Universidad de Buenos Aires
Facultad de
Ciencias Exactas y Naturales
Departamento de Computación

Teoría de Lenguajes

Primer Cuatrimestre de 2013

Trabajo práctico

Micro HTML Prettyprint

Integrante	LU	Correo electrónico
Cammi, Martín	676/02	martincammi@gmail.com
De Sousa Bispo, Mariano	389/08	marian_sabianaa@hotmail.com
Felisatti, Ana	335/10	anafelisatti@gmail.com

1. Introducción

En el presente trabajo práctico definiremos la gramática para generar un lenguaje que interprete una versión simplificada de HTML.

2. Asunciones y aclaraciones

A continuación detallaremos una serie de asunciones que hemos tomado al respecto del trabajo:

- Asumimos que la sección **head** tiene que ir necesariamente antes que la sección **body**.
- Asumimos que los tags $< title >< /title >$ solo puede aparecer una sola vez y el tag $< script >< /script >$ puede aparecer varias veces además de que ambos son opcionales y pueden estar en cualquier orden.
- Asumimos que dentro de la sección **title** puede haber cualquier texto sin tags definidos por el lenguaje.
- Asumimos que dentro de la sección **script** no pueden contener otras subsecciones **script**
- Asumimos que los espacios entre tags son inicialmente filtrados por el analizador léxico y que no llegan a la gramática así como también los tags de comentarios.
- Asumimos en las **regular expressions** que los nombres de los tags son todos en minúsculas y sin ningún espacio.
- En los gráficos de árboles de derivación los ejemplos se basan en código de html para mejor legibilidad, pero en realidad la derivación se realizará en base a los tokens traducidos por el analizador léxico.

3. Gramática

$G = \langle V_N, V_T, P, \text{BEGIN} \rangle$ donde

$V_N = \{$ BEGIN, HTML, HEAD, LEER_HEAD, TITLE, SCRIPTS, MORE_SCRIPTS, BODY, LEER_BODY $\}$

$V_T = \{$ initHtml, endHtml, initHead, endHead, initTitle, endTitle, initScript, endScript, noScripts, initBody, endBody, initDiv, endDiv, initH1, endH1, initP, endP, br, noTags $\}$

y P está dada por:

BEGIN	$\longrightarrow$	initHtml HTML endHtml
HTML	$\longrightarrow$	HEAD BODY
HEAD	$\longrightarrow$	λ | initHead LEER_HEAD endHead
LEER_HEAD	$\longrightarrow$	SCRIPTS TITLE MORE_SCRIPTS | TITLE MORE_SCRIPTS | MORE_SCRIPTS
TITLE	$\longrightarrow$	initTitle noTags endTitle
SCRIPTS	$\longrightarrow$	initScript noScripts endScripts MORE_SCRIPTS
MORE_SCRIPTS	$\longrightarrow$	λ | initBody LEER_BODY endBody
BODY	$\longrightarrow$	λ | initBody LEER_BODY endBody
LEER_BODY	$\longrightarrow$	λ | noTags LEER_BODY | br LEER_BODY | initDiv LEER_BODY endDiv LEER_BODY | initP LEER_BODY endP LEER_BODY | initH1 LEER_BODY endH1 LEER_BODY

4. Expresiones regulares

Terminal	***Expresión regular***
initHtml	<html>
endHtml	</html>
initHead	<head>
endHead	</head>
initBody	<body>
endBody	</body>
initTitle	<title>
endTitle	</title>
initScript	<script>
endScript	</script>
initDiv	
endDiv	
initH1	<h1>
endH1	</h1>
initP	
endP	
br	
noTags	^((?!(TAGS_GRAMATICA)).)*$
noScripts	^((?!(<script> | </script>)).)*$

Donde TAGS_GRAMATICA = <html>|</html>|<head>|</head>|<body>|</body>|
<title>|</title>|<script>|</script>| | | | |<h1>|</h1>|

El tag *noTags* se representa con una expresión regular que identifica cualquier texto sin tags definidos por la gramática.

Por otro lado, la expresión regular para identificar los comentarios es la siguiente:

comment = <!–(.*?)–>

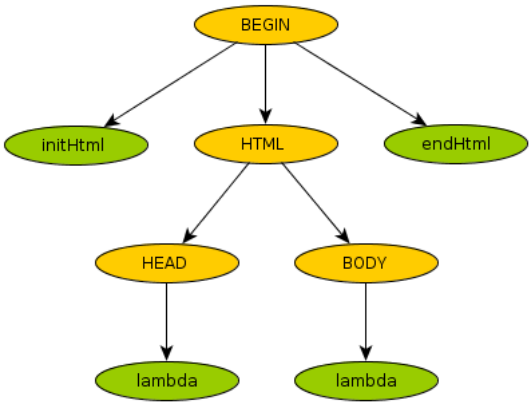
Los diferentes símbolos utilizados en las expresiones denotan:

- | ⟶ Indica disyunción entre dos expresiones.
- ^ ⟶ Matchea la expresión al comienzo de una línea.
- $ ⟶ Matchea la expresión al final de una línea.
- ?! ⟶ No machea la aparición denotada a su derecha.
- * ⟶ Indica una repetición de 0 o más veces.

