \{\epsilon, 0, 1, 11, 111, 1111, 10, 100, 1000, 10000, \dots\}$

(G) $\alpha_{10} = 1^*0^*(0|1)$ $L = \{0, 1, 10, 11, 110, 111, 11001, 1110001, \dots\}$

Erweiterte reguläre Ausdrücke

Mit der "Minimalausstattung" von drei Operatoren (ohne die Klammern) lassen sich im Prinzip schon alle regulären Ausdrücke schreiben. Bequemer, kürzer und übersichtlicher geht es, wenn man weitere Operatoren hinzunimmt, hier ist allerdings zu beachten, dass es unterschiedliche "Dialekte" solcher Erweiterungen gibt.

Die meisten Programmiersprachen verfügen über Funktionen oder Bibliotheken, um Strings anhand von regulären Ausdrücken zu finden. Meist gibt es auch weitere Funktionen wie Rückbezüge, mit denen sich komplexe Ersetzungen und Ähnliches realisieren lassen.

Typische Erweiterungen umfassen die folgenden Elemente:

- Eckige Klammern fassen Zeichengruppen zusammen. Der Ausdruck $[0-9]$ bedeutet "die Ziffern von 0 bis 9", $[0-9]^*$ bedeutet entsprechend "beliebig oft eines der Zeichen von 0-9. Bislang hätten wir dafür $(0|1|2|3|4|5|6|7|8|9)^*$ schreiben müssen.
- $[abc]$ – der Buchstabe a, b oder c.
- $[a-zA-Z]$ – die Buchstaben A bis Z, groß und klein.
- $.$ ²⁾ – ein beliebiges Zeichen.
- $[\wedge abc]$ – jedes Zeichen außer den Buchstaben a, b und c³⁾
- Wenn man ein Zeichen suchen will, das eigentlich eine Spezialbedeutung hat, muss man es escapen, meist mit einem Backslash: $[\backslash[a-z\backslash]]$ – erst ein [, dann ein Buchstabe a bis z dann ein].
- Der Anfang einer Zeile kann durch ein $^$ gematcht werden, das Ende durch ein $\$$.
- Ein $*$ bedeutet wie bisher, dass beliebig viele Zeichen der vorherigen Gruppe gesucht werden
 $^[\text{aeiou}]^*n\$$ – findet n, an, en, in, on, un, aan, aaan, aen, ein, ... sofern sie in einer Zeile stehen.
- $a\{2,5\}$ – 2 bis 5 Vorkommen des Buchstabens a.
- b^+ – mindestens ein Vorkommen des Buchstabens b⁴⁾.



(A2)

Schreibe die Ausdrücke $[abc]$, b^+ und $a\{2,4\}$ mit Hilfe der elementaren regulären Operatoren auf.

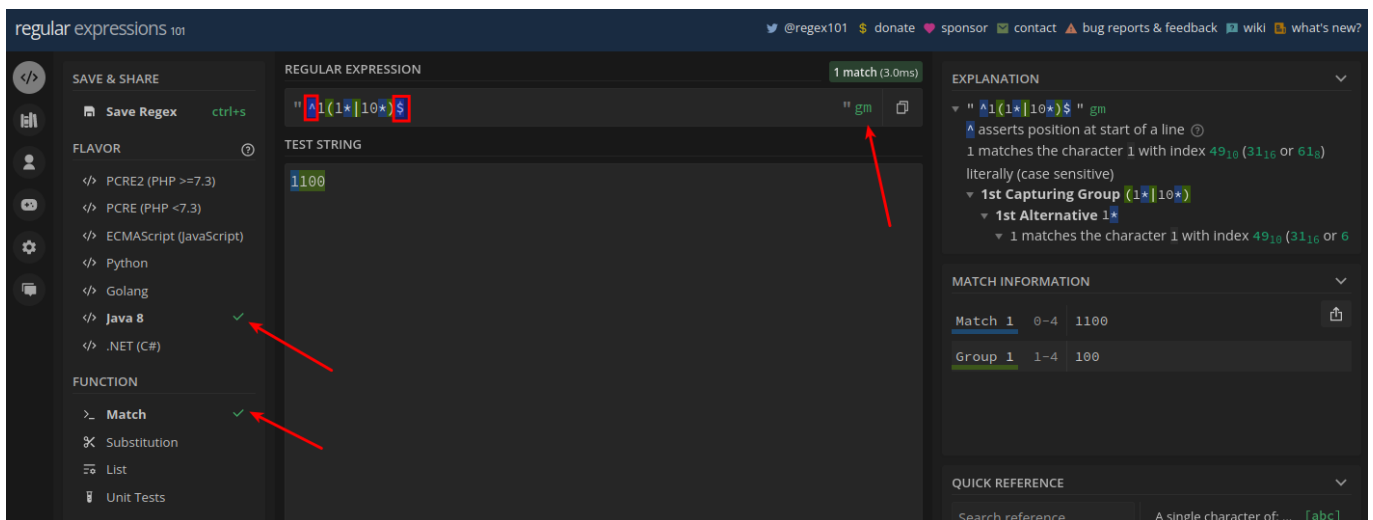


(A3)

Bearbeite das Tutorial unter <https://regexone.com/>, um mit den erweiterten regulären Operatoren vertraut zu werden.

