Vincent Tscherter

Regexp – Reguläre Ausdrücke

Suchen und Ersetzen von Texten, das können sehr viele Programme. Aber was, wenn man nach allen Telefonnummern oder E-Mail-Adressen in einem Text suchen möchte? Oder nach allen ISBN? Oder ohne grossen Aufwand nach möglichst viel Schreibweisen von Viagra suchen, um diese durch ##SPAM## zu ersetzen? Hier helfen reguläre Ausdrücke.

Ein regulärer Ausdruck (engl. regular expression, Abk. RegExp oder Regex) ist eine Zeichenkette, die der Beschreibung von Mengen von Zeichenketten mit Hilfe bestimmter syntaktischer Regeln dient. Reguläre Ausdrücke finden werden in der Programmierung sehr oft verwendet; für fast alle Programmiersprachen existieren Umsetzungen.

Reguläre Ausdrücke stellen eine Art Filterkriterium für Texte dar, indem der jeweilige reguläre Ausdruck in Form eines Musters mit dem Text abgeglichen wird. So ist es beispielsweise möglich, alle Wörter, die mit S beginnen und mit D enden, zu suchen, ohne die zwischenliegenden Buchstaben explizit vorgeben zu müssen.

Ein weiteres Beispiel für den Einsatz als Filter ist die Möglichkeit, komplizierte Textersetzungen durchzuführen, indem man die zu suchenden Zeichenketten durch reguläre Ausdrücke beschreibt.

1 Reguläre Ausdrücke festlegen

1.1 Einfache Textsuche (Zeichenliterale)

Diejenigen Zeichen, die direkt (wörtlich, literal) übereinstimmen müssen, werden auch direkt notiert.

Suchmuster	und
Text	Hier ist der Apfel. Man mache Raum - Er nehme seine Weite, Wie's Brauch ist - Achtzig Schritte geb ich ihm - Nicht weniger, noch mehr - Er rühmte sich, Auf ihrer h**und**ert seinen Mann zu treffen - Jetzt Schütze triff, **und** fehle nicht das Ziel!

1.2 Beliebiges Zeichen, der Joker

Ein Punkt . bedeutet, dass an seinem Platz ein beliebiges Zeichen stehen kann.

Suchmuster ih.

Text Hier ist der Apfel. Man mache Raum - Er nehme seine Weite, Wie's Brauch ist - Achtzig Schritte geb ich **ihm** - Nicht weniger, noch mehr - Er rühmte sich, Auf **ihr**er hundert seinen Mann zu treffen - Jetzt Schütze triff, und fehle nicht das Ziel!

1.3 Ein Zeichen aus einer Auswahl, eine Zeichenklasse [abc]

Mit eckigen Klammern lässt sich eine Zeichenauswahl definieren. Der Ausdruck in eckigen Klammern steht dann für genau ein Zeichen aus dieser Auswahl. [egh] bedeutet also eines der Zeichen e, g oder h.

Suchmuster [nm]..[hr]

Text Hier ist der Apfel. Man **mach**e Raum - Er nehme seine Weite, Wie's Brauch ist - Achtzig Schritte geb ich ihm - Nicht weniger, **noch** **mehr** - Er rühmte sich, Auf ihrer hu**nder**t seinen Mann zu treffen - Jetzt Schütze triff, und fehle **nich**t das Ziel!

Es gibt sog. vordefinierte Zeichenklassen, die als Abkürzung für die Zeichen einer Auswahl verwendet werden können: Zum Beispiel steht \d für alle Ziffern, also als Abkürzung für [0123456789]. \w steht für einen Buchstabe, eine Ziffer oder den Unterstrich; \w könnte also als [a-zA-Z_0-9] ausgeschrieben werden.

Es können auch Bereiche von Zeichen angeben werden: [a-j] steht für irgendeinen Buchstaben aus dem Bereich a...j.

1.4 Option X?

Soll ein Teil des Suchmusters optional sein, so ist dieser mit einem Fragezeichen ? zu markieren. Ein solcher Teil kann einmal vorkommen, muss es aber nicht. Das heisst, der Ausdruck kommt null- oder einmal vor. Anders gesagt: Er kommt höchstens einmal vor.

