

به نام خداوند بخشنده مهربان



فاز ۱ پروژه - درس اصول و طراحی کامپایلرها

دکتر رزازی

ترم پاییز ۱۴۰۰-۱۳۹۹ - دانشکده کامپیوتر، دانشگاه صنعتی امیرکبیر

لطفا قبل از شروع به حل کردن تمرین به نکات زیر توجه فرمایید:

۱. هدف از انجام تمرین‌ها، یادگیری عمیق‌تر مطالب درسی است. در نتیجه هرگونه کپی‌برداری موجب کسر نمره خواهد شد.
۲. تمام فایل‌های خواسته شده را در یک فایل فشرده قرار دهید. نام فایل نهایی را شماره دانشجویی خود قرار دهید. (برای مثال dlexer_Family_9631747.zip)
۳. در کنار این فایل ۳ ویدئو برای راهنمایی شما وجود دارد که توصیه می‌کنیم آن‌ها را مشاهده کنید.
۴. شما موظفید که ۳ فایل txt را به برنامه خود ورودی دهید و خروجی آن را در فایل‌ها جداگانه به همراه کد های برنامه در فایل فشرده قرار بدهید.
۵. در صورت وجود هرگونه سوال می‌توانید از طریق ایمیل با تدریسار در ارتباط باشید.

moh.robati@aut.ac.ir

sheykh@aut.ac.ir

rouzbehghasemi1998@gmail.com

sepehr.asgarian@gmail.com

amirali.sajjadi98@gmail.com

parsafarinnia@gmail.com

ما از شما در این بخش پروژه می خواهیم که یک تحلیل گر لغوی^۱ طراحی کنید که بتواند لغات^۲ را در برنامه تشخیص داده و نشانه^۳ مربوط به آن را برگرداند. همچنین تحلیل گر شما باید بتواند تشخیص بدهد که آیا لغت درست است یا خیر (یعنی آیا نشانه ای معتبر برای آن وجود دارد یا خیر). و در صورت غلط بودن بخش غلط را اعلام کند و به تحلیل ادامه بدهد.

Lexeme	Token Value
Identifier	ID
Integer number	INTEGERNUMBER
Float number	FLOATNUMBER
"int"	INTEGER
"float"	FLOAT
"bool"	BOOLEAN
"void"	VOID
"fun"	FUNCTION
"True"	TRUE
"False"	FALSE
"print"	PRINT
"return"	RETURN
"main"	MAIN
"if"	IF
"else"	ELSE
"elseif"	ELSEIF
"while"	WHILE
"on"	ON
"where"	WHERE
"for"	FOR
"and"	AND
"or"	OR
"not"	NOT
"in"	IN
"="	ASSIGN
"+"	SUM
"-"	SUB
"*"	MUL

¹Lexical analyzer

² Lexemes

³ Token

"/"	DIV
"%"	MOD
">"	GT
">="	GE
"<"	LT
"<="	LE
"=="	EQ
"!="	NE
"{"	LCB
"}"	RCB
"("	LRB
")"	RRB
"["	LSB
"]"	RSB
","	SEMICOLON
" "	COMMA
Error	ERROR!

Identifier:

شناسه^۴ می تواند شامل حروف بزرگ و کوچک لاتین، ارقام عددی و _ باشد.

شناسه نباید با عدد یا حروف بزرگ لاتین شروع بشود به طور مثال "3id" و یا "AMIR" غلط است و باید به عنوان نشانه "ERROR" برگردانده شود.

همچنین شناسه نمی تواند یکی از کلمات ذخیره^۵ شده مانند "main", "print", "for" و ... باشد.

Integer number :

عدد طبیعی شامل ارقام بین ۰ تا ۹ است و همچنین تعداد ارقام عدد طبیعی باید کمتر از ۱۰ باشد. توابعی که برای برگرداندن توکن های اعداد صحیح می نویسید نباید صفر های ابتدای^۶ را برگردانند، یعنی برای عدد ۰۰۲ مقدار ۲ برگردانده می شود.

Float number :

عدد حقیقی به صورت دو عدد طبیعی با یک "." در وسط است و همچنین تعداد ارقام بخش طبیعی عدد باید کمتر از ۱۰ باشد.

⁴ Identifier

⁵ Reserved

⁶ Leading Zero

همچنین توابعی که برای برگرداندن توکن های اعداد اعشاری می نویسید نباید صفر های ابتدای بخش طبیعی و صفر های انتهایی بخش اعشاری^۷ را برگردانند، یعنی برای عدد ۰۰۲۳۰۱,۲۰۳۰۰ مقدار ۲۳۰۱,۲۰۳ برگردانده می شود.

توجه داشته باشید که اعداد منفی در این مدل تحلیل نداریم.

همچنین در لکسر باید تضمین کنید که ۲ عملگر^۸ اصلی (/ - * +) نمی توانند در کنار هم قرار بگیرند مگر کاراکتری غیر از whitespace میان آن ها باشد و در این صورت کل عملگر ها باید در کنار هم در یک error باشند.

برنامه شما باید در مقابل ورودی های مقابل خروجی مورد نظر را بدهد.

خروجی	ورودی
ERROR SUM ID ASSIGN INTEGERNUMBER	19a + b = 9
WHILE LRB RRB	while ()
ID ERROR ID	_x-+-x_
ERROR SUB FLOATNUMBER	9.7.9 – 0.0
ERROR FOR RSB SUB INTEGERNUMBER LSB	For for]-90[
AND OR SUB EQ	and or==
ID ERROR SEMICOLON	c++ ;
ERROR SUM ERROR	ERROR + 1231231239.509
INTEGERNUMBER NE NOT TRUE	5!=not True
INTEGERNUMBER ERROR INTEGERNUMBER ID	9 - * - 8 _
VOID MUL LE INTEGERNUMBER	void * <= 0

⁷ Trailing Zero

⁸ operator