Universidad Simón Bolívar
Departamento de Computación y Tecnología de la Información
Traductores e Interpretadores
Trimestre Septiembre-Diciembre 2019

Etapa I: Análisis Lexicográfico

Esta entrega consiste en la implementación de un analizador lexicográfico para el lenguaje GuardedUSB, el cual constituye la primera fase del interpretador que se desea construir a lo largo del trimestre. Como entrada, debemos aceptar cualquier secuencia de caracteres, y como salida, los tokens relevantes reconocidos. Si no se reconocen los caracteres como parte de nuestro lenguaje, entonces debemos arrojar un error. Todo token debe ir acompañado de su número de fila y columna.

1. Tokens

Entre los tokens relevantes del lenguaje se encuentran:

- Cada palabra reservada del lenguaje GuardedUSB: `declare`, `if`, `do`, `od`, etc. En este caso, los tokens deben llamarse: `Tk<Palabra Clave>`, donde `<Palabra Clave>` es la palabra clave del lenguaje, con su primera letra en mayúscula. Por ejemplo, para la palabra clave `for`, su token sería `TkFor`.
- Los identificadores de variables. Todos los identificadores corresponden a un token llamado `TkId`. Este token siempre tendrá asociado como atributo el identificador reconocido. Ejemplo: al leer: `id`, su token correspondiente sería `TkId("id")`.
- Los números enteros. Ellos son agrupados bajo el token `TkNum`. Por ejemplo: `TkNum(30)`.
- Las cadenas de caracteres encerradas entre comillas. Debe ser un único *token* cuyo contenido sea la cadena leída. Ejemplo: al leer `"hola mundo"`, su token correspondiente sería `TkString("hola mundo")`.
- Las constantes booleanas. El usado para `true` será: `TkTrue` y el asociado a `false` será: `TkFalse`.
- Cada símbolo utilizado para denotar separadores, los cuales se presentan a continuación:
  - "`|[`" - `TkOBlock`
  - "`]|`" - `TkCBlock`
  - "`..`" - `TkSoForth`
  - "`,`" - `TkComma`
  - "`(`" - `TkOpenPar`
  - "`)`" - `TkClosePar`
  - "`:=`" - `TkAsig`
  - "`;`" - `TkSemicolon`
  - "`-->`" - `TkArrow`
- Cada símbolo utilizado para denotar operadores aritméticos, booleanos, relacionales o de manipulación de arreglos y cadenas de caracteres, los cuales se presentan a continuación:
  - "`+`" - `TkPlus`
  - "`-`" - `TkMinus`

- “*” - `TkMult`
- “/” - `TkDiv`
- “%” - `TkMod`
- “\/” - `TkOr`
- “/\” - `TkAnd`
- “!” - `TkNot`
- “<” - `TkLess`
- “<=” - `TkLeq`
- “>=” - `TkGeq`
- “>” - `TkGreater`
- “==” - `TkEqual`
- “!=” - `TkNEqual`
- “[” - `TkOBracket`
- “]” - `TkCBracket`
- “:” - `TkTwoPoints`
- “||” - `TkConcat`

- Cada símbolo utilizado para denotar funciones de conversión de tipos y embebidas, las cuales se presentan a continuación:
  - “atoi” - `TkAtoi`
  - “size” - `TkSize`
  - “max” - `TkMax`
  - “min” - `TkMin`

  Otras consideraciones:
- Los espacios en blanco, tabuladores, salto de línea y comentarios deben ser ignorados. Es inaceptable manejarlos como *tokens*.
- Se debe preservar la diferencia entre mayúsculas y minúsculas.

