Automata and Formal Languages (3)

Email:	christian.urban at kcl.ac.uk
Office:	S1.27 (1st floor Strand Building)
Slides:	KEATS (also home work and course-work is there)

Regular Expressions

In programming languages they are often used to recognise:

- symbols, digits
- identifiers
- numbers (non-leading zeros)
- keywords
- comments

http://www.regexper.com

Last Week

Last week I showed you a regular expression matcher which works provably correctly in all cases.

$$\textit{matcher}\ r\ s \quad \text{if and only if} \quad s \in L(r)$$

by Janusz Brzozowski (1964)

The Derivative of a Rexp

$$
\begin{array}{lcl}
der\,c\,(\varnothing) & \stackrel{\text{def}}{=} & \varnothing \\
der\,c\,(\epsilon) & \stackrel{\text{def}}{=} & \varnothing \\
der\,c\,(d) & \stackrel{\text{def}}{=} & \text{if } c = d \text{ then } \epsilon \text{ else } \varnothing \\
der\,c\,(r_1 + r_2) & \stackrel{\text{def}}{=} & der\,c\,r_1 + der\,c\,r_2 \\
der\,c\,(r_1 \cdot r_2) & \stackrel{\text{def}}{=} & \text{if } nullable(r_1) \\
 & & \text{then } (der\,c\,r_1) \cdot r_2 + der\,c\,r_2 \\
 & & \text{else } (der\,c\,r_1) \cdot r_2 \\
der\,c\,(r^*) & \stackrel{\text{def}}{=} & (der\,c\,r) \cdot (r^*)
\end{array}
$$

The Derivative of a Rexp

$$
\begin{array}{lcl}
der\,c\,(\varnothing) & \stackrel{\text{def}}{=} & \varnothing \\
der\,c\,(\epsilon) & \stackrel{\text{def}}{=} & \varnothing \\
der\,c\,(d) & \stackrel{\text{def}}{=} & \text{if } c = d \text{ then } \epsilon \text{ else } \varnothing \\
der\,c\,(r_1 + r_2) & \stackrel{\text{def}}{=} & der\,c\,r_1 + der\,c\,r_2 \\
der\,c\,(r_1 \cdot r_2) & \stackrel{\text{def}}{=} & \text{if } nullable(r_1) \\
 & & \text{then } (der\,c\,r_1) \cdot r_2 + der\,c\,r_2 \\
 & & \text{else } (der\,c\,r_1) \cdot r_2 \\
der\,c\,(r^*) & \stackrel{\text{def}}{=} & (der\,c\,r) \cdot (r^*) \\
ders\,[]\,r & \stackrel{\text{def}}{=} & r \\
ders\,(c::s)\,r & \stackrel{\text{def}}{=} & ders\,s\,(der\,c\,r)
\end{array}
$$

To see what is going on, define

$$Der\,c\,A \stackrel{\text{def}}{=} \{s \mid c::s \in A\}$$

For $A = \{"foo", "bar", "frak"\}$ then

$$\begin{array}{lcl} Der\,f\,A & = & \{"oo", "rak"\} \\ Der\,b\,A & = & \{"ar"\} \\ Der\,a\,A & = & \varnothing \end{array}$$

The Idea of the Algorithm

If we want to recognise the string "abc" with regular expression r then

1. $Der\, a\, (L(r))$

The Idea of the Algorithm

If we want to recognise the string $"abc"$ with regular expression r then

1. $Der\, a\, (L(r))$
2. $Der\, b\, (Der\, a\, (L(r)))$

The Idea of the Algorithm

If we want to recognise the string $"abc"$ with regular expression r then

1. $Der\,a\,(L(r))$
2. $Der\,b\,(Der\,a\,(L(r)))$
3. $Der\,c\,(Der\,b\,(Der\,a\,(L(r))))$

The Idea of the Algorithm

If we want to recognise the string $"abc"$ with regular expression r then

1. $Der\, a\, (L(r))$
2. $Der\, b\, (Der\, a\, (L(r)))$
3. $Der\, c\, (Der\, b\, (Der\, a\, (L(r))))$
4. finally we test whether the empty string is in this set

The Idea of the Algorithm

If we want to recognise the string $"abc"$ with regular expression r then

1. $Der\, a\, (L(r))$
2. $Der\, b\, (Der\, a\, (L(r)))$
3. $Der\, c\, (Der\, b\, (Der\, a\, (L(r))))$
4. finally we test whether the empty string is in this set

The matching algorithm works similarly, just over regular expression instead of sets.

