Trabajo Práctico
Teoría de Lenguajes
"Micro HTML Prettyprint - primera parte"

Primer cuatrimestre 2013

1. Introducción

HTML es un lenguaje de marcas utilizado para la inmensa mayoría de los contenidos en la World Wide Web. Básicamente, es un documento que contiene elementos (tags) HTML que pueden contener texto plano, otros elementos, o nada. Estos documentos son procesados por clientes comunmente conocidos como browsers.

El objetivo de este trabajo práctico es realizar un parser que interprete la estructura de los archivos HTML y que genere otro documento HTML que permita visualizar el HTML original de manera elegante (Prettyprint). Ciertos tags se verán de manera anidada mediante la indentación y los comentarios, tags, atributos y otros elementos de control se verán con diferentes colores.

Para cumplir con este objetivo dividimos el trabajo práctico en dos partes, la primera consiste en identificar cual será la gramática y los componentes léxicos de nuestro lenguaje y la segunda parte consistirá en realizar el parser junto con las reglas semánticas que permitirán generar el HTML resultante.

2. Lenguaje a reconocer

Lo primero a definir son los Tokens Léxicos, nuestro parser recibirá como entrada un archivo de texto que no deja de ser una secuencia de caracteres, el analizador léxico transforma esos caracteres agrupándolos en tokens que son a su vez los símbolos terminales de nuestra gramática. Dichos tokens deberán ser expresados como una expresión regular. Otra tarea del analizador léxico es la de ignorar los espacios o los comentarios, por lo que hay que definir tokens para estos casos.

Para simplificar el lenguaje ya que HTML es muy completo y extenso, sólo tendremos un subconjunto de tags válidos, no permitiremos atributos y todos los tags tienen su tag de cierre correspondiente (salvo
)

Los tags que vamos a admitir son: <html>, <head>, <body>, <title>, <script>, , <h1>, <p> y
. con las siguiente restricciones:

- Es obligatorio que el documento completo esté rodeado por tags de apertura y cierre <html>
- dentro del html puede tener una sección opcional <head> y otra sección también opcional <body>
- dentro de la sección head pueden aparecer indistintamente y en orden no preestablecido secciones <title> y <script>
- <title> contiene texto sin tags
- dentro de la sección <script> puede haber cualquier cosa salvo el tag de cierre de dicha sección

- dentro de laquilese seccion <body> tendremos bloques de texto con subsecciones (div, h1 o p) o con tags
 sueltos, las subsecciones se comportan como la sección <body>

3. Ejemplo

Esto es un HTML válido:

<html> <head><title>Esto es un título</title> <!– esto es un comentario –> <script> print("Hola mundo")</script></head> <body> Esto es una <h1>prueba</h1>
 </body></html>

El documento "prettyprinted"sería:

```
<html>
<head>
</title>Esto es un titulo</title>
<!– esto es un comentario –>¿
<script>
print("Hola mundo")
</script>
</head>
<body>
Esto es <p>una <h1>prueba</h1></p> <br>
</body>
</html>
```

4. Detalles de la entrega

Deben entregar un informe breve que incluya:

- descripción de la gramática,
- decisiones tomadas y su justificación,
- ejemplos de árboles de derivación con entradas correctas

El informe debe entregarse impreso, y además en un archivo comprimido a la dirección tptleng@gmail.com.

Fecha de entrega: 29 de Mayo de 2013.