Laboratorio di Linguaggi Formali e Traduttori
Corso di Studi in Informatica
A.A. 2015/2016

Ferruccio Damiani e Jeremy Sproston
Dipartimento di Informatica — Università degli Studi di Torino

Versione del 25 novembre 2015

Sommario

Questo documento descrive le esercitazioni di laboratorio e le modalità d'esame del corso di *Linguaggi Formali e Traduttori* per l'A.A. 2015/2016.

Svolgimento e valutazione del progetto di laboratorio

È consigliato sostenere l'esame nella prima sessione d'esame dopo il corso.

Forum di discussione e supporto on-line al corso

Sulla piattaforma I-learn sono disponibili forum di discussione dedicati per gli argomenti affrontati durante il corso e per scambiare opinioni tra i vari gruppi di lavoro e con il docente. L'iscrizione al forum principale è effettuata automaticamente, è possibile disiscriversi ma è consigliabile farlo solo a seguito del superamento dell'esame per poter sempre ricevere in modo tempestivo le comunicazioni effettuate dal docente.

Progetto di laboratorio

Il progetto di laboratorio consiste in una serie di esercitazioni assistite mirate allo sviluppo di un semplice traduttore. Il corretto svolgimento di tali esercitazioni presuppone una buona conoscenza del linguaggio di programmazione Java e degli argomenti di teoria del corso Linguaggi Formali e Traduttori.

Modalità dell'esame di laboratorio

Per sostenere l'esame a un appello è necessario prenotarsi. L'esame di laboratorio è **orale** e **individuale**, anche se il codice è stato sviluppato in collaborazione con altri studenti. Durante l'esame vengono accertati: il corretto svolgimento della prova di laboratorio; la comprensione della sua struttura e del suo funzionamento; la comprensione delle parti di teoria correlata al laboratorio stesso.

Note importanti

- Per poter discutere il laboratorio è *necessario* aver prima superato la prova scritta relativa al modulo di teoria.

- La presentazione di codice "funzionante" non è condizione sufficiente per il superamento della prova di laboratorio. In altri termini, è possibile essere respinti presentando codice funzionante (se lo studente dimostra di non avere adeguata familiarità con il codice e i concetti correlati).
- Anche se il codice è stato sviluppato in collaborazione con altri studenti, i punteggi ottenuti dai singoli studenti sono indipendenti. Per esempio, a parità di codice presentato, è possibile che uno studente meriti 30, un altro 25 e un altro ancora sia respinto.
- Dal momento che durante la prova è possibile che venga richiesto di apportare modifiche al codice del progetto, è opportuno presentarsi all'esame con un'adeguata conoscenza del progetto e degli argomenti di teoria correlati.

Calcolo del voto finale

I voti della prova scritta e della prova di laboratorio sono espressi in trentesimi. Il voto finale è determinato calcolando la media pesata del voto della prova scritta e del laboratorio, secondo il loro contributo in CFU, e cioè

$$\text{voto finale} = \frac{\text{voto dello scritto} \times 2 + \text{voto del laboratorio}}{3}$$

La lode è attribuita agli studenti con voto finale pari a 30 e che abbiano dimostrato particolare brillantezza nello svolgimento delle esercitazioni di laboratorio.

Validità del presente testo di laboratorio

Il presente testo di laboratorio è valido sino alla sessione di settembre 2016.

1 Implementazione di un DFA in Java

Lo scopo di questo esercizio è l'implementazione di un metodo Java che sia in grado di discriminare le stringhe del linguaggio riconosciuto da un automa a stati finiti deterministico (DFA) dato. Il primo automa che prendiamo in considerazione, mostrato in Figura 1, è definito sull'alfabeto $\{0, 1\}$ e riconosce le stringhe in cui compaiono almeno 3 zeri consecutivi.