5. Árboles de derivación

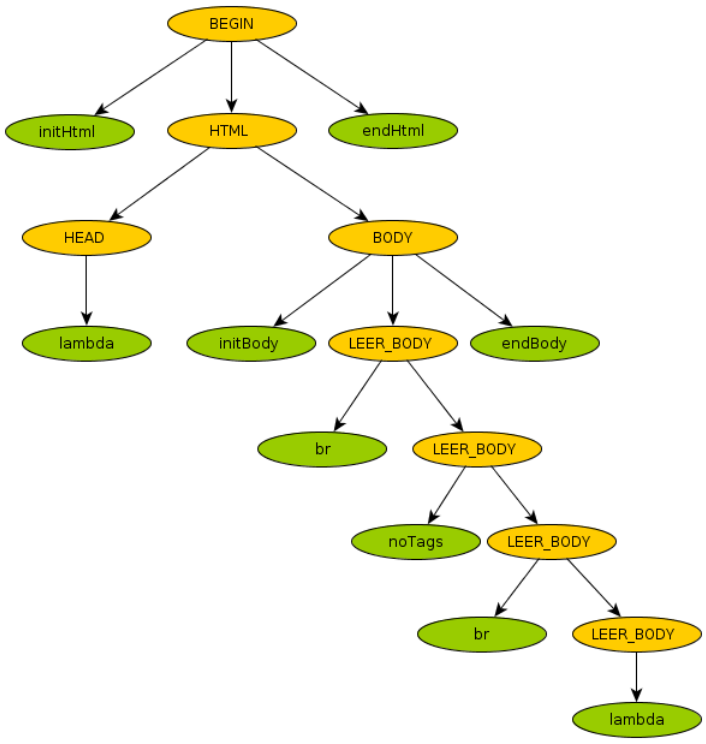
- Ejemplo 1: Html Empty.

```
<html>
</html>
```



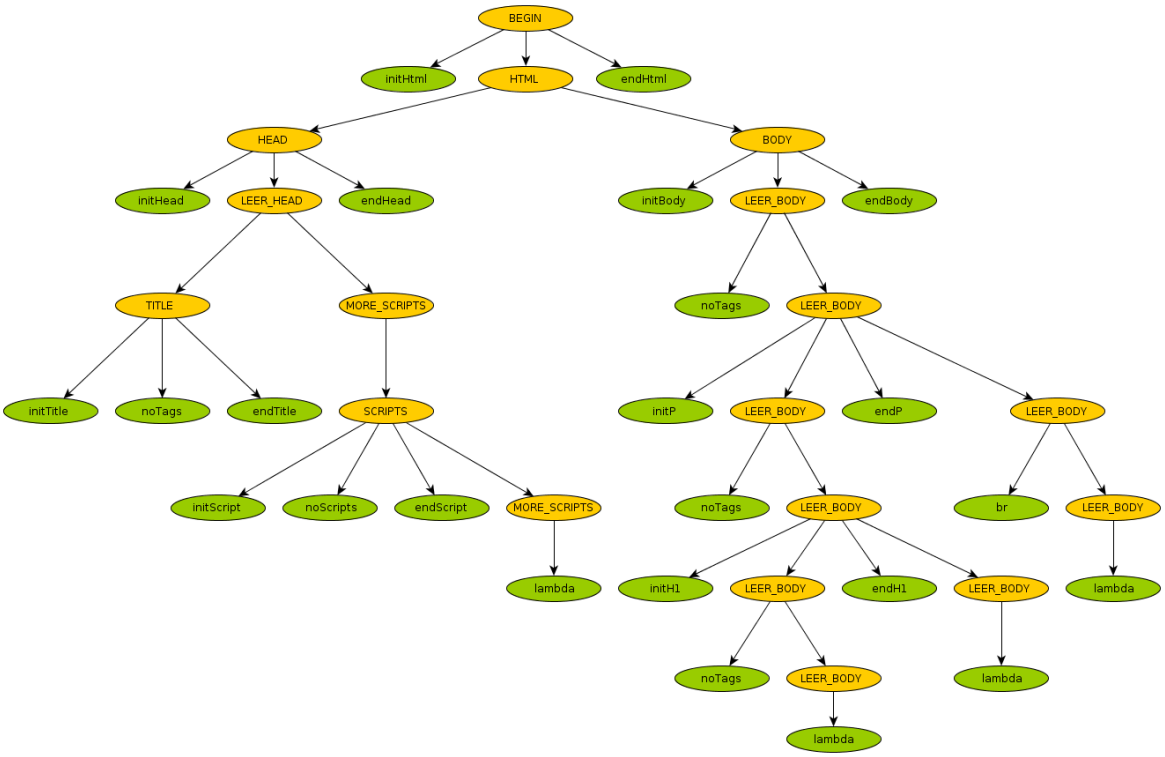
- Ejemplo 2: Empty head body with text.

```
<html>
   <body>
      <br> Body text <br>
   </body>
</html>
```



