Übungen 4: Reguläre Mustererkennung

Programmiertechniken in der Computerlinguistik II · Sommersemester 2005

1. Reguläre Suchmuster

Unterstreiche in folgenden Zeichenketten denjenigen Teil, der vom Suchmuster gematcht wird. Falls das Suchmuster nicht matcht, streiche die Zeichenkette durch. Die doppelten Hochkommata begrenzen die Zeichenketten, gehören aber nicht dazu. Du kannst deine Intuitionen unter http://www.cl.unizh.ch/clab/regex/ilap_regextut selbst überprüfen.

Muster	Zeichenketten		
`/^[0-9]/`	`"33"`	`" 78"`	`"-"`
`/\d+/`	`"a 019 9"`	`"1"`	`""`
`/[Dd](u|ü)fte?/`	`"feine Düfte"`	`"der Duft"`	`"dufter Duft"`
`/d.e/`	`"Dd.e"`	`"diee"`	`"..."`
`/.\.$/`	`"..."`	`"z.T."`	`"Fertig.!"`
`/^\^[^\^]/`	`"2^4"`	`"^44^"`	`"\^44"`
`/[a-z].+/`	`"a1"`	`"Aa"`	`"AA aa 4"`
`/[A-Z][^aeiou].*/`	`"B!"`	`"fliessen"`	`"blauer Ski"`
`/(\d{2,3}\s?){2,4}/`	`"Tel. 01 364 72 18"`		`"01 02 03 1 "`

2. Englische Datums- und Zeitangaben erkennen

Unter http://www.cl.unizh.ch/clab/regex/ilap_datum sollen im englischen Text zu Apollo möglichst alle Angaben markiert werden, welche ein Datum, einen Zeitpunkt oder eine Zeitdauer bezeichnen.

3. Die Sprache des Prolog-Tokenizers

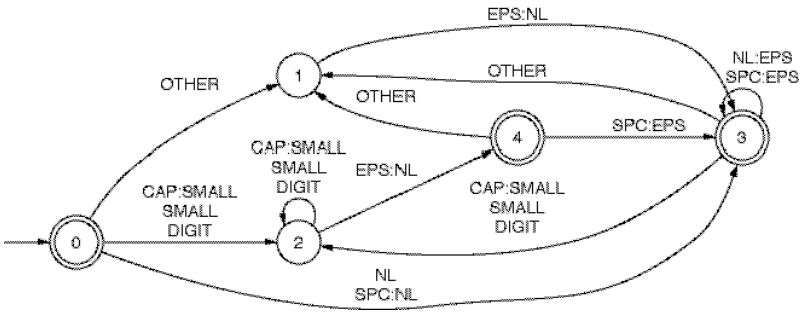
Regula S. behauptet, dass alle Tokens, welche der Prolog-Tokenizer aus der Vorlesung als Resultat zurückliefern kann, sich durch folgenden regulären Ausdruck beschreiben lassen:

```
([a-z0-9]+|[^a-z0-9\n\r ])
```

Was meinst du zu dieser Meinung?

4. Zu vertikalisiertem Text tokenisieren

Der Onkel von Regula S., eine Person namens X. Rox, behauptet, dass er einen Transduktor mit nur 5 Zuständen erfunden hat, welcher die gleiche Funktionalität wie der Prolog-Tokenizer aufwiese. Der einzige Unterschied wäre, dass sein Transduktor keine Prolog-Liste der Token zurückliefern würde, sondern sogenannten vertikalisierten Text. – d.h. jedes Token steht in Textform auf einer separaten Zeile.
Der Onkel hat Regula nebst einer Skizze auf einer Papierserviette nur noch folgende Hinweise gegeben: EPS steht für Epsilon, NL für Zeilenende, SPC für Leerzeichen (auch Tabulator, Wagenrücklauf etc.), DIGIT für Ziffern, CAP für Grossbuchstaben und SMALL für Kleinbuchstaben. CAP:SMALL für alle Paare von Grossbuchstaben mit dem entsprechenden Kleinbuchstaben. Etwas speziell sind die Übergänge, welche mit OTHER beschriftet sind: dort sind alle Symbole erlaubt, welche nicht unter eine der andern im Transduktor vorkommenden Kanten-Beschriftungen fallen.



Kann der Transduktor wirklich das leisten, was Regulas Onkel verspricht?