

„Einführung in den Umgang mit regulären Ausdrücken“

Sitzung 2

Thorsten Fritze

`t.fritze@ub.uni-frankfurt.de`

UB Frankfurt / Fachinformationsdienst Linguistik

17. Mai 2019

Gliederung

- 1 Wiederholung
- 2 Weitere Bestandteile (fort.)

Was sind Reguläre Ausdrücke

Zweck & Verwendung

Reguläre Ausdrücke erlauben es, Textmuster zu beschreiben und diese zum Suchen/Finden, Ersetzen, Validieren, etc. textbasierter Daten zu verwenden.

Metazeichen	Bedeutung
^	Zeichkettenanfang
\$	Zeichenkettenende
\b	Wortgrenzen
.	Beliebiges Zeichen
[a-zA-Z0-9]	Zeichenauswahl
[^a-zA-Z0-9]	Negierte Zeichenauswahl
\s	Leerzeichen & Tabulatoren
\d	Ziffern 0-9
\w	Buchstaben, Ziffern, ggf. Bindestrich
\S	Alles außer Leerzeichen & Tabulatoren
\D	Alles außer Ziffern 0-9
\W	Alles außer Buchstaben, Ziffern, ggf. Bindestrich
*	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

Tabelle: Bestandteile regulärer Ausdrücke (in Auswahl)

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 Ausdrücke 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:]]`, ...)