CS 331 Programming Languages, Spring 2013

Lexeme Description

Monday, February 4, 2013

We wish to write a lexical analyzer for a language with the following lexemes.

No lexeme contains a **whitespace** character (blank, tab, vertical-tab, new-line, carriage-return, form-feed). A whitespace character, or any contiguous group of whitespace characters, is considered as a separator between lexemes. However, pairs of lexemes are not *required* to be separated by whitespace.

Comments, which begin with a pound sign ("#") that is not part of a lexeme that began earlier, and end with the next newline, or the end of the input if there is no following newline, are treated as whitespace.

Legal characters outside comments are whitespace and printable ASCII (values 32 [blank] to 126 [tilde]). Any other characters outside comments are illegal.

Once a lexeme has begun, the complete lexeme is considered to be the longest substring beginning from the starting point that can be interpreted as a lexeme.

There are five tokens: **Keyword**, **Identifier**, **Operator**, **Punctuation**, **Number**.

Keyword

> One of "`func`", "`return`".

Identifier

> Begins with letter (upper- or lower-case) or underscore ("_"), contains only letters, digits, underscores, and is not a **Keyword**.

Operator

> One of "+", "-", "++", "--", "+=", "-=", ".", "=".

Punctuation

> Any single legal character that is not a letter, digit, underscore ("_"), or whitespace. And is not an **Operator**.

Number

> Where D represents a nonempty string of digits, a number has the form "D", "D.", ".D", "D.D" (the Ds may represent different strings), or any of these with a single "+" or "-" prepended.