

Linguaggi e Traduttori – Seconda parte

Compito del 6/9/2004

Si costruisca, utilizzando la coppia di programmi LEX e YACC (o Flex e Bison), un traduttore guidato dalla sintassi in grado di analizzare le tabelle in un documento HTML.

Linguaggio di ingresso

Un documento HTML è costituito da un testo ASCII annotato mediante opportune parole chiave. Tutte le parole chiave sono racchiuse tra i simboli "<" e ">". All'interno di tali simboli possono essere presenti anche eventuali modificatori e parametri delle parole chiave. Per le parole chiave ed i parametri è indifferente l'uso di lettere maiuscole o minuscole. I due caratteri delimitatori con il loro contenuto costituiscono un **tag**.

Un documento è composto da due sezioni: **intestazione** e **corpo**. L'intero documento è compreso tra le parole chiave "html" e "/html"; l'intestazione è racchiusa tra le parole chiave "head" e "/head"; il corpo è racchiuso tra le parole chiave "body" e "/body".

All'interno dell'intestazione può comparire il titolo del documento racchiuso tra le parole chiave "title" e "/title".

All'interno del corpo compare il testo del documento intercalato da varie parole chiave. Le parole chiave possono iniziare con un carattere alfabetico o con il carattere "/".

Le tabelle sono descritte dal tag "table" con parametro "border=number" che indica lo spessore del bordo della tabella stessa. Altri parametri sono possibili all'interno del tag "table" e l'ordine dei parametri non è fisso. Si supponga per semplicità che il parametro border sia obbligatorio e che compaia sempre per primo. Si noti anche che il valore numerico non compare tra virgolette "", come invece nel caso di altri parametri. Il valore 0 attribuito al bordo è lecito, ed indica una tabella senza bordi. Tutto ciò che segue il tag di apertura "table" e precede il tag di chiusura "/table" è il contenuto della tabella. La tabella può contenere all'interno altre tabelle, dette tabelle annidate, con la stessa sintassi.

Il traduttore deve riconoscere i commenti che seguono la sintassi dell'HTML, ovvero iniziano con la serie di caratteri "<!--" e terminano con i caratteri "-->", per evitare di considerare tabelle eventualmente presenti in un campo commentato.

Scopo del programma.

Il traduttore deve riportare su stdout alcune statistiche sulle tabelle.

In particolare deve visualizzare il numero totale di tabelle, lo spessore massimo, minimo e medio dei bordi e la massima "profondità" di tabelle annidate, ovvero il livello massimo di annidamento all'interno del documento.

Quando il programma incontra un errore, sia sintattico che semantico, deve segnalarlo e terminare l'esecuzione: non è richiesta la gestione ed il recupero degli errori. Ciò significa che il programma deve verificare che il documento HTML segua le regole del linguaggio d'ingresso (come specificate al punto precedente).

Esempio

Dato il seguente documento in ingresso:

```
<HTML><HEAD><TITLE>Prova</TITLE></HEAD>
<BODY>
<!-- .... <table>Finta tabella (da non contare)</table> -->
<H1>Titolo_1</h1>
<h2>Titolo_1_1</h2>
Testo <b>vario</b>
<H1>Titolo_2</h1>
<h2>Titolo_2_1</h2>
<table border=2><tr><td>Idem</td></tr></table>
<a href="top.html"></a>
<h2>Titolo_2_2</h2>
<table border=0><tr><td>
  <table border=0><tr><td> Tabella annidata livello 1</td></tr>
</table>
</td>
</tr>
</table>
<ul>
<li><a href="pippo.htm">pippo</a>
<li><a href="pluto.htm">pluto <i>pi&ugrave;</i> pippo</a>
</ul>
<table border=0><tr><td>
  <table border=0><tr><td>
    <table border=0><tr><td>Tabella annidata livello 2</td></tr>
    </table>
  </td></tr>
</table>
</table>

<hr>
<!-- Fine dell'esempio -->
</body></HTML>
```

Il traduttore dovrebbe visualizzare su stdout il seguente output (l'ordine non è vincolante, il contenuto sì):

Numero totale tabelle: 6

Spessore massimo: 2
Spessore minimo: 0
Spessore medio: 0.333333

Livello massimo annidamento: 2

Suggerimenti:

- l'opzione *-i* di *flex* consente un'analisi lessicale case in sensitive
- per determinare il livello di annidamento, è sufficiente contare il numero di tabelle "contenute" in un'altra.