Diseño e Implementación de un Analizador Léxico y Analizador Semántico para el Lenguaje AVISMO

Aramis E. Matos

Lenier Gerena

Angel Berrios Pellot

Segundo Semestre, 2022-2023

Tabla de Contenido

1	Intr	oducción	2						
2	Analizador Léxico								
	2.1	Gramatica del Lenguaje AVISMO	4						
	2.2	Diseño del del Analizador Léxico	12						
		2.2.1 Autómatas Finitos Deterministas	12						
		2.2.2 Tabla de Símbolos	13						
	2.3	Implementación del Analizador Léxico	14						
3	Imp	lementación del Analizador Sintáctico	19						
4	Conclusiones y Recomendaciones								
Re	Referencias Bibliográficas 23								

Introducción

La visualización molecular puede ser considerada como una de las áreas mas importante dentro de la bioinformática. Entre sus aplicaciones mas relevantes se destacan el diseño de nuevo fármacos... (Narciso Farias, Rios, Hidrobo, & Vicuña, 2012). Este enunciado fue escrito hace mas de una década. Sin embargo, hoy día en un mundo pospandemia, reconocemos que tan sabio fue. El desarollo de la vacuna contra el COVID-19 tan rápido fue gracias a herramientas de visualización como el Ambiente de visualización Molecular (AVISMO) (Narciso Farias et al., 2012) El propósito de este proyecto es definir el automata de estado finito del lenguaje AVISMO, los patrones el cual caracterizan los lexemas del lenguaje, los atributos de los lexemas y la implementación del analizador léxico y sintáctico del lenguaje AVISO en Python. La implementación léxica y sintáctica fue desarrollada utilizando Python Lex Yacc (PLY (Python Lex-Yacc) — Ply 4.0 Documentation, n.d.) con el lenguaje MAPL (PL-Project-LGM-YVV-AMN/MAPL, n.d.)

como base.

Analizador Léxico

2.1 Gramatica del Lenguaje AVISMO

- <SENTENCIAS> ::= <FIN_DE_LINEA> <SENTENCIAS> | <SENTEN-CIA> <FIN_DE_LINEA>
- <FIN_DE_LINEA> ::= ":" | ";"
- <SENTENCIA> ::= "defina" <ID> "como" <TIPO> | <ID> "=" <MOD-ELO_MOLECULAR> | <OPERACION> "(" <ID> ")"
- <ID> ::= "A" | "B" | "C" | "D" | "E" | "F" | "G" | "H" | "I" | "J" | "K" | "L" |

 "M" | "N" | "O" | "P" | "Q" | "R" | "S" | "T" | "U" | "V" | "W" | "X" | "Y" |

 "Z" | "a" | "b" | "c" | "d" | "e" | "f" | "g" | "h" | "i" | "j" | "k" | "l" | "m" | "n"

 | "o" | "p" | "q" | "r" | "s" | "t" | "u" | "v" | "w" | "x" | "y" | "z" | <LETRA>

 <IDCONT>

- <IDCONT> ::= "A" | "B" | "C" | "D" | "E" | "F" | "G" | "H" | "I" | "J" | "K" | "L" | "M" | "N" | "O" | "P" | "Q" | "R" | "S" | "T" | "U" | "V" | "W" | "X" | "Y" | "Z" | "a" | "b" | "c" | "d" | "e" | "f" | "g" | "h" | "i" | "j" | "k" | "1" | "m" | "n" | "o" | "p" | "q" | "r" | "s" | "t" | "u" | "v" | "w" | "x" | "y" | "z" | <LETRA> <IDCONT> | "0" | "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9" | <DIGITO> <IDCONT>
- <LETRA> ::= "A" | "B" | "C" | "D" | "E" | "F" | "G" | "H" | "I" | "J" | "K" | "L" | "M" | "N" | "O" | "P" | "Q" | "R" | "S" | "T" | "U" | "V" | "W" | "X" | "Y" | "Z" | "a" | "b" | "c" | "d" | "e" | "f" | "g" | "h" | "i" | "j" | "k" | "l" | "m" | "n" | "o" | "p" | "q" | "r" | "s" | "t" | "u" | "v" | "w" | "x" | "y" | "z"
- <DIGITO> ::= "0" | "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9"
- <TIPO> ::= "modelo"
- <OPERACION> ::= "graficar2d" | "graficar3d" | "pesomolecular"
- <MODELO_MOLECULAR> ::= "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" | "Be" | "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" | "V" | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh" | "Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" | "Cu" | "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" | "C" | "Si" | "Ge" | "Sn" | "Pb" | "N" | "P" | "As" | "Sb" | "Bi" | "O" | "S" | "Se" | "Te" | "Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" | "Xe" | "Rn" | <ELEMENTO_QUIMICO> <VALENCIA> | <ELE-

