
Quiz

CS-320, Computer Language Processing, Fall 2015

Wednesday, November 18th, 2015

Version 1

General notes about this quiz

- Have your CAMIPRO card ready on the desk.
- You are allowed to use any printed material (using standard fonts, no cursive or script fonts) that you brought yourself to the exam. You are not allowed to use any notes that were not typed-up. Also, you are not allowed to exchange notes or anything else with other students taking the quiz.
- Use separate sheets, for each question, to write your answers. No sheet of paper should contain answer to two or more questions at the same time.
- Make sure you write your name on each sheet of paper.
- Use a permanent pen with dark ink.
- It is advisable to do the questions you know best first.
- You have in total **3 hours 40 minutes**.

Exercise	Points	
1	25	
2	25	
3	25	
4	25	
Total	100	

Prelude. In this quiz, you consider the scenario where you are hired by a new Swiss start-up company that aims to disrupt the smartwatch market. The main product is a wearable device that runs apps in a variant of HTML5 so energy-efficiently that it can be powered solely by absorbing the heat of the person wearing it. You are in charge of building the compiler for this platform. You soon discover that using off-the-shelf regular expression libraries depletes a fully charged watch battery before processing `<title>Textme</title>`, so you need to build the compiler making use of what you learned in the class.

Problem 1: Lexer for XML (25 points)

The goal of this exercise is to develop a lexer for XML. The input to the lexer is a stream of characters belonging to the set $\Sigma = \{a, \dots, z, A, \dots, Z, 0, \dots, 9, <, >, /, \{, :, \}, ", !, -, WS\}$, where *WS* denotes a white space character. As a shorthand, we will use *letter* to denote any character in $\{a, \dots, z, A, \dots, Z\}$, *digit* to denote any of $0, \dots, 9$, and *special* to denote any of $\{, \}, \cdot$. (Note that *special* does not include other non-alpha-numeric characters like $<, >, ", !, -$.)

Your goal is to construct a lexer (i.e. an automaton) that tokenizes the input stream into the tokens listed below. The tokens are shown on the left and their defining regular expressions are shown on the right.

Token name	Regular expression
OP	<code><</code>
CL	<code>></code>
OPSL	<code>< /</code>
CLSL	<code>/ ></code>
EQ	<code>=</code>
NAME	<code>letter(letter digit)*</code>
NONNAME	<code>(digit special)(letter digit special)*</code>
STRING	<code>"(letter digit special)*"</code>
COMMENT	<code><!-- (letter digit special)* --></code>
SKIP	<code>WS</code>

- a) [15 pts] Construct the labelled DFA described in the lectures for the tokens define above. Note that every final state should be labelled by the token class(es) it accepts.

Consider the following XML string.

```
<jsonmessage>
  <!--CommunicationOfJSoNObjects-->
  <from ip="">EPFLserver</from>
  <message>{"field":1}</message>
</jsonmessage>
```

- b) [10 pts] Show the list of tokens that should be generated by the lexer for the above XML string. You should use the *Maximum Munch Rule* to tokenize. You **need not** show SKIP tokens, which correspond to whitespaces.

Problem 2: Parser for XML (25 points)

In this exercise, you are required to design an LL(1) parser for XML. Consider the following grammar that describes syntactically correct XML strings. The terminals of the grammar are the tokens that would be generated by the lexer (which you designed in the previous exercise).

element	→	OP NAME attributes CL content OPSL NAME CL
		OP NAME attributes CLSL
attributes	→	attributes attr ϵ
attr	→	NAME EQ STRING
content	→	textOpt body
textOpt	→	NAME NONNAME STRING ϵ
body	→	body elemOrComment elemOrComment
elemOrComment	→	element COMMENT

Remark: the above grammar is a simplified version of the Antlr v4 XML grammar available at <https://github.com/antlr/grammars-v4/xml/XMLParser.g4>

- a) [25 pts] Convert the above grammar to an equivalent grammar that is LL(1). Show the *Nullable* non-terminals of **your** grammar, and the *First* and *Follow* sets of each non-terminal.

Problem 3: CNF grammar for JSON (25 points)

Consider the following grammar that accepts JSON (JavaScript Object Notation) strings.

json	→	object
object	→	{ pairs } { }
pairs	→	pairs STRING : value ϵ
value	→	STRING NUMBER object

- a) [10 pts] Convert the grammar to Chomsky's Normal Form. Recall that a grammar in CNF should satisfy the following properties.

1. terminals t occur alone on the right-hand side: $X ::= t$
2. no productions of arity more than two
3. no nullable symbols except for the start symbol
4. no single non-terminal productions $X ::= Y$
5. no unproductive non-terminal symbols
6. no non-terminals unreachable from the start symbol

It is sufficient if you only write the final grammar that you obtain.

