



Trabajo Práctico

Generador de analizadores sintácticos

18 de junio, 2009

Teoría de lenguajes

Integrante	LU	Correo electrónico
Rodrigo Campos	561/06	rodrigo@sdfg.com.ar
Martín Fernandez	539/06	bondi007@gmail.com
Matías Pérez	002/05	elmaildematiaz@gmail.com

Reservado para la cátedra

Instancia	Corrector	Nota
Primera entrega		
Segunda entrega		



**DEPARTAMENTO
DE COMPUTACION**

Facultad de Ciencias Exactas y Naturales - UBA

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2160 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (54 11) 4576-3359

<http://exactas.uba.ar/>

Consideraciones generales

El trabajo práctico fué realizado en *Python2.6*. Y el mismo genera código en *C++*.

Análisis del lenguaje de especificación de gramáticas

Tal como planteaba el enunciado de este trabajo práctico, el conjunto de producciones pertenecientes a la gramática que definirá el lenguaje para el cual finalmente se generará el analizador sintáctico pertinente, se encontraba expresado en un lenguaje que especifica producciones de gramática. Para dicho lenguaje se conocía, mediante el enunciado, su conjunto de producciones. Así, por ejemplo la cadena $A:aB / C$; escrita en el lenguaje mencionado, representaba la producción $A \rightarrow aB / C$, para cualquier gramática que la incluyera entre sus producciones.

Dada esta situación, se tornaba fundamental comprender qué producción estaba representando efectivamente cada cadena escrita en el mencionado lenguaje de especificación. Además de esto, era indispensable poder verificar si dada una cadena escrita con los símbolos que define el lenguaje de especificación en cuestión, la misma pertenece o no al lenguaje.

Ante estas necesidades, decidimos utilizar un analizador sintáctico llamado *pyparsing*, para el lenguaje de programación *python*. Este analizador nos ofrecía las funcionalidades requeridas y por ende nos permitía verificar si dada una cadena pertenecía al lenguaje de gramáticas que debemos aceptar la mismo tiempo que nos permite construir el grafo correspondiente a la misma.

Breve reseña acerca de pyparsing

Pyparsing es un analizador sintáctico que, como tal, permite enunciar un conjunto de reglas sobre las cuales se construyen las cadenas válidas de un lenguaje. Estas reglas guardan similitud con las producciones que usualmente se utilizan para describir el comportamiento de una gramática.

Mientras que en el lenguaje natural de gramáticas el símbolo \rightarrow se utiliza para denotar la producción derecha que genera un símbolo no terminal, en pyparsing este símbolo se representa con $=$, con lo cual una producción de la forma $A \rightarrow aB$, en pyparsing se reescribiría como $A = a + B$ (en donde el símbolo '+' representa la concatenación en el lenguaje usual de producciones).

Por otra parte, el operador '|' se representa en pyparsing con el mismo símbolo. Además, el valor λ es representado con la sentencia *Empty()*. La funcionalidad del símbolo '*', por otra parte, es representada por la función *ZeroOrMore(cadena)*, mientras que las acciones del símbolo '+' son llevadas a cabo por la función *OneOrMore()*.

Por ser python un lenguaje de programación, no podemos enunciar la producción $A = a + B$ sin haber definido antes que es 'B'. Esto puede ocasionar un problema si el símbolo inicial de la gramática tiene en su producción un no terminal todavía no definido. Para solucionar esto, alcanza con definir que B es una "producción" a definir luego, esto se hace con la cláusula *Forward*. Y luego se define a B como ya se detalló pero usando $<<$ en vez de $=$.

Otras funciones que resultaron de utilidad a la hora de utilizar pyparsing son:

- `setParseAction(nombreDeFuncion)`, que fue utilizada en el momento de la creación del grafo correspondiente a la gramática que se parsea. Así, por ejemplo, `OneOrMore(valor).setParseAction(esConcat)` invocará a la función `esConcat` cuando ocurrir que la cadena valor aparezca una o mas veces consecutivamente .
- `Literal(caracter)`, esta función sirve para definir caracteres a matchear (Es *case sensitive*).