Suchmuster Schüler(in)?

Text Die **Schülerin** Hermine Granger ist Bulmahns literarisches Ebenbild. Sie hebt als Erste den Finger, wenn die Lehrer etwas wissen wollen. Sie interessiert sich für nahezu alle Fächer mit Inbrunst. ... So illustriert der dritte Harry-Potter-Film die aktuelle Diskussion um die auseinander driftenden Leistungen junger **Schüler**.

Die Klammern () dienen dabei der Gruppierung: Die beiden Buchstaben in sind zusammen optional. Der Ausdruck erkennt also Schüler und Schülerin, nicht aber zum Beispiel Schülern.

1.5 Alternative X|Y

Gibt es zwei oder mehrere Varianten des Suchmusters, so werden diese mit dem sog. Pipezeichen | voneinander getrennt angegeben.

Suchmuster Sch(ü|ue)ler(in)?

Text War es die **Schülerin**, oder waren es die **Schueler**?

1.6 Repetition X*, X+, {n}, {min,}, {,max}, {min,max}

Soll ein Teil des Suchmusters beliebig oft wiederholt werden können, so werden diese mit einem * markiert.

Suchmuster eh*|so*|dol*

Text dad w**e**tter is **ehhhhhhhh** n**e**d **sooo** **dollll**

Varianten

+ Der voranstehende Ausdruck muss mindestens einmal vorkommen, darf aber auch mehrfach vorkommen. (Dies entspricht {1,})

{n} Der voranstehende Ausdruck muss exakt n-mal vorkommen.

{min,} Der voranstehende Ausdruck muss mindestens min-mal vorkommen.

{,max} Der voranstehende Ausdruck darf maximal max-mal vorkommen.

{min,max} Der voranstehende Ausdruck muss mindestens min-mal und darf maximal max-mal vorkommen.

Gierig oder genügsam?

Die obigen sog. Quantoren sind „gierig“: Mit ihnen wird nach der grösstmöglichen Übereinstimmung gesucht. Ein Beispiel: „.*David“ würde im Text „*Hallo David, wie geht es Dir? David*, ich wünsche Dir alles Gute für Deinen Match morgen Abend! Sarah“ den gesamten kursiven Text erkennen.

Dieses Verhalten ist nicht immer gewollt. Die meisten Umsetzungen von regulären Ausdrücken erlauben es daher, dieses Verhalten zu übersteuern und einen Quantor als „genügsam“ zu deklarieren. Im obigen Beispiel würde mit „.*?David“ im Text „*Hallo David*, wie geht es Dir? David, ich wünsche Dir alles Gute für Deinen Match morgen Abend! Sarah“ nur noch die kursive Anrede erkennen. Das ? bedeutet, bei der Auswertung des regulären Ausdrucks soll der unmittelbar vorangehende Quantor * „genügsam“ sein, also so wenig wie möglich erkennen.

1.7 Sonderzeichen \()[]*+?{}.|^$-

Da die Zeichen \()[]*+?{}.|^$- eine bestimmte Bedeutung haben, muss diesen Zeichen ein Backslash \ vorangestellt werden, wenn man nach diesen Zeichen selbst suchen will.

Suchmuster \+

Text Muster Hans
Friedheimstrasse 5
8000 Zürich
+41-44-234 34 23

1.8 Zusammenfassung

a	Das Zeichen a
.	Ein beliebiges Zeichen
[abc]	Ein beliebiges Zeichen aus der Menge {a, b, c}
\d	eine Ziffer [0-9]
\w	Buchstabe, Ziffer oder Unterstrich – [a-zA-Z_0-9]
X|Y	X oder Y
X*,X*?	Eine beliebige Wiederholung von X, gierig bzw. genügsam
X+, X+?	Mindestens einmal X, gierig bzw. genügsam
X?	Höchstens einmal X
{n}	der vorangehende Ausdruck muss exakt n mal vorkommen
{m,n}	der vorangehende Ausdruck muss mindestens m mal vorkommen und darf höchstens n mal vorkommen
(xyz)	Gruppierung: xyz müssen miteinander vorkommen
[^xyz]	Nicht x, nicht y, nicht z
\X	X ein Sonderzeichen \()[]*+?{}.|^$-