MENTO> <GRUPO_FUNCIONAL> | <COMPUESTO> <ELEMENTO> <GRUPO_FUNCIONAL> | <COMPUESTO> <COMPUESTO>

- <COMPUESTO> ::= "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" | "Be" |

 "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" | "V"

 | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh"

 | "Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" |

 "Cu" | "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" |

 "C" | "Si" | "Ge" | "Sn" | "Pb" | "N" | "P" | "As" | "Sb" | "Bi" | "O" | "S" |

 "Se" | "Te" | "Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" |

 "Xe" | "Rn" | <ELEMENTO | QUIMICO > <VALENCIA > | <ELEMENTO >

 <GRUPO_FUNCIONAL > | <ELEMENTO > <GRUPO_FUNCIONAL > <EN
 LACE > | <ELEMENTO > <ENLACE >
- <COMPUESTOS> ::= <COMPUESTO> <COMPUESTO> | <COMPUESTOS>
- <ELEMENTO> ::= "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" | "Be" | "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" | "V" | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh" | "Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" | "Cu" | "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" | "C" | "Si" | "Ge" | "Sn" | "Pb" | "N" | "P" | "As" | "Sb" | "Bi" | "O" | "S" | "Se" | "Te" | "Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" | "Xe" | "Rn" | <ELEMENTO_QUIMICO> <VALENCIA>
- <ELEMENTO_QUIMICO> ::= "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" |

"Be" | "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" |

"V" | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh" |

"Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" | "Cu"

| "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" | "C" | "Si"

| "Ge" | "Sn" | "Pb" | "N" | "P" | "As" | "Sb" | "Bi" | "O" | "S" | "Se" | "Te" |

"Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" | "Xe" | "Rn"

- <VALENCIA> ::= "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9"
- <GRUPO_FUNCIONAL> ::= <GRUPO_FUNCIONAL_INFERIOR>
 <GRUPO_FUNCIONAL_SUPERIOR> | <GRUPO_FUNCIONAL_SUPERIOR>
 <GRUPO_FUNCIONAL_INFERIOR> | "(" <MODELO_GRUPO_FUNCIONAL>
 ")" | "[" <MODELO_GRUPO_FUNCIONAL> "]"
- <GRUPO_FUNCIONAL_SUPERIOR> ::= "[" < MODELO_GRUPO_FUNCIONAL> "]"
- <GRUPO_FUNCIONAL_INFERIOR> ::= "(" < MODELO_GRUYPO_FUNCIONAL> ")
- <MODELO_GRUYPO_FUNCIONAL> ::= <ENLACE> <MODELO_MOLECULAR> | "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" | "Be" | "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" | "V" | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh" | "Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" | "Cu" | "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" | "C" | "Si" | "Ge" | "Sn" | "Pb" | "N" | "P"

| "As" | "Sb" | "Bi" | "O" | "S" | "Se" | "Te" | "Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" | "Xe" | "Rn" | <ELEMENTO_QUIMICO> <VA-LENCIA> | <ELEMENTO> <GRUPO_FUNCIONAL> | <COMPUESTO> <ELEMENTO> | <COMPUESTO> <COMPUESTO> <COMPUESTO> <

En la tabla 2.1, en la columna de patrones, note que cuando dice {TOKEN} donde TOKEN se refiere a el patrón asociado a token. Por ejemplo, si un patrón dice {ELEMENTO_QUIMICO}, esto significa que inserta el patrón asociado al token ELEMENTO_QUIMICO. Esto no significa que el analizador léxico espera un token de por si, sencillamente se hizo con el propósito de evitar redundancias.