Descripción del Lenguaje de gramáticas

El lenguaje de gramáticas dado en el enunciado fué:

- $G \rightarrow P G \mid P$

- $P \rightarrow n \text{ '}' PD \text{ '}'$
- $PD \rightarrow n \mid t \mid \lambda \mid PD \text{ '|' } PD \mid PD PD \mid PD \text{ '*' } \mid PD \text{ '+' } \mid PD \text{ '?' } \mid \text{'(' } PD \text{ ')}'$

Donde 'n' son letras mayúsculas y 't' son minúsculas.

Para explicarle a *pyarsing* cómo se compone nuestro lenguaje, primero empezamos por definirle las átomos del mismo. Es decir, empezamos por definirle las cosas más pequeñas que son un elemento a detectar por si mismo. Estos son:

- Los caracteres en minúscula: Estos son nuestros nodos Terminales.
- Los caracteres en mayúscula: Estos son nuestros nodos No-Terminales.
- Los símbolos '?', '*', '+'
- La barra invertida '\ ' que representa λ
- Y por último diferenciamos al símbolo distinguido del resto de los No-Terminales por cuestiones prácticas.

Una vez definidos los átomos de nuestro lenguaje pasamos a definirle la sintaxis:

- 1 $\text{valorMin} \rightarrow \text{noterminal} \mid \text{'(' produccionDer ')}' \mid \text{terminal} \mid \lambda$
- 2 $\text{valor} \rightarrow \text{valorMin? signo+} \mid \text{valorMin}$
- 3 $\text{concat} \rightarrow \text{valor+}$
- 4 $\text{produccionDer} \rightarrow \text{'|'* concat? ('|' concat?)*}$
- 5 $\text{prod} \rightarrow \text{noterminal ':' produccionDer ':'}$
- 6 $\text{produccionInicial} \rightarrow \text{simbDisting ':' produccionDer ':'}$
- 7 $\text{gramaticaInicio} \rightarrow \text{produccionInicial prod*}$

Los items 5 y 6 son casi el mismo, los diferenciamos para poder diferenciar fácilmente la producción inicial del resto de las producciones. Por otro lado el "símbolo distinguido" sería *gramaticaInicio*, ya que por esa se empieza a parsear. Se puede ver entonces que si *produccionDer* se comporta como PD en la gramática del enunciado se genera el mismo lenguaje. Por otro lado también se puede observar que de no haber conflictos en el parseo de *produccionDer* tampoco va a haber problemas en el parseo de la gramática, ya que el parseo de los items 5 a 7 es trivial.

Por otro lado los items 1 y 3 si se asume que *valor* y *produccionDer* no tienen conflictos en el parseo es fácil ver que estos tampoco los tendrán.

Entonces los que pueden llegar a tener conflictos son los items 2 y 4:

- Por un lado el item 2 (*valor*) puede tener problemas porque tiene recursión a izquierda, pero esto para *pyarsing* no genera conflictos porque el '|' de *pyarsing* tiene prioridad a izquierda. Es decir, primero tratará de "matchear" con la primera opción (*valorMin? signo+*) y luego con la segunda (*valorMin*).

También vale aclarar que el '+' seguirá reconociendo los signos hasta que no haya más. En caso contrario se podrían generar conflictos de parseo, por ejemplo, al parsear *a+**. Pero esta cadena se reconocería de la siguiente manera:

Primero se reconoce la 'a' como un *valorMin*
 luego el signo '+'
 y luego el signo '*'.

Por otro lado una concatenación de letras del estilo 'ab' se reconoce ambas veces con la segunda opción de la producción 2, ya que no hay signos.

- Por otro lado el item 4 también genera varias derivaciones posibles para una misma cadena. Pero esto no genera problemas de parseo, ya que el símbolo '*' también se comporta como el '+' (reconociendo siempre que pueda). Esto hace que primero reconozca *todos* los '|' que haya antes de tratar de reconocer las concatenaciones (o concat). Luego si no hay nada más termina y sino trata de reconocer una concatenación, y luego no hay más posibles problemas. Ya que lo que queda es ('|' concat?)* que no trae inconvenientes.

