„Einführung in den Umgang mit regulären Ausdrücken“

Sitzung 1

Thorsten Fritze
t.fritze@ub.uni-frankfurt.de

UB Frankfurt / Fachinformationsdienst Linguistik

15. Januar 2020

Gliederung

Beispiel I: Leerzeichen entfernen

```
␣Ein␣Werwolf␣eines␣Nachts␣entwich
␣␣von␣Weib␣und␣Kind,␣und␣sich␣begab
␣␣␣an␣eines␣Dorfschullehrers␣Grab
␣␣␣und␣bat␣ihn:␣Bitte,␣beuge␣mich!

␣␣Der␣Dorfschulmeister␣stieg␣hinauf
auf␣seines␣Blechschilds␣Messingknauf
und␣sprach␣zum␣Wolf,␣der␣seine␣Pfoten
␣␣␣geduldig␣kreuzte␣vor␣dem␣Toten:

"Der␣Werwolf",-␣sprach␣der␣gute␣Mann,
␣␣␣"des␣Weswolfs,␣Genitiv␣sodann,
dem␣Wemwolf,␣Dativ,␣wie␣man's␣nennt.
␣den␣Wenwolf,-␣damit␣hat's␣ein␣End'."

...

(Christian␣Morgenstern,␣"Der␣Werwolf")
```

Regex

```
'^␣+'
```

```
http://www.zeno.org/Literatur/M/Morgenstern,

+Christian/Gedichte/Galgenlieder/5./Der+

Werwolf
```

Beispiel II: Datumsangaben finden

Ausschnitt aus dem Wikipedia-Artikel: „Römische Republik“

Durch die Einrichtung der Provinzen Gallia cisalpina (203 v. Chr.) sowie Hispania citerior und Hispania ulterior (197 v. Chr.) stieg die Zahl der außeritalischen Verwaltungsgebiete auf fünf an, 168 v. Chr. kam mit der Provinz Illyricum ein sechstes hinzu.
Bereits 200 v. Chr. hatte Rom in Griechenland gegen die Hegemonie Makedoniens unter dem Antigoniden Philipp V. interveniert und ihn im Zweiten Makedonisch-Römischen Krieg besiegt. 196 v. Chr. wurde Griechenland vom philhellenisch gesinnten Titus Quinctius Flamininus für frei erklärt. ...
(`https://de.wikipedia.org/w/index.php?title=R%C3%B6mische_Republik&oldid=186811996`)

Regex

```
\d{3}\s*v\.\s*Chr\.
```

Beispiel III: XML/HTML Tags entfernen

```
...
<lg>
<lb xml:id="tg346.2.16"/>
<pb n="275" xml:id="tg346.2.17"/>
<l xml:id="tg346.2.18">
    Dem Werwolf schmeichelten die Fälle,
</l>
<l xml:id="tg346.2.19">
    er rollte seine Augenbälle.
</l>
<l xml:id="tg346.2.20">
    Indessen, bat er, füge doch
</l>
<l xml:id="tg346.2.21">
    zur Einzahl auch die Mehrzahl noch!
</l>
</lg>
...
```

Regex

```
<[^>]+>
```

http://hdl.handle.net/hdl:11858/00-1734-0000-0004-3A69-5

Abbildung: https://xkcd.com/208/

Einführung I

Zweck & Verwendung

Was sind „reguläre Ausdrücke“ und wozu kann man sie verwenden?

Reguläre Ausdrücke...

- ... sind eine Notation zur formalen beschreibung von Zeichenketten
- ... gestatten das Suchen nach und Ersetzen von Mustern in Texten
- ... werden von vielen Standardprogrammen unterstützt
- ... werden oft als „Regex“ abgekürzt
- ... existieren in diversen Varianten (sog. „flavours“)

Einführung II

Anwendungszwecke

„Reguläre Ausdrücke“ sind zwar vor allem in der IT verbreitet, jedoch für alle Personen nützlich, die regelmäßig mit text-basierten Daten arbeiten.

- … Normalisieren von Text
- … Komplexe Suchen
- … Validieren (z. B. E-Mail-Adressen, ISBNs, Telefonnummern, etc.)
- … …

Einzelne Zeichen

Wie bei einer gewöhnlichen Suche, ist es auch bei der Verwendung von regulären Ausdrücken möglich, nach konkreten Zeichen zu suchen.
Um *ein beliebiges* Zeichen zu suchen, wird stattdessen ein Punkt verwendet.