Sandkasten bei regex101.com

Auf der Seite <https://regex101.com/> kannst du reguläre Ausdrücke testen. Außerdem werden die Ausdrücke und Matches erklärt.

**(A4)**

Öffne die Seite <https://regex101.com/> und Sorge dafür, dass die Einstellungen wie im Screenshot zusehen angepasst sind⁵⁾

Überprüfe jetzt deine Ergebnisse aus Aufgabe 1, indem du den regulären Ausdruck in die Eingabezeile "regular Expression" schreibst und die zu testenden Wörter der regulären Sprache in das Feld "Test String", pro Zeile ein Wort.

Achtung: Weil die regulären Ausdrücke nach der Greedy Strategie angewandt werden, musst du deine regulärer Ausdruck inklusive Zeilenanfang (^) und Zeilenende (\$) testen, sonst klappts nicht - siehe Screenshot. Probiere einfach aus, was passiert, wenn du diese beiden Zeichen weglässt. Wenn du Groß und Kleinschreibung nicht unterscheiden möchtest, musst du die Option i für Case-Insensitive setzen.



(A5) Übungen

Füge die etwa 180000 Worte, die in

dieser Datei

enthalten sind in das "Test String"-Feld bei regex101.com ein.

Formuliere anschließend reguläre Ausdrücke, um die folgenden Wörter zu finden. Groß- und Kleinschreibung soll nicht berücksichtigt werden.

1. Alle Wörter, die die Zeichenfolge mc enthalten (zu erwarten: 35 Treffer)
2. Alle Wörter, die mit lau beginnen und danach noch genau einen weiteren Buchstaben enthalten (12 Treffer)
3. Alle Wörter, die mit L beginnen und mit cht enden (31 Treffer)
4. Alle Wörter, die mit einem Vokal enden (36218 Treffer)
5. Alle Wörter, die mit d, r oder t beginnen und mit a, g oder l enden (1258 Treffer)
6. Alle Wörter, die mindestens zwei u enthalten (13749 Treffer)
7. Alle Wörter, die ein doppeltes u enthalten (67 Treffer)
8. Alle Wörter, die aus sechs Buchstaben bestehen (6723 Treffer)
9. Alle Wörter, die höchstens fünf Buchstaben lang sind (7019 Treffer)
10. Alle Wörter, die mindestens 15 Buchstaben lang sind (38518 Treffer)
11. Alle Wörter, die aus sechs Buchstaben bestehen und deren dritter Buchstabe ein t ist (344 Treffer)
12. Alle Wörter, die die Vokale a, e, i, o und u in dieser Reihenfolge und genau einmal enthalten. (1 Treffer)
13. Alle Wörter, die ein y enthalten, aber kein a. (885 Treffer)



(A6) Gültige Telefonnummern

Es soll erkannt werden, ob ein String das korrekte Format für eine Telefonnummer hat.

Für uns soll eine korrekte Telefonnummer zunächst folgende Kriterien aus folgenden Bestandteilen zusammengesetzt sein:

- Einer Vorwahl, die mit einer 0 beginnt und danach mindestens eine andere Ziffer enthält.
- Einem der folgenden Zeichen zur Abtrennung der Vorwahl: -, / oder Leerzeichen
- Einer mindestens einstelligen Zahl, die nicht mit 0 beginnt.

Erstelle einen regulären Ausdruck, der in der folgenden Liste von Ziffernfolgen alle korrekten Telefonnummern erkennt.

Liste

```
07071-12345
17071-12345
07071 12345
07071 02345
0721 1654
171-8867524
07473/138652
0049-12345
```

1)

"Erst beliebig viele c's, dann beliebig viele h's" - wobei "beliebig viele auch "keines" bedeuten kann.

2)

Punkt

3)

Hier bedeutet das ^ gewissermaßen "nicht"

4)

Unterschied zu *: Beliebig oft, aber **nicht** kein mal

5)

Regex Flavor Java 8, Funktion Match, regex Optionen gm

From:

<https://www.info-bw.de/> -

Permanent link:

https://www.info-bw.de/faecher:informatik:oberstufe:automaten:regulaere_ausdruecke:start?rev=1678128267

Last update: **06.03.2023 18:44**