Entonces se puede concluir que problemas de parseo con pyparsing no va a haber (a no ser que se ingrese una cadena inválida).¹.

Lo único que queda por ver es que ambas gramáticas definan el mismo lenguaje. Esto es más complejo y no conocemos un metodo formal para probarlo. Pero se puede ver, como ya dijimos antes, que si `produccionDer` genera lo mismo que PD ambas generan lo mismo.

`produccionDer` intenta dividir en dos casos las producciones de PD; cuando es una lista de disyunciones o cuando es una lista de concatenaciones. Si es una lista de disyunciones se puede ver que cumple, y tambien cumple si es una lista de conjunciones ya que `'|'*` y `('|' concat?)*` son ambos anulables, quedando *concat* que intenta representar justamente eso, una lista de conjunciones.

Entonces *concat* debe representar todas las otras opciones de PD que no sea `PD'|PD`. Es decir: `n | t | λ | PD PD | PD '*' | PD '+' | PD '?' | '(' PD ')'`.

Por otro lado *valor* se encargará de representar `PD '*' | PD '+' | PD '?'`. Mientras que *valorMin* se encargará de representar `n | t | '(' PD ')'` | `λ`.

Para comprender las reglas que se le definen a pyparsing quizás es útil pensar en la siguiente gramática, que si bien no es igual a las reglas definidas, las mismas se basan en esta:

- $G \rightarrow P G \mid P$
- $P \rightarrow n \mid ' PD '$
- $PD \rightarrow PC \mid ' PD \mid PC$
- $PC \rightarrow V PC \mid V$
- $V \rightarrow VM S \mid VM$
- $VM \rightarrow '(' PD ') \mid n \mid t \mid \lambda$
- $S \rightarrow '*' \mid '+' \mid '?'$

Comprobar que la gramática sea ELL(1)

Para asgurarnos que la gramática sea ELL(1), hacemos dos chequeos en dos momentos distintos.

El primer chequeo que realizamos es ver si la gramática tiene simbolos inútiles (es decir, verificamos que todos los no-terminales sean alcanzables y activos). En caso contrario la gramática se rechaza informando el motivo. Luego, por el teorema visto en clase que dice:

- Para toda gramatica G reducida (osea, una en las cuales todos lo no-terminales son alcanzables y activos), si G es independiente de contexto y LL(1), entonces G no es recursiva a izquierda.

Checkeamos si tiene recursión a izquierda, y si es así, concluimos que no es ELL(1).

Es decir, suponemos que la gramática es ELL(1). Sabemos entonces que no tiene recursión a izquierda. Si al checkear vemos que tiene recursión a izquierda, entonces llegamos a un absurdo que vino de suponer que era ELL(1). Es decir que podemos concluir que la gramática no es ELL(1).

Para poder aplicar el teorema anterior es necesario ver que la gramática es independiente de contexto. Para ver esto, podemos ver cómo se reescribiría cada símbolo que agregan las expresions regulares en producciones de una gramática independiente de contexto y ver entonces que cada gramática cuyas producciones son expresiones regulares se puede transformar en una gramática independiente de contexto que genera el mismo lenguaje.

- Por cada expresion $A \rightarrow B \mid C$, se puede transformar en dos producciones: $A \rightarrow B$ y $A \rightarrow C$.
- Por cada expresion $A \rightarrow B^*$, se puede transformar en: $A \rightarrow BA$ y $A \rightarrow \lambda$

¹Para más información a cerca de pyparsing consultar <http://pyparsing.wikispaces.com/HowToUsePyparsing>, <http://onlamp.com/pub/a/python/2006/01/26/pyparsing.html> y <http://pyparsing.wikispaces.com>

- Por cada expresión $A \rightarrow B+$, se puede transformar en: $A \rightarrow BA'$ y $A' \rightarrow B$ y $A' \rightarrow \lambda$
- Por cada expresión $A \rightarrow B?$, se puede transformar en: $A \rightarrow B$ y $A \rightarrow \lambda$

Es fácil ver que ambas expresiones tienen el mismo lenguaje.