Problem 4: Intersection Types (25 points)

In this exercise, we will consider the notion of intersection of types. Let T_1 and T_2 be two types belonging to our language. An expression has an intersection type $T_1 \wedge T_2$ iff it can be typed as both T_1 and T_2 . Therefore, we have the following type rules.

$$\frac{\Gamma \vdash e : T_1 \quad \Gamma \vdash e : T_2}{\Gamma \vdash e : T_1 \wedge T_2} \quad \frac{\Gamma \vdash e : T_1 \wedge T_2}{\Gamma \vdash e : T_1} \quad \frac{\Gamma \vdash e : T_1 \wedge T_2}{\Gamma \vdash e : T_2}$$

We consider $T_1 \wedge T_2$ and $T_2 \wedge T_1$ to be the same. In the above rules T_1 and T_2 can also be function types like $R \rightarrow S$.

- a) [5 pts] If T_1 and T_2 are arbitrary types, consider the following three expressions denoting types: $T_1 \wedge T_2$, T_1 , and T_2 . State all subtyping relations that you believe should hold among the 3×3 possible pairs of expressions.

Enter <: if the type corresponding to the row is a subtype of the type corresponding to the column; enter / if this is not necessarily the case.

$\prec:$	T_1	T_2	$T_1 \wedge T_2$
T_1			
T_2			
$T_1 \wedge T_2$			

In the next part of the exercise, you are required to come up with a type derivation involving intersection types. Consider a language, similar to the one described in lecture 12, with arithmetic operations, if-else expressions, assignment and block statements, that has the following types: *Pos*, *Neg*, *Int* and *Bool*. (We provide all the type rules that you may need for this exercise at the end of this question.)

Consider the function f shown below. Γ_0 is the initial type environment before the beginning of the function.

```

 $\Gamma_0 = \{ \text{divk} : (Pos \rightarrow Pos) \wedge (Neg \rightarrow Neg) \}$ 
def  $f(x : \text{Int}) : \text{Int}$  {
  if( $x > 0$ )  $\text{divk}(x)$ 
  else
    if( $x < 0$ )  $\text{divk}(x)$ 
    else  $x$ 
}
```

divk is a function (e.g. like $x \Rightarrow 10/x$) that maps positive integers to positive integers and negative integers to negative integers. Observe that with intersection types we can type the function as $(Pos \rightarrow Pos) \wedge (Neg \rightarrow Neg)$.

- b) [20 pts] Complete the type derivation for the body of the function f , shown on page 5, by filling in the holes marked with ?. If the expression will not type check, show the step where the type derivation cannot proceed. You will need to use only the type rules of intersection types and the types rules shown below.

Type Rules:

$$\begin{array}{c}
\frac{(x, T) \in \Gamma}{\Gamma \vdash x : T} \quad \text{Pos} <: \text{Int} \quad \text{Neg} <: \text{Int} \quad \frac{\Gamma \vdash e : T_1 \quad T_1 <: T_2}{\Gamma \vdash e : T_2} \\
\\
\frac{\Gamma \vdash e : T}{\Gamma \vdash \{e\} : T} \quad \frac{s_1 : \text{Unit} \quad \Gamma \vdash \{s_2; \dots s_n\} : T}{\Gamma \vdash \{s_1; \dots ; s_n\} : T} \quad \frac{\Gamma \oplus \{(x, T)\} \vdash \{s_2; \dots s_n\} : T}{\Gamma \vdash \{\text{var } x : T; s_2; \dots ; s_n\} : T} \\
\\
\frac{\Gamma \vdash x : \text{Int} \quad \Gamma \oplus \{(x, \text{Pos})\} \vdash e_1 : T \quad \Gamma \vdash e_2 : T}{\Gamma \vdash \text{if}(x > 0) e_1 \text{ else } e_2 : T} \\
\\
\frac{\Gamma \vdash x : \text{Int} \quad \Gamma \oplus \{(x, \text{Neg})\} \vdash e_1 : T \quad \Gamma \vdash e_2 : T}{\Gamma \vdash \text{if}(x < 0) e_1 \text{ else } e_2 : T} \quad \frac{\Gamma \vdash b : \text{Bool} \quad \Gamma \vdash e_1 : T \quad \Gamma \vdash e_2 : T}{\Gamma \vdash \text{if}(b) e_1 \text{ else } e_2 : T} \\
\\
\frac{\Gamma \vdash e_1 : T_1 \quad \dots \quad \Gamma \vdash e_n : T_n \quad \Gamma \vdash g : (T_1 \times \dots \times T_n) \rightarrow T}{\Gamma \vdash g(e_1, \dots, e_n) : T}
\end{array}$$

$\Gamma_1 \vdash x : Pos$	$\frac{\quad}{?}$	$\Gamma_2 \vdash x : Neg$	$\frac{\quad}{?}$
	$\Gamma_1 \vdash \text{divk} : ?$		$\frac{\quad}{?}$
	$\frac{\quad}{\Gamma \vdash x : Int}$	$\frac{\quad}{?}$	
	$\frac{\quad}{?}$	$\Gamma_2 \vdash \text{divk}(x) : ?$	
	$\Gamma_1 \vdash \text{divk}(x) : Int$	$\Gamma \vdash \text{if}(x < 0) \text{ divk}(x) \text{ else } x : Int$	

$$\Gamma \vdash \text{if}(x > 0) \text{ divk}(x) \text{ else if}(x < 0) \text{ divk}(x) \text{ else } x : Int$$

where,

$$\Gamma_1 = ?$$

$$\Gamma_2 = ?$$