Advanced Compilers

COMP 240 Fall 2002

Written Assignment #1

Assigned: Fri Aug 30, 2002
Due: Thu Sep 5, 2002

Suppose we have a collection of regular expressions r specifying a scanner over some alphabet Σ. The combined length (in symbols) of the regular expressions is $|r|$. We also have a string $w \in \Sigma^*$ to be scanned into tokens.

(a) The scanner NFA M constructed from r will have $O(|r|)$ states, while the scanner DFA M' constructed from M may have $O(2^{|r|})$ states. Give an example specification r for which this upper bound is achieved, that is, the minimal scanner DFA for r has $\theta(2^{|r|})$ states. Discuss whether this asymptotically worst-case DFA size is likely to occur with the lexical structure of a typical programming language.

(b) The scanner DFA M' may take $O(|w|^2)$ time to tokenize w, because it may read more input than the length of the token it returns each time, even without right context. Give an example specification r (without right context) and input string w for which this upper bound is achieved, i.e. the scanning time is $\theta(|w|^2)$. Discuss whether this worst case scanning behavior can occur with the lexical structure of a typical programming language and a typical input.