<fin_de_linea></fin_de_linea>	;1:	:	Simbolo reservado
<palabra< td=""><td>defina I como</td><td>defina</td><td>Palabra reservada</td></palabra<>	defina I como	defina	Palabra reservada
_RESERVADA>			
<id></id>	[A-Za-z][A-Za-z0-9]*	var1	Modelo molecular aso-
			ciado
<idcont></idcont>	[A-Za-z0-9]+	1ar	ID asociado
<letra></letra>	[A-Za-z]	a	ID asociado
<digito></digito>	[0-9]	7	Valor numérico, lexema
			asociado
<tipo></tipo>	modelo	modelo	ID asociado
<operacion></operacion>	graficar2d graficar3d pesomolecular	pesomolecular	ID asociado
<modelo< td=""><td>({ELEMENTO _QUIMICO} {ELEMENTO _QUIMICO} {VALEN-</td><td>СН3(СН3)СНН</td><td>ID asociado</td></modelo<>	({ELEMENTO _QUIMICO} {ELEMENTO _QUIMICO} {VALEN-	СН3(СН3)СНН	ID asociado
_MOLECULAR>	CIA} {ELEMENTO} {GRUPO _FUNCIONAL} {ELEMENTO}		
	{GRUPO_FUNCIONAL} {ENLACE} {ELEMENTO} {ENLACE})		

Atributos

Lexema

Token

Patrón

<compuesto></compuesto>	COMPUESTO ({ELEMENTO _QUIMICO} {ELEMENTO _QUIM-	CH3::	Modelo molecular aso-
	ICO} {VALENCIA} {ELEMENTO} {GRUPO_FUNCIONAL} {ELE-		ciado, enlaces, valencias
	MENTO} {GRUPO_FUNCIONAL} {ENLACE} {ELEMENTO} {EN-		
	LACE})		
<compuestos></compuestos>	{COMPUESTO}+	CH3::(OH)3	Modelo molecular aso-
			ciado, enlaces, valencias
<elemento></elemento>	{ELEMENTO _QUIMICO} {VALENCIA}?	Ag3	Elemento, valencia
<elemento< td=""><td>("H" "Li" "Na" "K" "Rb" "Cs" "Fr" "Be" "Mg" "Ca" "Sr" </td><td>I</td><td>Elemento</td></elemento<>	("H" "Li" "Na" "K" "Rb" "Cs" "Fr" "Be" "Mg" "Ca" "Sr"	I	Elemento
_QUIMICO>	"Ba" "Ra" "Sc" "Y" "Ti" "Zr" "Hf" "Db" "V" "Nb" "Ta" "Ji"		
	"Cr" "Mo" "W" "Rf" "Mn" "Tc" "Re" "Bh" "Fe" "Ru" "Os"		
	"Hn" "Co" "Rh" "Ir" "Mt" "Ni" "Pd" "Pt" "Cu" "Ag" "Au"		
	"Zn" "Cd" "Hg" "B" "Al" "Ga" "In" "Ti" "C" "Si" "Ge" "Sn"		
	"Pb" "N" "P" "As" "Sb" "Bi" "O" "S" "Se" "Te" "Po" "F"		
	"Cr" "Br" "I" "At" "He" "Ne" "Ar" "Kr" "Xe" "Rn")		
<valencia></valencia>	[1-9]	2	Valor

<grupo _fun-<="" th=""><th>({GRUPO _FUNCIONAL _INFERIOR} {GRUPO _FUNCIONAL</th><th>(CH3){Ag2}</th><th>Grupos funcionales,</th></grupo>	({GRUPO _FUNCIONAL _INFERIOR} {GRUPO _FUNCIONAL	(CH3){Ag2}	Grupos funcionales,
CIONAL>	_SUPERIOR} {GRUPO _FUNCIONAL _SUPERIOR} {GRUPO		grupo funcional infe-
	_FUNCIONAL_INFERIOR} "(" {MODELO _GRUPO _FUN-		rior, grupo funcional
	CIONAL} ")" "[" MODELO _GRUPO _FUNCIONAL "]")		superior
<grupo _fun-<="" td=""><td>"[" {MODELO _GRUPO _FUNCIONAL} "]"</td><td>[CVHe3]</td><td>Elementos, valencias</td></grupo>	"[" {MODELO _GRUPO _FUNCIONAL} "]"	[CVHe3]	Elementos, valencias
CIONAL _INFE-			
RIOR>			
<grupo _fun-<="" td=""><td>"(" {MODELO _GRUPO _FUNCIONAL} ")"</td><td>(CVHe3)</td><td>Elementos, valencias</td></grupo>	"(" {MODELO _GRUPO _FUNCIONAL} ")"	(CVHe3)	Elementos, valencias
CIONAL _SUPE-			
RIOR>			
<modelo< td=""><td>({ELEMENTO _QUIMICO}+ {VALENCIA}?)+ ({ELEMENTO}+</td><td>FeH=C3Si4</td><td>Elementos, enlaces, va-</td></modelo<>	({ELEMENTO _QUIMICO}+ {VALENCIA}?)+ ({ELEMENTO}+	FeH=C3Si4	Elementos, enlaces, va-
_GRUPO _FUN-	{ENLACE} {ELEMENTO}+)+		lencias
CIONAL>			
<enlace></enlace>	("-" "=" ":" "::")	-	Valencia