2. Formato de salida y entrada del programa

Su analizador lexicográfico debe llamarse `lexer` y recibirá como primer argumento el nombre del archivo a analizar con extensión `.gusb` (su programa `lexer` debe revisar si la extensión del archivo es correcta). La salida debe mostrar todos los tokens reconocidos de manera legible e incluyendo para cada uno, la línea y columna donde se encontró. Por ejemplo:

Suponiendo que el contenido del archivo `programa.gusb` es:

```
|[
  declare
    a, b, c : int;
    d, e, f : array[0..2]
    a := b + 3;
    read e
    // Esto es un comentario. Debe ser ignorado.
]|
```

Entonces al correr `lexer` con el argumento `programa.gusb`, se debe arrojar como salida la secuencia de tokens, todos acompañados de dos números (el primero de ellos la fila en donde se encuentra el token y el segundo la columna en donde empieza el mismo):

```
TkOBlock 1 1
TkDeclare 2 3
TkId("a") 3 5
TkComma 3 6
TkId("b") 3 8
TkComma 3 9
TkId("c") 3 11
TkTwoPoints 3 13
TkInt 3 15
TkSemicolon 3 18
TkId("d") 4 5
TkComma 4 6
TkId("e") 4 8
TkComma 4 9
TkId("f") 4 11
TkTwoPoints 4 13
TkArray 4 15
TkOBracket 4 20
TkNum("0") 4 21
TkSoForth 4 22
TkNum("2") 4 24
TkCBracket 4 25
TkId("a") 5 5
TkAsig 5 7
TkId("b") 5 10
TkPlus 5 12
TkNum("3") 5 14
TkSemicolon 5 15
TkRead 6 5
TkId("e") 6 10
TkCBlock 8 1
```

El analizador lexicográfico solo es capaz de reconocer secuencias arbitrarias de tokens, a pesar de que esas secuencias pueden estar sintácticamente incorrectas.

En cuanto a los errores; los únicos errores detectables a nivel lexicográfico, corresponden a frases incorrectas o mal formadas, que no corresponden a ningún token válido. Por ejemplo para el programa:

```
|[
  declare
   a, b, c : int;
   d, 1e, f : array[0..2]
   a :,= b + 3;
   read e
   // Esto es un comentario. Debe ser ignorado.
]|
```

Se debe imprimir

```
Error: Unexpected character "1" in row 4, column 7
Error: Unexpected character "=" in row 5, column 8
```

Su analizador lexicográfico debe reportar todos los errores léxicos, en caso de haberlos. Cuando un error es encontrado, los tokens se hacen irrelevantes (ya que no corresponden a un programa correcto), por lo que no deben ser mostrados.

3. Tecnologías

A continuación se explica cuáles son los lenguajes y herramientas permitidas, para que usted escoja cuál de ellas usará para la implementación de este proyecto. El lenguaje que usted seleccione para esta etapa será el lenguaje que utilizará para las siguientes etapas, no podrá cambiarlo:

- Python: Lenguaje de scripting, orientado a objetos.

  Para Python la herramienta generadora de analizadores lexicográficos a utilizar es a la vez la misma que genera analizadores sintácticos. Por lo tanto, deberán manejarse algunas nociones de gramáticas libres de contexto antes de tiempo para poder trabajar con la misma. Esta herramienta se llama *PLY* y puede ser encontrada desde la siguiente dirección Web: `http://www.dabeaz.com/ply/`

- Ruby: Lenguaje de scripting, orientado a objetos.

  En el caso particular de Ruby no hay una buena herramienta generadora de analizadores lexicográficos, por lo que el trabajo deberá hacerse manualmente a través de las expresiones regulares que provee el lenguaje.

- Java: Lenguaje imperativo, orientado a objetos.

  Para Java, la herramienta generadora de analizadores lexicográficos a utilizar se llama *ANTLR* y puede ser encontrada en la siguiente dirección Web: `http://www.antlr.org`

4. Detalles de la entrega

La entrega del proyecto es el martes de la semana 4 (8 de Octubre) por Aula Virtual. Su entrega debe incluir lo siguiente:

- La "Declaración de Autenticidad para Entregas" firmada por los integrantes del equipo.
- Un archivo comprimido `tar.gz` con el código fuente de su proyecto, debidamente documentado, colocado en el Aula Virtual. El nombre del archivo debe ser **Etapa1-XX-YY.tar.gz** donde **XX-YY** son los carné de los integrantes del grupo.

  El no cumplimiento de los requerimientos podría resultar en el rechazo de su entrega.

Federico Flaviani / federico.flaviani@gmail.com / Septiembre 2019