Regular Expressions in JLex

To define a token in JLex, the user to associates a regular expression with commands coded in Java.

When input characters that match a regular expression are read, the corresponding Java code is executed. As a user of JLex you don't need to tell it *how* to match tokens; you need only say *what* you want done when a particular token is matched.

Tokens like white space are deleted simply by having their associated command not return anything. Scanning continues until a command with a return in it is executed.

The simplest form of regular expression is a single string that matches exactly itself.

For example,

```
if    {return new Token(sym.If);}
```

If you wish, you can quote the string representing the reserved word (**"if"**), but since the string contains no delimiters or operators, quoting it is unnecessary.

For a regular expression operator, like +, quoting is necessary:

```
"+"    {return
          new Token(sym.Plus);}
```

Character Classes

Our specification of the reserved word if, as shown earlier, is incomplete. We don't (yet) handle upper or mixed-case.

To extend our definition, we'll use a very useful feature of Lex and JLex—*character classes*.

Characters often naturally fall into classes, with all characters in a class treated identically in a token definition. In our definition of identifiers all letters form a class since any of them can be used to form an identifier. Similarly, in a number, any of the ten digit characters can be used.

Character classes are delimited by `[` and `]`; individual characters are listed without any quotation or separators. However `\`, `^`, `]` and `-`, because of their special meaning in character classes, must be escaped. The character class `[xyz]` can match a single `x`, `y`, or `z`.

The character class `[\])]` can match a single `]` or `)`.

(The `]` is escaped so that it isn't misinterpreted as the end of character class.)

Ranges of characters are separated by a `-`; `[x-z]` is the same as `[xyz]`. `[0-9]` is the set of all digits and `[a-zA-Z]` is the set of all letters, upper- and lower-case. `\` is the escape character, used to represent unprintables and to escape special symbols.

Following C and Java conventions, **`\n`** is the newline (that is, end of line), **`\t`** is the tab character, **`\\`** is the backslash symbol itself, and **`\010`** is the character corresponding to octal 10.

The **`^`** symbol complements a character class (it is JLex's representation of the Not operation).

`[^xy]` is the character class that matches any single character *except* **`x`** and **`y`**. The **`^`** symbol applies to all characters that follow it in a character class definition, so **`[^0-9]`** is the set of all characters that aren't digits. **`[^]`** can be used to match all characters.

Here are some examples of character classes:

Character Class	Set of Characters Denoted
`[abc]`	Three characters: **a**, **b** and **c**
`[cba]`	Three characters: **a**, **b** and **c**
`[a-c]`	Three characters: **a**, **b** and **c**
`[aabbcc]`	Three characters: **a**, **b** and **c**
`[^abc]`	All characters except **a**, **b** and **c**
`[\^\-\]]`	Three characters: **^**, **-** and **]**
`[^]`	All characters
`"[abc]"`	Not a character class. This is one five character *string*: **[abc]**

Regular Operators in JLex

JLex provides the standard regular operators, plus some additions.

- Catenation is specified by the juxtaposition of two expressions; no explicit operator is used. Outside of character class brackets, individual letters and numbers match themselves; other characters should be quoted (to avoid misinterpretation as regular expression operators).

Regular Expr	Characters Matched
`a b cd`	Four characters: **`abcd`**
`(a)(b)(cd)`	Four characters: **`abcd`**
`[ab][cd]`	Four different strings: **`ac`** or **`ad`** or **`bc`** or **`bd`**
`while`	Five characters: **`while`**
`"while"`	Five characters: **`while`**
`[w][h][i][l][e]`	Five characters: **`while`**

Case *is* significant.

- The alternation operator is `|`. Parentheses can be used to control grouping of subexpressions.
  If we wish to match the reserved word `while` allowing any mixture of upper- and lowercase, we can use
  `(w|W)(h|H)(i|I)(l|L)(e|E)`
  or
  `[wW][hH][iI][lL][eE]`

Regular Expr	Characters Matched
`ab|cd`	Two different strings: **`ab`** or **`cd`**
`(ab)|(cd)`	Two different strings: **`ab`** or **`cd`**
`[ab]|[cd]`	Four different strings: **`a`** or **`b`** or **`c`** or **`d`**

- Postfix operators:
  ***** Kleene closure: 0 or more matches.
  (ab)* matches λ or **ab** or **abab** or **ababab** ...

  + Positive closure: 1 or more matches.
  (ab)+ matches **ab** or **abab** or **ababab** ...

  ? Optional inclusion:
  expr?
  matches `expr` zero times or once.
  expr? is equivalent to **(expr)** | λ and eliminates the need for an explicit λ symbol.

  [-+]?[0-9]+ defines an optionally signed integer literal.

- Single match:
  The character "." matches any single character (other than a newline).
- Start of line:
  The character ^ (when used outside a character class) matches the beginning of a line.
- End of line:
  The character $ matches the end of a line. Thus,

  ^A.*e$

  matches an entire line that begins with A and ends with e.

Overlapping Definitions

Regular expressions map overlap (match the same input sequence).

In the case of overlap, two rules determine which regular expression is matched:

- The *longest possible* match is performed. JLex automatically buffers characters while deciding how many characters can be matched.
- If two expressions match *exactly* the same string, the earlier expression (in the JLex specification) is preferred. Reserved words, for example, are often special cases of the pattern used for identifiers. Their definitions are therefore placed before the expression that defines an identifier token.

Often a “catch all” pattern is placed at the very end of the regular expression rules. It is used to catch characters that don’t match any of the earlier patterns and hence are probably erroneous. Recall that " ." matches any single character (other than a newline). It is useful in a catch-all pattern. However, avoid a pattern like `.*` which will consume all characters up to the next newline.

In JLex an unmatched character will cause a run-time error.

The operators and special symbols most commonly used in JLex are summarized below. Note that a symbol sometimes has one meaning in a regular expression and an *entirely different* meaning

in a character class (i.e., within a pair of brackets). If you find JLex behaving unexpectedly, it's a good idea to check this table to be sure of how the operators and symbols you've used behave. Ordinary letters and digits, and symbols not mentioned (like @) represent themselves. If you're not sure if a character is special or not, you can always escape it or make it part of a quoted string.

Symbol	Meaning in Regular Expressions	Meaning in Character Classes
(	Matches with) to group sub-expressions.	Represents itself.
)	Matches with (to group sub-expressions.	Represents itself.
[	Begins a character class.	Represents itself.
]	Represents itself.	Ends a character class.
{	Matches with } to signal macro-expansion.	Represents itself.
}	Matches with { to signal macro-expansion.	Represents itself.
"	Matches with " to delimit strings (only \ is special within strings).	Represents itself.
\	Escapes individual characters. Also used to specify a character by its octal code.	Escapes individual characters. Also used to specify a character by its octal code.
.	Matches any one character except \n.	Represents itself.
|	Alternation (or) operator.	Represents itself.

Symbol	Meaning in Regular Expressions	Meaning in Character Classes
*	Kleene closure operator (zero or more matches).	Represents itself.
+	Positive closure operator (one or more matches).	Represents itself.
?	Optional choice operator (one or zero matches).	Represents itself.
/	Context sensitive matching operator.	Represents itself.
^	Matches only at beginning of a line.	Complements remaining characters in the class.
$	Matches only at end of a line.	Represents itself.
-	Represents itself.	Range of characters operator.