Regex

de. →der, dem, des, den
.ag →lag, Tag, sag, mag ...

Selbstdefinierte Zeichenklassen

Statt beliebiger Zeichen können auch Zeichenmengen definiert werden.
Diese werden in [eingeschlossen].

Regex

de[mnrs]→der, dem, des, den

Selbstdefinierte Zeichenklassen (fort.)

Selbstdefinierte Zeichenklassen können negiert werden

Regex

de[^mn]→**nicht** dem, den

Vordefinierte Zeichenklassen

Metazeichen	Bedeutung
\s	Leerzeichen & Tabulatoren
\d	Ziffern 0-9
\w	Buchstaben, Ziffern, ggf. Bindestrich

Regex

\d\d\d →123, 023, 999

\d\d\w →12a, 19x, 00a

Vordefinierte Zeichenklassen (fort.)

Metazeichen	Bedeutung
\S	Alles außer Leerzeichen & Tabulatoren
\D	Alles außer Ziffern 0-9
\W	Alles außer Buchstaben, Ziffern, ggf. Bindestrich

Regex

\S\s\S →A b, b 1, 3 3

Quantoren

Oft möchte man mehrere Zeichen(klassen) wiederholen. Dazu verfügen reguläre Ausdrücke über sog. **Quantoren**.

Metazeichen	Bedeutung
*	Null- oder mehrmals
?	Null- oder einmal
+	Eins- oder mehrmals
{5}	genau fünf mal
{2, 5}	Zwei bis fünf mal
{5, }	Fünf mal oder mehr
{, 5}	Höchstens fünf mal

Quantoren (fort.)

Lange Wörter

rumathunaradidillifaititillibumullunukkunun
neverheedthemhorseluggarsandlisteltomine
Milkinghoneybeaverbrooker
morphomelosophopancreates
ppatupperstrippuckputtanach
quasicontribusodalitarian
...
(Aus: James Joyce, *Finnegan's Wake*
`https://archive.org/details/in.ernet.dli.2015.463592`)

Reg-Ex

```
\W[\w-]{25,}\W
```

Anker

Anker erlauben es, reguläre Ausdrücke relativ zu bestimmten Regionen in einer Zeichenkette zu definieren.

Metazeichen	Bedeutung
^	Zeichekettenanfang
$	Zeichenkettenende

Optionen

Datumsangaben

Den 3. September 1786.
Muenchen, den 6. September.
Mittenwald, den 7. September, abends.
..
Venedig, den 14. Oktober, 2 Stunden in der Nacht.
...
Rom, den 2. Dezember 1786.
...
Den 16. Februar 1787.
(Aus: Johann Wolfgang von Goethe, *Italienische Reise: Band 1*
`https://archive.org/details/italienischereis02404gut`)

Regex

```
^.*\d{1,2}\.\s+(?:Januar|Februar|März|April|Mai|Juni
|Juli|August|September|Oktober|November|Dezember).*$
```

Steuerzeichen

Metazeichen	Bedeutung
\n	Zeilenvorschub
\t	Tabulator
\r	Wagenrücklauf

Capturing Groups & Referenzen

Reguläre Ausdrücke erlauben es, sich einzelne Teiltreffer zu merken und diese wieder zu verwenden. Dazu schreibt man den Suchbegriffe in (Klammern). Auf die einzelnen „Matches“ kann dann so zugegriffen werden:
\1, \2, ...\N

Regex

1. (bla)\1 →blabla
2. (Senckenberg), (Johann Christian) →\2 \1 = Johann Christian Senckenberg

Ausblick

Reguläre Ausdrück bieten eine Reihe weiterer Möglichkeiten, die hier nicht angesprochen worden sind. Etwa:

- Kontext-sensitive Suche mit sog. „Look-Around-Assertions“ (z. B. „Positive-Look-Ahead:“ \\w+\s(?=Bach))
- „Genügsame Quantoren“, die möglichst minimal matchen (z. B. *?)
- Weitere Anker (z. B. \B, \<, \>, ...)
- Andere Zeichenklassendefinition (z. B. Posix-Klassen: [:lower:], [:cntrl:], ...)