Para transformar una expresión regular cualquiera, lo que se hace es por cada ' B^* ' que tenga, se puede introducir un nuevo no-terminal A , tal que $A \rightarrow BA$ y $A \rightarrow \lambda$, y reemplazar ' B^* ' en la expresión regular por A . Haciendo lo mismo para el resto de los símbolos siguiendo las transformaciones explicadas arriba, se puede ver que cualquier expresión regular se puede transformar en una gramática independiente de contexto equivalente.

Se puede notar también, de la forma de transformar las expresiones regulares, que si la expresión regular tenía recursión a izquierda, entonces luego de transformarlas también tendrán.

Es decir, si $A \rightarrow B^*$ tiene recursión a izquierda, entonces es $A \rightarrow A^*$, y las producciones generadas serán $A \rightarrow AA$ y $A \rightarrow \lambda$, es decir que hay recursión a izquierda. Es fácil ver que para el resto de los casos sucede lo mismo.

El segundo chequeo para ver que sea ELL(1), es el que vimos en clase, que lo hacemos luego de calcular anulables, primeros, siguientes y los símbolos directrices. Es decir, chequeamos que para cada '|', los símbolos directrices de sus hijos sean disjuntos y para cada '+', '*', '?', chequeamos que los primeros del nodo hijo sea disjunto con los siguientes del nodo y que ningún hijo sea anulable.

Chequeo de símbolos Inútiles

Para chequear que la gramática no tenga símbolos inútiles primero chequeamos que todos los símbolos sean alcanzables y luego que sean activos.

Para chequear que los símbolos sean inalcanzables simplemente se recorre la componente conexa del nodo del símbolo distinguido marcando los nodos. Y una vez que se terminó de recorrer simplemente se verifica que todos los nodos se hayan marcado, de no ser así se busca un no-terminal y se notifica que éste no cumple produciendo una excepción.

Para chequear que todos los símbolos sean activos, recorremos el grafo partiendo desde el símbolo distinguido. Y por cada nodo, si es un terminal, λ , '*' o '?', lo marcamos como activo (pues siempre producen algo, ese terminal o λ). A cada no-terminal o '|', lo marcamos como activo si alguno de sus hijos es activo. Y al '.' y '+' los marcamos como activo sólo si todos sus hijos son activos. Esto lo hacemos hasta que recorramos el grafo entero sin marcar ningún nodo nuevo como activo.

Una vez que conocemos a todos los símbolos activos lo único que queda por ver es si algún símbolo de la gramática no lo es, y en caso de serlo se produce una excepción notificando el no terminal inactivo.

Finalmente, si pasa estos chequeos, el grafo se corresponde al de una gramática reducida (pues todos sus símbolos no-terminales son activos y alcanzables)

Checkear la recursión a izquierda

Para analizar la recursión a izquierda, usamos una idea muy similar a la que utilizamos para realizar el cálculo de "primeros". Recorremos el grafo de la misma manera, las diferencias son:

- Si el nodo actual es un terminal o λ , lo ignoramos.
- Si es '|', '*', '?', '+' o un no-terminal, hacemos lo mismo que hace primeros y además, para cada hijo si es un no-terminal que no pertenece a `rec_iz` del nodo actual, lo agregamos y marcamos que cambió algo en esta iteración.
- Si es '.', hacemos lo mismo que hace primeros para cada hijo, solo que además también si el hijo es un no-terminal lo agregamos a `rec_iz` del nodo actual.

Como consecuencia de esto, al finalizar tendremos un diccionario, el cual dado un nodo nos dice con qué no-terminales puede comenzar (por esto es que es muy similar a primeros). Si algún nodo puede comenzar con él mismo, entonces tiene recursión a izquierda. Si, por el contrario, ninguno de ellos puede comenzar con si mismo, entonces no se tiene recursión a izquierda presente en la gramática.