Input: string $"abc"$ and regular expression r

1. $der\,a\,r$
2. $der\,b\,(der\,a\,r)$
3. $der\,c\,(der\,b\,(der\,a\,r))$

Input: string "abc" and regular expression r

1. $der\, a\, r$
2. $der\, b\, (der\, a\, r)$
3. $der\, c\, (der\, b\, (der\, a\, r))$

4. finally check whether the last regular expression can match the empty string

We proved already

$$nullable(r) \text{ if and only if } "" \in L(r)$$

by induction on the regular expression.

We proved already

$$nullable(r) \text{ if and only if } \text{""} \in L(r)$$

by induction on the regular expression.

Any Questions?

We need to prove

$$L(der\, c\, r) = Der\, c\, (L(r))$$

by induction on the regular expression.

Proofs about Rexps

- P holds for $\varnothing$, ϵ and c
- P holds for $r_1 + r_2$ under the assumption that P already holds for r_1 and r_2.
- P holds for $r_1 \cdot r_2$ under the assumption that P already holds for r_1 and r_2.
- P holds for r^* under the assumption that P already holds for r.

Proofs about Natural Numbers and Strings

- P holds for 0 and
- P holds for $n+1$ under the assumption that P already holds for n

- P holds for "" and
- P holds for $c::s$ under the assumption that P already holds for s

Languages

A language is a set of strings.

A regular expression specifies a language.

A language is regular iff there exists a regular expression that recognises all its strings.

Languages

A language is a set of strings.

A regular expression specifies a language.

A language is regular iff there exists a regular expression that recognises all its strings.

not all languages are regular, e.g. $a^n b^n$.

Regular Expressions

$$\begin{array}{rcll} r & ::= & \varnothing & \text{null} \\ & | & \epsilon & \text{empty string / "" / []} \\ & | & c & \text{character} \\ & | & r_1 \cdot r_2 & \text{sequence} \\ & | & r_1 + r_2 & \text{alternative / choice} \\ & | & r^* & \text{star (zero or more)} \end{array}$$

How about ranges $[a\text{-}z]$, r^+ and $\sim r$? Do they increase the set of languages we can recognise?

Negation of Regular Expr's

- $\sim r$ (everything that r cannot recognise)
- $L(\sim r) \stackrel{\text{def}}{=} UNIV - L(r)$
- $nullable(r) \stackrel{\text{def}}{=} not(nullable(r))$
- $der\,c\,(\sim r) \stackrel{\text{def}}{=} \sim (der\,c\,r)$

Negation of Regular Expr's

- $\sim r$ (everything that r cannot recognise)
- $L(\sim r) \stackrel{\text{def}}{=} UNIV - L(r)$
- $nullable(r) \stackrel{\text{def}}{=} not(nullable(r))$
- $der\, c\, (\sim r) \stackrel{\text{def}}{=} \sim (der\, c\, r)$

Used often for comments:

$$/ \cdot * \cdot (\sim ([a\text{-}z]^* \cdot * \cdot / \cdot [a\text{-}z]^*)) \cdot * \cdot /$$

Regular Exp's for Lexing

Lexing separates strings into "words" / components.

- Identifiers (non-empty strings of letters or digits, starting with a letter)
- Numbers (non-empty sequences of digits omitting leading zeros)
- Keywords (else, if, while, ...)
- White space (a non-empty sequence of blanks, newlines and tabs)
- Comments

Automata

A deterministic finite automaton consists of:

- a set of states
- one of these states is the start state
- some states are accepting states, and
- there is transition function

  which takes a state as argument and a character and produces a new state

  this function might not always be defined

Automata

A deterministic finite automaton consists of:

- a set of states
- one of these states is the start state
- some states are accepting states, and
- there is transition function

  which takes a state as argument and a character and produces a new state

  this function might not always be defined