

Introducción

Las tareas básicas de un analizador lexicográfico consisten en buscar y reconocer los componentes léxicos de un lenguaje de programación en el programa fuente, según unas reglas o patrones. El analizador del léxico tiene que dividir la cadena de caracteres en palabras con significado propio y después convertirlo a una secuencia de símbolos terminales para el analizador sintáctico. El reconocimiento de las palabras se puede trabajar en función de una gramática regular, implementable mediante expresiones regulares.

Requerimiento

Diseñar y construir un programa para computadora que reciba como entrada un texto con la estructura de un lenguaje de programación y realice las tareas básicas de un analizador lexicográfico. El programa deberá estar diseñado para la sintaxis básica de alguno de los siguientes lenguajes de programación: C++, Python, C#, Swift, Perl, R o Kotlin.

Se debe analizar, como mínimo: uso de identificadores, palabras reservadas, operadores, caracteres o palabras de inicio y fin de bloque, constantes, cadenas de caracteres, entre otros. También se deben rechazar los elementos que no pertenecen a la sintaxis y omitir los comentarios del lenguaje seleccionado.

Adicionalmente al programa tokenizador, se deberá generar una tabla de símbolos utilizando una tabla hash o árboles de búsqueda. El objetivo de esta tabla es registrar los datos/atributos de cada símbolo del código fuente, que esté asociado con información como el tipo de datos, el ámbito de cada variable, valor de la constante, etc.

El resultado del programa resultante debe mostrar la lista de elementos lexicográficos encontrados y la tabla de símbolos construida. En caso que el programa detecte elementos que no pertenecen al léxico del lenguaje, entonces el grupo tiene la decisión de abortar la ejecución, indicando el error, o omitir el elemento y seguir buscando los otros tokens.

La mayoría de los lenguajes que se pueden elegir, tienen una sintaxis bastante diversa y compleja, además de poseer extensiones y librerías para ampliar su funcionamiento, favor limitarse a la sintaxis básica y estructuras de programación elementales del lenguaje.

Datos de equipo y entrega

Como lenguaje implementador para el analizador del léxico, pueden utilizar cualquier lenguaje de programación distinto al que seleccionaron para tokenizar. Favor seleccionar aquellos que sean interpretados o multiplataforma para asegurar la compatibilidad del software resultante.

La fecha y datos de entrega son:

Antes del 6 de octubre a las 23:59horas. Deberán subir su solución (fuentes, ejecutables, ejemplos, notas, documentos, etc.) a moodle.

Notas:

-> Es necesario entregar un documento con la mínima información de las reglas de los elementos del lenguaje que están tokenizando.

-> Seleccione 7 palabras reservadas como mínimo, máximo 15.

-> En caso de hacer un subconjunto de los elementos del lenguaje que vaya tokenizar, deberán entregar la documentación de cuáles tipos de elementos han tomado en cuenta.

-> Pueden utilizar cualquiera de las versiones de los lenguajes que se dan opción a elegir. Deberán escribir cuál versión están trabajando en los documentos que entreguen.