Si se encuentra recursión a izquierda se levanta una excepción explicando el porqué de la misma, es decir, donde se encontró la recursión. Vale recordar en este punto que como la gramática fué alterada para quitar producciones inútiles puede ser que la excepción hable de un λ que no existía en la gramática, pero que es consecuencia del procedimiento antes explicado.

Simbolos directrices

Para calcular los simbolos directrices, primero calculamos para cada nodo, los anulables, los primeros y los siguientes.

Luego creamos un diccionario que tenga una entrada por cada nodo del grafo cuyo valor sea los primeros del mismo, y si dicho nodo es anulable, le agregamos los siguientes del nodo.

Para todos estos algoritmos utilizamos la misma forma de recorrer el grafo: recorreremos el grafo hasta que en una recorrida entera del grafo no hayamos agregado nada a lo que estamos calculando (anulables, primeros o siguientes). Si no cambio nada en una recorrida completa del grafo, es porque entonces ya hemos calculado lo que queríamos calcular.

Anulables

En cada paso (o nodo) entonces, lo que hacemos es fijarnos si el nodo actual está en el conjunto de nodos anulables y sino y cumple ciertos requisitos, lo agregamos y marcamos que cambió algo en esta iteración.

Los requisitos dependen del nodo actual, si es un terminal, nunca es anulable, por lo que no lo agregamos. Si es '*', '?' o λ , siempre es anulable por lo que lo agregamos. Si es un no-terminal o un '|', es anulable si alguno de sus hijos es anulable. Y si es un '.' o '+', es anulable si todos sus hijos son anulables.

Al terminar de recorrer el grafo, tenemos el conjunto de nodos anulables.

Primeros

En cada paso, lo que hacemos es fijarnos cierta relacion entre los primeros del nodo actual y sus hijos, si no la cumplen los agregamos y marcamos que cambió algo en esta iteración.

Si el nodo actual es un no terminal y el caracter no pertenece a los primeros del nodo actual, lo agregamos. Si es un λ , lo ignoramos. Si es un '|', '*', '?', '+' o un no-terminal, entonces nos fijamos que contenga a los primeros de todos sus hijos. Si no los contiene, los agregamos y marcamos que algo cambio en esta iteracion. Y si es '.', nos fijamos si contiene a los primeros de su primer hijo (el de más a la izquierda), si no los contiene lo agregamos y marcamos que cambió algo. Si el primer hijo es anulable, nos fijamos que contenga a los primeros del segundo hijo, si no los contiene los agregamos y marcamos que cambió algo en esta iteración. Y así con los sucesivos hijos hasta que no tenga más o encontrar uno no anulable.

Cuando hayamos recorrido el grafo entero sin cambiar nada, entonces sabemos que todas las condiciones se cumplen, por lo que es fácil ver que se ha calculado los primeros de cada nodo correctamente.

Siguientes

En cada paso, lo que hacemos es fijarnos cierta relacion entre los siguientes del nodo actual y sus hijos, si no la cumplen los agregamos y marcamos que cambió algo en esta iteración.

Si el nodo actual es un terminal o λ , lo ignoramos. Si es un '|', '*', '?', '+' o un no-terminal, entonces nos fijamos que los siguientes de todos sus hijos incluyan a los siguientes del nodo actual. Si no es así, los incluimos y marcamos que algo cambió en esta iteración. También, solo si es '*' o '+', nos fijamos que primeros del hijo esté incluido en siguientes del hijo. Si no es así, lo agregamos y marcamos que cambio el grafo en esta iteración. Si es un '.' nos fijamos que los siguientes del '.' sean también siguientes de su último hijo (el de más a la derecha), si no es así los agregamos y marcamos que cambió. Si el último hijo es anulable, entonces nos fijamos que los siguientes de '.' también sean siguientes del nodo anterior al último y lo agregamos si no los incluye. Y lo mismo con el anterior y así sucesivamente hasta llegar a uno que no es anulable. También, para cada dos hijos 'x' e 'y', consecutivos ('x' más a la izquierda que 'y'), nos fijamos que los siguientes de 'x' contengan a los primeros de 'y', si no es así los agregamos. Y si 'y' es anulable, nos fijamos que los siguientes de 'x' contenga a los a los siguientes de 'y'.

