Analizador HTML

Presentado por:

Leonar Fabian Gonzales Cod: 1150894

Nelson Andrés Sepúlveda Cod: 1150831

Presentado A:

MSc. Marco Antonio Adarme Jaimes

Universidad Francisco de Paula Santander

Facultad de Ingeniería

Ingeniería de Sistemas

Cucuta-2014

Explicación

HTML, siglas de *Hyper Text Markup Lenguaje* («lenguaje de marcas de hipertexto»), hace referencia al lenguaje de marcado para la elaboración de páginas web. Es un estándar que sirve de referencia para la elaboración de páginas web en sus diferentes versiones, define una estructura básica y un código (denominado código HTML) para la definición de contenido de una página web, como texto, imágenes, etc. Es un estándar a cargo de la W3C, organización dedicada a la estandarización de casi todas las tecnologías ligadas a la web, sobre todo en lo referente a su escritura e interpretación. Es el lenguaje con el que se definen las páginas web. [1]

En el validador HTML se podrá editar documentos HTML donde el mismo usuario podrá agregar o incluir etiquetas como el desee siempre y cuando la etiqueta sea válida. También se podrá cargar la url de una página web siempre y cuando esta cuente con la extensión htm o html. Una vez el código fuente se encuentre en el editor el usuario podrá rectificarlo cuya función carga todo el texto en una fila de archivos que representa una Lista Circular Doble de datos de tipo String, luego se recibe el código fuente, donde se separa en un vector por medio de saltos de línea. Una vez en el vector termine con separar los saltos de línea recorre para ir agregando cada dato del vector dentro de la fila de archivos. Si el editor esta vacío se lanzara una excepción que indique el hecho. Luego de tener separadas todas las filas del editor se procede a identificar las etiquetas, para esto convertimos las filas de archivos en una cadena de caracteres, luego ubicamos los paréntesis angulares que nos permiten saber dónde inicia una etiqueta y donde tiene su fin. Así obtenemos todas las etiquetas del editor y nuevamente las incluimos dentro de las filas de archivos.

Luego de obtener todas las etiquetas debidamente separadas se proceden a su respectiva validación. Para ello se debe tomar en cuenta que el código fuente de un HTML cuenta con etiquetas tanto binarias como unarias, así que se procede a separarlas en una Lista Circular Doble que contiene dos Colas de datos de tipo String, la primera cola contendrá las etiquetas binarias y la segunda las etiquetas unarias. Primero se confirma si el doctype se encuentra dentro de la estructura básica el HTML si en algún momento no se encuentra esta etiqueta se enviara un error de etiqueta no reconocida a la pila de errores del sistema.

También se cuestiona si la forma básica del documento HTML se encuentra en orden, donde estas etiquetas se encuentran en una Cola de datos de tipo String, si en el caso que la cola de etiquetas básicas tenga algún dato se enviara un error de que no contiene etiqueta de estructura a la pila de errores del sistema.

Luego de separar cada una de las etiquetas binarias y unarias se procede a validar cada una de las colas primero se validan las binarias y luego damos paso a las unarias.

Para validar la binarias se crearon Pilas de tipo String conforme a la cantidad de errores que se pudiesen presentar, al momento de sacar una etiqueta pasara por una serie de comprobaciones que nos permitirá saber si dicha etiqueta atrae algún error o se encuentra bien escrita, por ejemplo si de la cola sale una etiqueta de apertura se comprueba utilizando expresiones regulares que nos dicen si dicha etiqueta se compone de los paréntesis angulares y un cuerpo valido, así mismo se comprueba si el cuerpo de la etiqueta se encuentra bien escrito recorriendo los tags cargados desde sandbox, así se almacenaría en una pila de datos de tipo String . De no ser así se incluiría en la pila de etiquetas no válidas. En el caso de una etiqueta de cierre se deberá comprobar si es válida utilizando expresiones regulares, luego se modifica dicha etiqueta para que al momento de buscarla en la Secuencia<TagGeneral> tags reporte que si se encuentra la etiqueta, así la etiqueta se almacenaría en una pila. Para modificar la etiqueta se retira el slash que referencia como etiqueta de cierre. Una vez la cola de etiquetas binarias se encuentre vacía se procederá a validar dos tipos de errores los cuales son que alguna etiqueta no tenga etiqueta de fin o alguna etiqueta no contenga etiqueta de inicio.

Ahora bien las pilas respectivas de cierre y apertura pueden o no contener elementos, en el caso que la pila de apertura no contenga elementos y la pila de cierre si contenga se generan tantos errores como sean la cantidad de elementos de la pila de cierre. En el caso contrario se generan tantos errores como sean la cantidad de elementos de la pila de apertura. De no cumplirse alguna de esas condiciones, las pilas contendrán los elementos ordenados respectivamente. Luego se comprueba sacando un elemento a la vez de la cola de etiquetas de cierres, así el dato será modificado, para luego sacar una etiqueta de apertura un comprobar si son iguales. En caso de no serlo se almacenara en una pila auxiliar y el dato seguirá comprobando en la pila de apertura hasta que encuentre su respectiva etiqueta. Si el dato no llega a encontrar su contraparte de apertura se genera un error de Inicio de etiqueta que se almacenara en la pila de errores del sistema.

Luego si la pila de apertura contiene elementos se generaran errores de que no contiene etiqueta de cierre, así se generaran tantos errores como elementos contenga la pila y serán almacenados en la pila de errores del sistema.

Una vez validadas las etiquetas binarias se procede a las unarias allí se recorre la cola de etiquetas unarias cuestionando si la etiqueta es válida, luego se modifica de modo

que su slash no lo contenga, después se pregunta si el cuerpo de la etiqueta resultante es válido. De no serlo se envía un error de etiqueta no reconocida a la pila de errores del sistema.

Referencias

[1] http://es.wikipedia.org/wiki/HTML