Maîtrise d'informatique Bordeaux I — Année 2002-2003

Analyse Syntaxique et Compilation
Sujet de programmation en Lex et Yacc.

Le sujet

On vous demande d'écrire un programme Lex et un programme Yacc pour analyser un document écrit en HTML. Nous ne considérerons ici qu'un sous-ensemble des règles de la grammaire de HTML.

Le travail demandé

L'analyseur syntaxique généré par Yacc devra pouvoir être appelé avec deux options: `-h` et `-m`.
Un appel avec option `-h` doit calculer la hauteur de l'arbre de dérivation du document.
Un appel avec option `-m` générera un nouveau document dans lequel on aura fait correspondre une marque $</x>$ à toute marque $<x>$ qui n'en avait pas (x étant ici uniquement `li` ou `p`).

Question supplémentaire (à traiter seulement après avoir traité les questions précédentes):
on demande d'ajouter une option `-e` hi qui permet d'extraire (en l'affichant, par exemple) la section "couverte" par le titre indiqué par la première marque hi du document (avec $1 \leq i \leq 6$) .

Le rapport papier devra contenir :

- la définition précise de la syntaxe des fonctions traitées
- des résultats :
  - le source Lex,
  - le source Yacc,
  - des exemples significatifs de fonctions traitées,
  - les éventuels exemples posant problèmes.
- des explications et/ou des commentaires sur :
  - les règles de la grammaire,
  - les fonctions C utilisées,
  - les difficultées rencontrées et les solutions retenues,
  - les messages générés par Yacc concernant les conflits.

Quelques contraintes

Attention *Il vous est bien demandé d'utiliser un minimum de fonctions C, en effet le projet doit bien être considéré comme un projet de Yacc et non pas comme un projet de programmation en C.*

Le travail est à faire en binôme.
Le rapport papier devra être **remis au secrétariat avant le lundi 13 janvier 2002 à 16h.**

Annexe 1: La grammaire

Les symboles auxiliaires ont l'initiale en majuscule (Document est l'axiome); *chaine*, $< x >$ et $< /x >$ désignent les symboles terminaux.

$$
\begin{array}{lcl}
\text{Document} & \longrightarrow & < html > \text{Entete Corps} < /html > \\
\text{Entete} & \longrightarrow & < head > \text{Titre} < /head > \\
\text{Titre} & \longrightarrow & < title > chaine < /title > \\
\text{Corps} & \longrightarrow & < body > \text{Liste-Sections} < /body > \\
\text{Liste-Sections} & \longrightarrow & \text{Entete-De-Section Liste-Sections} \\
 & | & \text{Texte Liste-Sections} \\
 & | & \varepsilon \\
\text{Entete-De-Section} & \longrightarrow & < h1 > chaine < /h1 > \\
 & | & < h2 > chaine < /h2 > \\
 & | & < h3 > chaine < /h3 > \\
 & | & < h4 > chaine < /h4 > \\
 & | & < h5 > chaine < /h5 > \\
 & | & < h6 > chaine < /h6 > \\
\text{Texte} & \longrightarrow & \text{Element Texte} \\
 & | & \varepsilon \\
\text{Element} & \longrightarrow & chaine \\
 & | & < em > \text{Texte} < /em > \\
 & | & < b > \text{Texte} < /b > \\
 & | & < p > \text{Texte} < /p > \\
 & | & < p > \text{Texte} \\
 & | & < ol > \text{Liste-Ord} < /ol > \\
\text{Liste-Ord} & \longrightarrow & \text{Liste-Items Liste-Ord} \\
 & | & \varepsilon \\
\text{Liste-Items} & \longrightarrow & < li > \text{Texte} < /li > \\
 & | & < li > \text{Texte}
\end{array}
$$

Annexe 2: Quelques explications

Un document HTML est composé de texte et de commandes permettant de mettre en forme le texte. Les commandes HTML ont une marque de début et une marque de fin (mais certaines marques de fin sont facultatives). Les marques sont de la forme $< x >$ et $< /x >$, où x est une opportune chaîne de caractères (pouvant être aussi bien majuscules que minuscules, pour ne pas alourdir la grammaire de l'exemple nous n'avons choisi que les minuscules, mais votre programme Lex pourra facilement traiter le cas général).
Par exemple les marques $< body >$ et $< /body >$ entourent le texte qui sera le "corps" du document, les marques $< b >$ et $< /b >$ indiquent que le texte entre les deux doit être affiché en caractères gras.
La marque $< ol >$ indique le début d'une liste numérotée; la marque $< p >$ désigne un saut de paragraphe, la marque $< li >$ désigne un item d'une liste.
En HTML les titres des sections peuvent être visualisés en fonction de leur importance: ils sont représentés

par des fontes de caractères de différente taille. Le symbole Entete-De-Section de la grammaire permet de générer des titres de sections de taille décroissante de $< h1 >$ à $< h6 >$.
Le terminal *chaine* de la grammaire ci-dessus désigne toute chaîne de caractères ne contenant pas de marques.
Un document pourra aussi contenir des commentaires qui seront entourés par les deux séquences $<!$ et $- >$ et ne seront pas imbriqués.

Annexe 3: Exemples

Trois exemples de documents générés par la grammaire ci-dessus:
Exemple 1:

```
<html>
<head>
   <title> Un premier exemple simple </title>
<! 20-12-2002 ->
</head>
<body>

<ol>
    <li> un
    <li> deux
    <li> trois
</ol>
<b> c'est fini! </b>
</body>
</html>
```

Exemple 2:

```
<html>
<head>
   <title> Un deuxième exemple simple </title>
<! 20-12-2002 ->
</head>
<body>
<h1> Ceci est une entête de section! </h1>
<ol>
    <li> un
    <li> deux
    <li> trois
</ol>
<p>
<em> c'est fini! </em>
</body>
</html>
```

Exemple 3:

```
<html>
<head>
   <title> Un deuxième exemple simple </title>
<! 20-12-2002 ->
</head>
<body>
<h1> Ceci est la première entête de section </h1>
<ol>
    <li> un
    <li> deux
    <li> trois
</ol>
<h4> Ceci est une  entête de sous-section </h4>
<li> (en italien)
<ol>
    <li> uno
    <li> due
    <li> tre
</ol>
<h1>  Ceci est une nouvelle entête de section </h1>
<ol>
    <li> premier
    <li>  deuxième
    <li>  troisième
</ol>
<h4> Ceci est encore une  entête de sous-section </h4>
<li> (toujours en italien)
<ol>
    <li> primo
    <li> secondo
    <li> terzo
</ol>

<p>
<em> l'exemple est fini! </em>
<p>
<em> l'esempio é finito! </em>

</body>
</html>
```