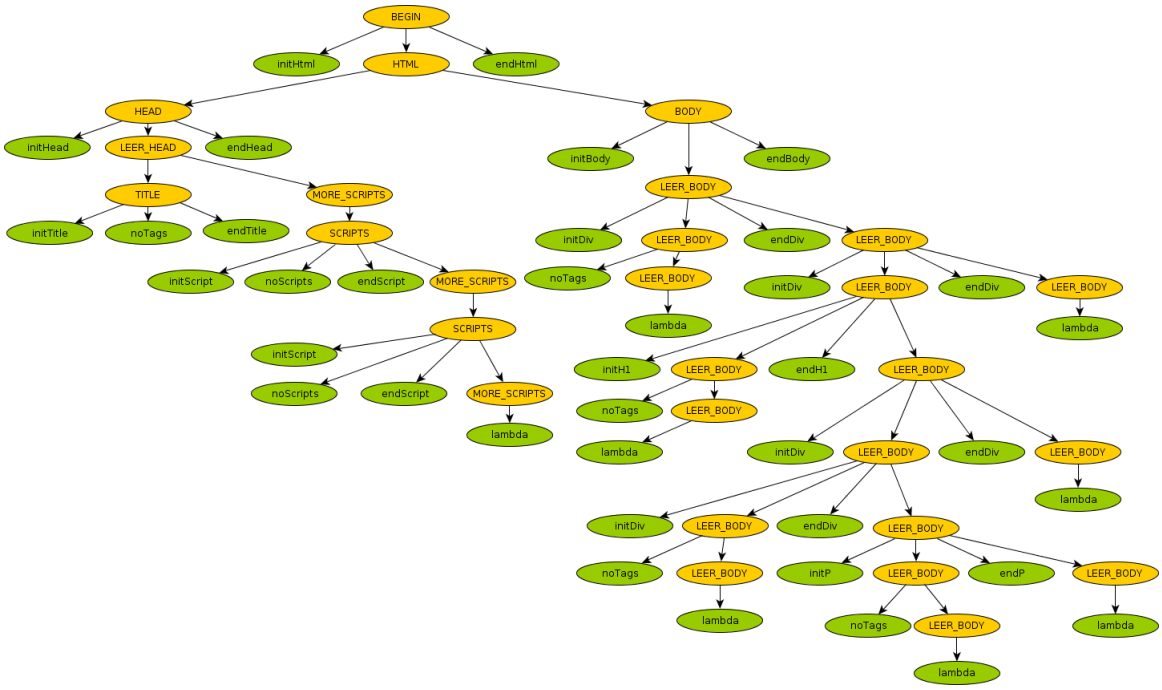
- Ejemplo 3: Tp example.

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



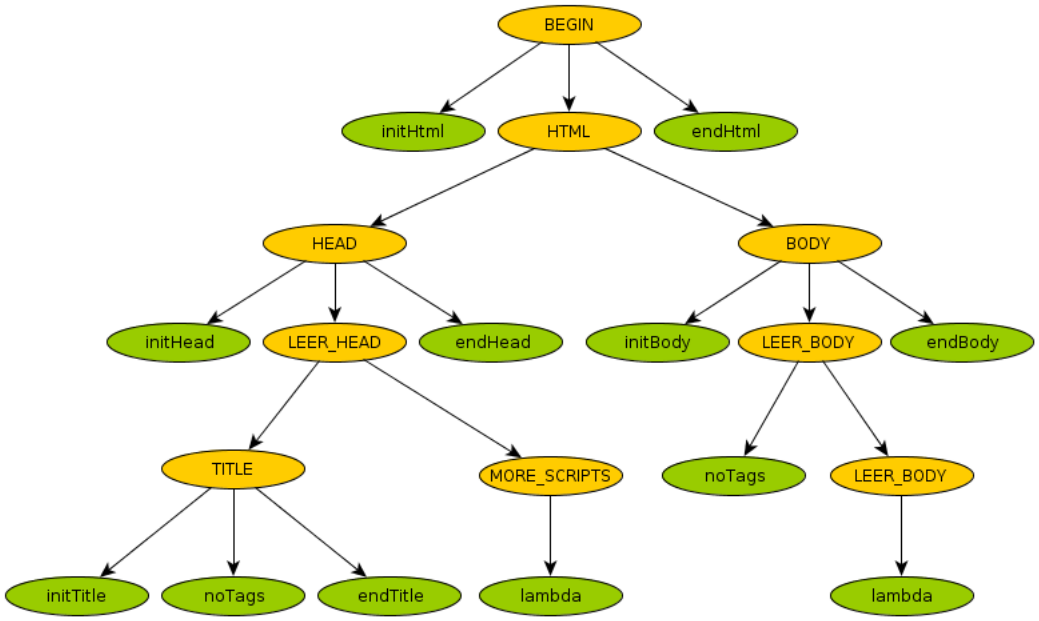
- Ejemplo 4: Huge Body.

```
<html>
   <head>
      <title> Something's Gone Terribly Wrong </title>
      <script> </script>
      <script> bN_cfg = {p: {"dL_ch":"us.hpmguncat",
               "dL_dpt":"error","cobrand":"HuffPost"}};
      </script>
   </head>
   <body>
      <div> </div>
      <div>
         <h1> Oh, Noes! A 404! </h1>
         <div>
            <div> </div>
            <p> or </p>
         </div>
      </div>
   </body>
</html>
```