Tabla 2.1: Tabla de Componentes Léxicos de AVISMO

2.2 Diseño del del Analizador Léxico

2.2.1 Autómatas Finitos Deterministas

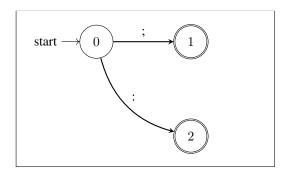


Figura 2.1: Automata del patrón para el token <FIN_DE_LINEA>

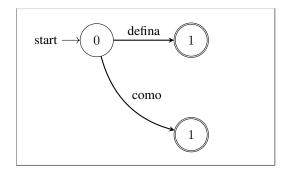


Figura 2.2: Automata del patrón para el token <PALABRAS_RESERVADA>

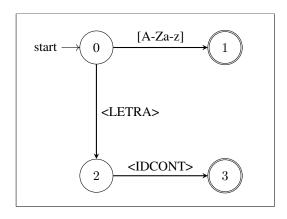


Figura 2.3: Automata del patrón para el token <ID>

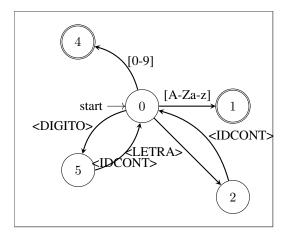


Figura 2.4: Automata del patrón para el token <IDCONT>





Figura 2.5: Automata del patrón para el token <ASIGNACION>

Figura 2.6: Automata del patrón para el token <LETRA>



Figura 2.7: Automata del patrón para el token <DIGITO>

2.2.2 Tabla de Símbolos

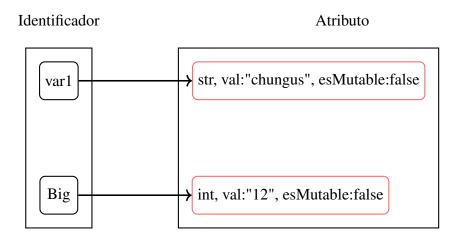


Figura 2.8: Tabla de símbolos implementada como un diccionario

2.3 Implementación del Analizador Léxico

Como se ha mencionado anteriormente, la implementación lexica del proyecto fue inspirada por la MAPL (*PL-Project-LGM-YVV-AMN/MAPL*, n.d.) y adaptada para la gramática de AVISMO.

Para ejecutar el analizador léxico, primero se tiene que instalar *Python 3*. Luego, se ejecuta pip install ply por la línea de comando. Finalmente, se ejecuta python lexer.py test_prog.txt

Al ejecutarse el comando anterior, el programa procede a leer **cada caracter** del programa e identificar si una serie de caracteres sigue un patrón que forma parte del lenguaje AVISMO. Al encontrar un patrón reconocido, tales como un identificador o modelo molecular, lo clasifica con un *token* correspondiente, lo emprime en el archivo de *output.txt* y lo devuelve al analizador sintáctico. Note, que el patrón de identificador reconoce palabra reservadas también. Esto crea ambiguedad semántica debido a que la gramatica no tiene un mecanismo para diferenciar entre una palabra reservada y un identificador. Por esta razón, si una serie de caracteres se identifica como un lexema de categoría identificador, se compara con los valores ya existentes del diccionario *variables*. Al inicializar el programa lexer.py, este se encarga de abrir el archivo de palabras reservadas (*keywords.txt*) y añadir las palabras reservadas antes que cualquier variable se pueda inicializar. Mas aún, a las palabras reservadas se les asigna el valor de la cadena vacía. Esto se hace con el propósito de poder diferenciar entre palabras reservadas e identificadores, ya que al nivel sintáctico, no es posible asignarle a

una identificador una cadena vacía, como se puede apreciar a continuación:

```
def t_ID(t):  #identifica el token ID

    pero tambien idetifica el token de palabra

    reservada

    r"[A-Za-z]+\d*"

    isPR = reserved.get(t.value,"ID")

    if isPR != "ID":

        t.type = "PALABRA_RESERVADA"

    else:

        t.type = isPR

        variables[t.value] = ""

    return t
```