Cuando hayamos recorrido el grafo entero sin cambiar nada, entonces sabemos que todas las condiciones se cumplen, por lo que es fácil ver que se ha calculado los siguientes de cada nodo correctamente.

Generación de código

Para generar el código que parsee las cadenas del lenguaje de la gramática que se obtiene como entrada cumpliendo con los requisitos pedidos se decidió elegir el lenguaje C++. Para hacerlo se procede de la siguiente forma.

Por un lado se tienen dos archivos *ParserEll1.cpp* y *Utilitario.cpp* estos contienen un main general y estandar que lo único que hace es leer la cadena de entrada, setear un par de variables y llamar a la función "parsear()" que será generada en el archivo *codigoParser.cpp* que es el que se generará. El archivo *Utilitario.cpp* es el que contiene la lógica del match, que se encarga de ver si la letra pasada para realizar el match es la misma del *TC* (Token Corriente), en caso de serlo aumenta el *TC* y en caso contrario tira el error correspondiente.

Por otro lado se genera el código de parseo específico para cada gramática que debe estar en "*codigoParser.cpp*".

El programa lo que hace es primero obtener los diccionarios *siguientes* y *primeros*, y una vez que se obtienen se imprimen una serie de líneas que corresponden a ciertos include genericos de las librerías que usará. Y luego se imprime para cada nodo No-Terminal una función, que tiene como nombre *Porc_ <Nombre Nodo>()* y para escribir el cuerpo de la misma se recorre el subgrafo que define el nodo llegando hasta las hojas o los nodos no terminales que se encuentren, para cada hijo se escribe²:

- Si es un Terminal se hace un *match(<NombreNodo>);*
- Si es un No Terminal se hace un *Porc_ <Nombre Nodo>()*
- Si es un "[" se hace una serie de *if* para cada uno de sus hijos preguntando si el *TC* está en los simbolos directrices de ese nodo, y dentro de cada *if* se coloca el cuerpo de los hijos y se sigue. Y se termina con un *else* para el caso de error en el que se encuentra un caracter inesperado.
- Si es un "." se escribe directamente el código de todos sus hijos en orden.
- Si es un "*" se escribe un *while* preguntando si en *TC* se encuentra en los Primeros de ese nodo.
- Si es un "+" se actua de manera análoga al "*" pero con un *do while* en vez del *while*.

De esta manera se escribe todo el código correspondiente a esta gramática.

Notas de utilización del trabajo

Para generar el programa que analiza una gramática dada se debe:

- Primero se debe generar el código del programa, para lo cual hay que ejecutar: *python parser.py > "codigoParser.cpp"* (recordar que se necesita *python2.6* o compatible).
- Luego hay que compilar el código, simplemente se debe compilar *ParserEll1.cpp* con los archivos *Utilitario.cpp* y *codigoParser.cpp* en el mismo directorio. Esto se puede realizar ejecutando *g++ ParserEll1.cpp -o parser*

²Si el No-Terminal tiene más de una producción se lo toma como si fuera un "[".

- Una vez obtenido el programa, este recibe la cadena a reconocer por consola. Por lo que se debe ejecutar `./parser "cadena"` para que reconozca la cadena *cadena*.

Testing

El trabajo fué testeado con 11 gramáticas, las cuales algunas no se aceptan por no ser ELL1 o tener símbolos inútiles y en los otros casos se prueban para cada una un set de cadenas válidas e inválidas.

Este set de test se encuentra en la carpeta *gramaticas*. Y para ejecutarlo hay un script llamado *testear.sh*, con lo que ejecutando `./testear.sh` se corren todos los casos de test.