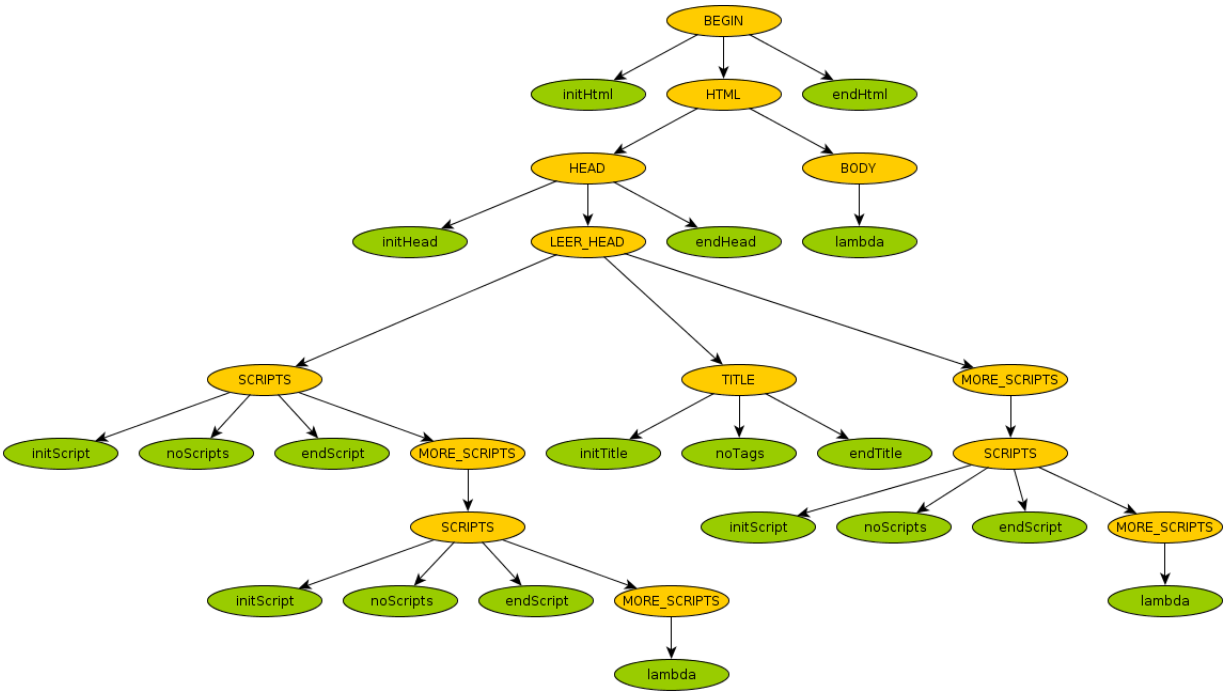
- Ejemplo 5: Just title empty body.

```
<html>
   <head>
      <title> reddit </title>
   </head>
   <body>
      Posts
   </body>
</html>
```



- Ejemplo 6: Several scripts middle title no body.

```
<html>
   <head>
      <script> $(".container").append("<h1> Header </h1>"); </script>
      <script> Tenemos varios scripts </script >
      <title> Esto es un titulo </title>
      <script> print("Title en medio (Y)") </script>
   </head>
</html>
```



- Ejemplo 7: Just scripts nested tags and text.

```
<html>
   <head>
      <script> <div> Solo scripts </div> </script>
      <script> <div> Sin title </div> </script>
   </head>
   <body>
      <br>
      <div>
         <p> Tags anidados </p>
         aca
      </div>
      Y texto suelto
   </body>
</html>
```