Figura 2.9: Patrón que cezga identificadores de palabras reservadas

Con el propósito de visualizar los lexemas generados por lexer.py, colocado en el archivo *lexer.py*, se utiliza la siguiente función:

```
with open(test_file, "r") as f:
      with open("output.txt", "w") as o:
           for data in f:
               lexer.input(data)
               for tok in lexer:
                   tokenTable.add_row([tokenNum,tok.type]
                      ,tok.value,tok.lineno,tok.lexpos,

    test_file])

                   tokenNum += 1
          o.write(str(tokenTable))
          line = "\n\nTABLA DE SIMBOLOS"
          o.write(line+"\n")
10
          for val in variables:
               symbofigable.add_row([val])
          o.write(str(symbofigable)+"\n")
          line = "\n\nPALABRAS RESERVADAS"
          o.write(line+"\n")
          for val in reserved:
               reservedWords.add_row([val])
17
          o.write(str(reservedWords)+"\n")
```

Figura 2.10: Código para imprimir los tokens encontrados

Figura 2.11: Gestión de errores

Esto imprime el *token* resultante en el archivo *output.txt*. En el caso de un error léxico, se ejecuta el siguiente código: Cuando se encuentra un error léxico, no se retorna al analizador sintáctico. Los patrones que se utilizan están en lexer.py (líneas 58-103) son una adaptación de la gramática en la tabla 2.1. Note la tupla *tokens*. Esta contiene todos los *tokens* del lenguaje AVISMO. Sin embargo, no todos están definidos a el nivel léxico. La mayoría de los *tokens* se construyen en la etapa sintáctica, particularmente los que tienen que ver con compuestos quimicos y modelos moleculares. Mas aún, note el objeto *tok* en la línea 5 de el listado 2.10. Este contiene los atributos *type*, *value*, *lineno* y *lexpos*. Estos devuelven, respectivamente, el tipo, lexema, en que línea del archivo se encuentra y la posición del primer caracter de el *token* llamado tok. Otro detalle importante de la implementación léxica es el orden de aplicación de los patrones. Los patrones en forma de variables, como en el listado 2.12, tienen que estar escrito antes que los patrones escritos en forma de funciones, como en el listado 2.9 y 2.13

```
t_ENLACE = r"(-|=|:|::)" #define los tokens para

→ diferentes tipos de enlaces quimicos
```

Figura 2.12: Patrón en forma de variable

Figura 2.13: Patrón en forma de función

Implementación del Analizador Sintáctico

Conclusiones y Recomendaciones

Lista de Figuras

2.1	Automata del patrón para el token <fin_de_linea></fin_de_linea>	12
2.2	Automata del patrón para el token <palabras_reservada></palabras_reservada>	12
2.3	Automata del patrón para el token <id></id>	12
2.4	Automata del patrón para el token <idcont></idcont>	12
2.5	Automata del patrón para el token <asignacion></asignacion>	13
2.6	Automata del patrón para el token <letra></letra>	13
2.7	Automata del patrón para el token <digito></digito>	13
2.8	Tabla de símbolos implementada como un diccionario	13
2.9	Patrón que cezga identificadores de palabras reservadas	15
2.10	Código para imprimir los <i>tokens</i> encontrados	16
2.11	Gestión de errores	17
2.12	Patrón en forma de variable	17
2.13	Patrón en forma de función	18

Lista de Tablas

2 1	Table de Com		T áminan	J. AVICMO						1.1
Z. I	Tabla de Com	ponentes :	Lexicos	de AVISMO						11

Referencias Bibliográficas

- Narciso Farias, F., Rios, A., Hidrobo, F., & Vicuña, O. (2012, May). Una gramática libre de contexto para el lenguaje del ambiente de visualización molecular AVISMO..
- *PL-Project-LGM-YVV-AMN/MAPL.* (n.d.). https://github.com/PL-Project-LGM-YVV-AMN/MAPL.
- PLY (Python Lex-Yacc) ply 4.0 documentation. (n.d.). https://ply.readthedocs.io/en/latest/index.html.