Referenzen und weiterführende Artikel: http://de.wikipedia.org/wiki/Regul%C3%A4rer_Ausdruck

2 Aufgaben

Um die Lösungen zu den Aufgaben am Computer zu testen, kann das Tool auf folgender Website verwendet werden:

http://www.regexe.de/

2.1 Muster Suchen

Finde im folgenden Text die im Suchmuster beschriebenen Textstellen.

> Er kauft ihr einen breiten Hut, der wär' wohl für die Sonne gut, für fünfzehn Pfennige. Behalt dein Gut, lass mir mein Mut - kein' and'rer doch dich nehmen tut für fünfzehn Pfennige.

1. Hut|Gut|Mut
2. [HGMtg]ut
3. [mds]ein\s[GMH]ut

Hinweis: \s bedeutet Leerzeichen.

Lösungen

1. Hut|Gut|Mut

> Er kauft ihr einen breiten **Hut**, der wär' wohl für die Sonne gut, für fünfzehn Pfennige. Behalt dein **Gut**, lass mir mein **Mut** - kein' and'rer doch dich nehmen tut für fünfzehn Pfennige.

2. [HGMtg]ut

> Er kauft ihr einen breiten **Hut**, der wär' wohl für die Sonne **gut**, für fünfzehn Pfennige. Behalt dein **Gut**, lass mir mein **Mut** - kein' and'rer doch dich nehmen **tut** für fünfzehn Pfennige.

3. [mds]ein\s[GMH]ut

> Er kauft ihr einen breiten Hut, der wär' wohl für die Sonne gut, für fünfzehn Pfennige. Behalt **dein Gut**, lass mir **mein Mut** - kein' and'rer doch dich nehmen tut für fünfzehn Pfennige.

2.2 Muster bestimmen

Im Text sind einige Stellen markiert. Welcher der vier Muster hat diese Markierungen erzeugt?

1. Die Roten Waldam**eisen** bilden **einen** dauerhaften Staat, der sich in **einem** Kuppelbau vor allem aus Fichtennadeln befindet. Das Nest enthält meist nur eine Königin und etwa eine halbe Million Arbeiterinnen.

- ☐ eis|nn|m
- ☐ ei..[nm]
- ☐ ei[sn]en
- ☐ ei..n|m (würde bedeuten: „ei..n“ ODER „m“)

2. Die Roten Waldameisen bilden **einen** dauerhaften Staat, der sich in **einem** Kuppelbau vor allem aus Fichtennadeln befindet. Das Nest enthält meist nur **eine** Königin und etwa eine halbe Million Arbeiterinnen.

- ☐ ein(e[nm]?)?
- ☐ ein(e(n|m))?
- ☐ eine?[nm]
- ☐ eine?(n|m)?

3. **Die** **Roten** **Waldameisen** bilden einen dauerhaften **Staat**, der sich in einem **Kuppelbau** vor allem aus Fichtennadeln befindet. **Das** **Nest** enthält meist nur eine **Koenigin** und etwa eine halbe **Million** **Arbeiterinnen**.

- ☐ [DRWSKAMN]\w*
- ☐ [A-Z][a-z]*
- ☐ \b\w+
- ☐ [A-Z][a-z]+

Lösungen

1. ei..[nm]

2. ein(e[nm]?)? oder eine?(n|m)?

3. [DRWSKAMN]\w*

2.3 Zahlen in einem Text erkennen

Ein einfaches Muster zur Zahlenerkennung wäre [0-9]+ bzw. in der abgekürzten Schreibweise \d+. Erweitere dieses Muster Schritt für Schritt, so dass ...

