Analizadores léxicos.
Aplicación a entornos web.

Teoría de Autómatas y lenguajes formales
Federico Simmross Wattenberg (fedesim@infor.uva.es)
Universidad de Valladolid

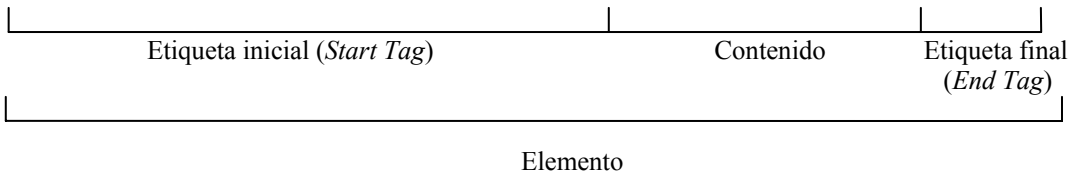
Para finalizar el tema del análisis léxico, vamos a utilizar la herramienta lex en un contexto que difiere considerablemente del análisis del código fuente C: el lenguaje HTML.

1. Introducción al HTML

HTML (HyperText Markup Language) es un lenguaje de marcas. Para que un lexema sea considerado palabra clave, debe ir delimitado por los caracteres *principio de marca* '<' y *fin de marca* '>'. Todo lo que no se encuentre entre estos dos caracteres se considera texto normal que no hay que interpretar.

Un documento en HTML consta de texto y elementos estructurales. El texto forma el contenido de la página web, y los elementos indican cómo organizarlo en la ventana del navegador. En general, un elemento de HTML tiene la siguiente sintaxis:

```
<nombre atrib1="valor1" atrib2="valor2"… >Contenido del elemento</nombre>
```



HTML no diferencia entra mayúsculas y minúsculas, así que un elemento <nombre> es igual que otro llamado <Nombre> ó <NOMBrE>.

Hay elementos que no necesitan la etiqueta final, otros no llevan contenido, e incluso para algunos elementos es ilegal escribir alguna de estas partes. De igual manera, para algunos elementos es obligatorio definir ciertos atributos. Cada elemento concreto tiene su propia sintaxis.

HTML es un subconjunto de un lenguaje más general: SGML (Standard Generalized Markup Language), y como tal, utiliza una parte de sus elementos estructurales. La especificación completa de HTML se puede encontrar en http://www.w3.org. Para esta práctica, sin embargo, vamos a centrarnos en los más imprescindibles.

Un documento en HTML debe estar encerrado en un elemento <HTML>. A su vez, el elemento <HTML> debe contener siempre una cabecera y un cuerpo, denotados, respectivamente, por los elementos <HEAD> y <BODY>. Veamos un primer ejemplo de documento en HTML:

```
<HTML>
    <HEAD>
        <TITLE>Documento Hola Mundo</TITLE>
    </HEAD>
    <BODY>
        Hola, Mundo
    </BODY>
</HTML>
```

Todo documento en HTML debería de incluir un elemento <TITLE> dentro de la sección de cabecera, que normalmente se muestra en el espacio reservado para el título de la ventana del navegador.

Para esta práctica, nos interesa especialmente el elemento <A> (*Anchor*), que sirve para establecer enlaces con otros documentos. El elemento <A> tiene la siguiente estructura:

```
< a …atributos… href="documento_destino" …atributos… >Descripción< /a >
```

Por ejemplo, el código HTML para la frase "Pulse en este enlace para continuar" sería:

```
…
Pulse en <a href="doc_destino.html">este enlace</a> para continuar
…
```

2. Ejercicios

1- Construir un programa que, apoyándose en un analizador léxico generado por lex, muestre por pantalla el destino y la descripción de los enlaces de un fichero en HTML. En la página de la asignatura hay un fichero de prueba.