

5 Theorieaufgaben 5 - Reguläre Ausdrücke

Zum Warmwerden

Diese „Aufgabe“ ist nicht bewertet, aber kann von jedem genutzt werden, um den Umgang mit regEx (und das Erkennen von Mustern) zu trainieren.

Viele Leute sehen in regEx eine ähnliche Herausforderung wie in Sudokus.

<http://regex.alf.nu/> und <https://regexcrossword.com/>

Um eigene regex an Beispieltexten zu testen, gibt es außerdem: <http://pythex.org/>

5.1 Theorie: Funktionen zu RE 1

Erklären Sie möglichst einfach (gerne durch Bilder oder Graphen, immer mit mind. einem Beispiel!) die folgenden Funktionen und jeweils den Unterschied zwischen den Paaren:

- `re.findall()` und `re.finditer()`
- `re.sub()` und `re.subn()`
- `re.search()` und `re.match()`
- `regex.finditer()` und `re.finditer()`

5.2 Theorie: Funktionen zu RE 2

- `regex.search()` und `regex.findall()`
- `regex.groups` und `regex.match()`
- `regex.match()` und `regex.fullmatch()`

5.3 Theorie: Pairs

Stellen Sie Beispiel 6.2.5.1 von <https://docs.python.org/3.5/library/re.html> vor (d.h. erklären Sie die regulären Ausdrücke und die Funktion der Methoden).

5.4 Theorie: [2-Personen Aufgabe] Writing a Tokenizer

Stellen Sie Beispiel 6.2.5.9 von <https://docs.python.org/3.5/library/re.html> vor (d.h. erklären Sie einerseits die regulären Ausdrücke und andererseits die Funktion der Methoden).

Diese Aufgabe können Sie zu **zweit** vorbereiten und vorstellen.

5.5 Theorie: Greedy vs. Nongreedy

BBCode (<https://de.wikipedia.org/wiki/BBCode>) ist eine Auszeichnungssprache ähnlich wie html oder xml.

Sie wird oft in Foren oder Chats verwendet. Beispielsweise wird Text, der fett geschrieben werden soll von `[b]` und `[/b]` eingeschlossen. Beispielsweise ergibt „Ein `[b]gro`ser`[/b]` Mann und ein `[i]d`unner`[/i]` Hund.“ in der Auswertung „Ein **großer** Mann und ein *dünner* Hund.“ Möchte man mit Hilfe eines regulären Ausdrucks alle so markierten Teile eines Textes finden, schreibt man vielleicht `'(\\[.*\\])'`.

Das Problem: Normalerweise matchen reguläre Ausdrücke greedy - so viel gerade geht. **Was bedeutet das für den regulären Ausdruck im Beispiel - was wird im Beispieltext gematcht? Warum?**

`.*` statt `.{0,1}` bewirkt das gewünschte Ergebnis. **Warum?**

5.6 Theorie: Dave-to-girl-ratio

In einigen Bereichen der Informatik an Universitäten oder Unternehmen ist die Zahl der Frauen so niedrig, dass statt einer Frauenquote die „Dave to girl ratio“ verwendet wird. Grob gesagt: *„To track Carnegie Mellon University's progress in attracting female undergraduates to its School of Computer Science, faculty, staff and students there have kept count of the number of men named Dave versus the number of women.“*

Dabei werden auch Davids, Daveys, Davies, und Davys gezählt, aber keine Davidas, Davettas oder Davinas.

- Schreiben Sie einen regulären Ausdruck, der aus der Liste der Angestellten eines Unternehmens (Vor- und Nachnamen) die (vermutlich) männlichen Daves (und Davids, Daveys, Davies, und Davys) herausucht.
- Finden Sie einen entsprechenden Ansatz für ein klassisch vor allem von Frauen besetztes Feld (Sekretärinnen/e, Arzthelfer/innen, Linguistik-Student(inn)en?). Recherchieren Sie häufige Frauennamen und ihre abgewandelten Formen und schreiben Sie einen passenden regulären Ausdruck *oder* erklären Sie, wieso das Verfahren mit (welchen) häufigsten Frauenvornamen (in Deutschland) nicht funktionieren würde.