1. vor einer Zahl auch eine Minus stehen kann, z.B. -4342

2. die Zahl auch Nachkommastellen aufweisen kann. z.B. 3.14159 oder -3.14159. Verwende als Interpunktionszeichen einen Punkt.

3. nach einer Zahl ein Exponent mit Vorzeichen folgen kann z.B. 314159E-5 oder 3.14159E+5 oder 3.14159E5

4. die Ziffern vor dem Punkt in dreier Blöcken durch ein Apostroph getrennt sein können. z.B. -23'234'233.23E-23

Prüfen Sie Ihre Lösungen an mindestens den folgenden Zahlen:

```
5223445
214'352'134
-1234
-12353.3523
214'352'134.1323
234E12
-2'343.34E-23
3453.3455.345
3"3345'4.34
3.14159
-3.14159
3.14159E-5
3.14159E+5
3.14159E5
-31'415'987.123E-5
```

Lösungen

1. \-?\d+

2. \-?\d+(\.\d+)?

3. \-?\d+(\.\d+)?(E(\-|\+)?\d+)?

4. \-?\d{1,3}('\d\d\d)*(\.\d+)?(E(\-|\+)?\d+)?

2.4 Spamfilter: Viagra

Erstellen Sie einen regulären Ausdruck, der alle im folgenden Text enthaltenen Schreibweisen von Viagra entdeckt, aber nicht die Wörter Nigeria, Niagara, Hiragana:

```
Viagra  viagra  \/iagra
VIAGRAvi@gra  \/|/\gr/\
viaGgraviagr@  viiagrra
v|agra  v!agra  v|agra
viagRa  Via,gra  Viagrra
Nigeria Niagra  Hiragna
```

Lösung

Es gibt verschiedene Lösungen, unter anderem könnte auch (?i) verwendet werden, um Gross-/Kleinschreibung zu ignorieren. Ohne diese Abkürzung ist eine mögliche Lösung:

```
(V|v|\\/)+.?(i|I|!|\|)+.?(a|A|@|/\\)+.?(g|G)+.?(r|R)+.?(a|A|@|/\\)+.?
```

3 Ersetzen (replace)

Gefunde Textstellen können mit regulären Ausdrücken auch einfach ersetzt werden. Dazu braucht es zusätzlich zum Suchmuster ein Ersetzungsmuster, das beschreibt, wie die gefundenen Textstellen ersetzt werden. Die zu ersetzenden Textstellen werden mit runden Klammern () markiert und implizit durchnummeriert. Im Ersetzungsmuster können die Textstellen dann mit $1, $2 etc. adressiert werden:

Suchmuster	([a-z0-9_-]+)@([a-z0-9_.-]+)
Ersetzungsmuster	$1[at]$2
Text	Praesent in lacus nunc, sed blandit risus. Pellentesque quis ebenawa_432@yahho.com venenatis neque. Morbi placerat turpis vitae justo iaculis a luctus tellus lobortis. avadele_67@gmail.com Quisque eget tellus nec eros facilisis consectetur. Sed sed nulla lectus.
Text nach Ersetzen	Praesent in lacus nunc, sed blandit risus. Pellentesque quis **ebenawa_432[at]yahho.com** venenatis neque. Morbi placerat turpis vitae justo iaculis a luctus tellus lobortis. **avadele_67[at]gmail.com** Quisque eget tellus nec eros facilisis consectetur. Sed sed nulla lectus

3.1 Aufgaben zu Ersetzen

Mit Hilfe von regulären Ausdrücken sollen die Daten einer Wassermessung in ein Format gebracht werden, das in Wikis für Tabellen verwendet wird.

Input
Aare - Bern 5.52
Aare - Brienzwiler 4.20
Aare - Brugg 5.25
.. ..

Output
Aare - Bern	5.52 °C
Aare - Brienzwiler	4.20 °C
Aare - Brugg	5.25 °C
..	..

Quelle der Wassermessungsdaten: http://www.hydrodaten.admin.ch/lhg/T-Bulletin.html

Lösung

Suchmuster: (\w+ \- \w+) (\d\.\d\d)

Ersetzungsmuster: | $1 | $2 °C |