

Coursework 1

This coursework is worth 3% and is due on 12 November at 16:00. You are asked to implement a regular expression matcher and submit a document containing the answers for the questions below. You can do the implementation in any programming language you like, but you need to submit the source code with which you answered the questions. However, the coursework will *only* be judged according to the answers.

The task is to implement a regular expression matcher based on derivatives. The implementation should be able to deal with the usual regular expressions

$$\emptyset, \epsilon, c, r_1 + r_2, r_1 \cdot r_2, r^*$$

but also with

$[c_1c_2 \dots c_n]$	a range of characters
r^+	one or more times r
$r^?$	optional r
$r^{\{n,m\}}$	at least n -times r but no more than m -times
$\sim r$	not-regular expression of r

In the case of $r^{\{n,m\}}$ we have the convention that $0 \leq n \leq m$. The meaning of these regular expressions is

$$\begin{aligned} L([c_1c_2 \dots c_n]) &\stackrel{\text{def}}{=} \{ "c_1", "c_2", \dots, "c_n" \} \\ L(r^+) &\stackrel{\text{def}}{=} \bigcup_{1 \leq i} L(r)^i \\ L(r^?) &\stackrel{\text{def}}{=} L(r) \cup \{ "" \} \\ L(r^{\{n,m\}}) &\stackrel{\text{def}}{=} \bigcup_{n \leq i \leq m} L(r)^i \\ L(\sim r) &\stackrel{\text{def}}{=} UNIV - L(r) \end{aligned}$$

whereby in the last clause the set $UNIV$ stands for the set of *all* strings. So $\sim r$ means ‘all the strings that r cannot match’. We assume ranges like $[a-z0-9]$ are a shorthand for the regular expression

$$[abcd \dots z01 \dots 9].$$

Be careful that your implementations for *nullable* and *der* satisfies for every r the following two properties:

- $nullable(r)$ if and only if $"" \in L(r)$
- $L(\text{der } cr) = \text{Der } c(L(r))$