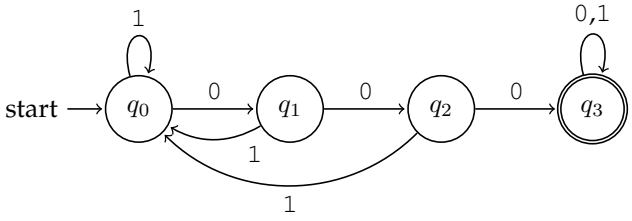

Figura 1: DFA che riconosce stringhe con 3 zeri consecutivi.

L'implementazione Java del DFA di Figura 1 è mostrata in Figura 2. L'automa è implementato nel metodo `scan` che accetta una stringa `s` e restituisce un valore booleano che indica se la stringa appartiene o meno al linguaggio riconosciuto dall'automa. Lo stato dell'automa è rappresentato per mezzo di una variabile intera `state`, mentre la variabile `i` contiene l'indice del prossimo carattere della stringa `s` da analizzare. Il corpo principale del metodo è un ciclo che, analizzando il contenuto della stringa `s` un carattere alla volta, effettua un cambiamento dello stato dell'automa secondo la sua funzione di transizione. Notare che l'implementazione assegna

```java
public class TreZeri
{
    public static boolean scan(String s)
    {
        int state = 0;
        int i = 0;

        while (state >= 0 && i < s.length()) {
            final char ch = s.charAt(i++);

            switch (state) {
            case 0:
                if (ch == '0')
                    state = 1;
                else if (ch == '1')
                    state = 0;
                else
                    state = -1;
                break;

            case 1:
                if (ch == '0')
                    state = 2;
                else if (ch == '1')
                    state = 0;
                else
                    state = -1;
                break;

            case 2:
                if (ch == '0')
                    state = 3;
                else if (ch == '1')
                    state = 0;
                else
                    state = -1;
                break;

            case 3:
                if (ch == '0' || ch == '1')
                    state = 3;
                else
                    state = -1;
                break;
            }
        }
        return state == 3;
    }

    public static void main(String[] args)
    {
        System.out.println(scan(args[0]) ? "OK" : "NOPE");
    }
}
```

Figura 2: Implementazione Java del DFA di Figura 1.

il valore -1 alla variabile `state` se viene incontrato un simbolo diverso da `0` e `1`. Tale valore non è uno stato valido, ma rappresenta una condizione di errore irrecuperabile.

Esercizio 1.1. Copiare il codice in Figura 2, compilarlo e testarlo su un insieme significativo di stringhe, per es. `010101`, `1100011001`, `10214`, ecc.

Come deve essere modificato il DFA in Figura 1 per riconoscere il linguaggio complementare, ovvero il linguaggio delle stringhe di `0` e `1` che **non** contengono 3 zeri consecutivi? Progettare e implementare il DFA modificato, e testare il suo funzionamento.

Esercizio 1.2. Progettare un DFA che riconosca il linguaggio delle costanti numeriche in virgola mobile. Esempi di tali costanti sono:

```
123    123.5    .567    +7.5    -.7    67e10    1e-2    -.7e2
```

Realizzare il DFA in Java seguendo la costruzione vista in Figura 2, assicurarsi che l'implementazione riconosca il linguaggio desiderato.

In base al particolare stato finale in cui si trova l'automa al termine del riconoscimento, cosa si può dire della costante numerica riconosciuta?

Esercizio 1.3. Modificare l'automa dell'esercizio precedente in modo che riconosca costanti numeriche precedute e/o seguite da sequenze eventualmente vuote di spazi. Modificare l'implementazione Java dell'automa conseguentemente.

Esercizio 1.4. Progettare e implementare un DFA che riconosca il linguaggio degli identificatori in un linguaggio in stile Java: un identificatore è una sequenza non vuota di lettere, numeri, ed il simbolo di sottolineatura _ che non comincia con un numero e che non può essere composto solo da un _.

Esercizio 1.5. Progettare e implementare un DFA che riconosca il linguaggio dei numeri binari (stringhe di `0` e `1`) il cui valore è multiplo di 3. Per esempio, `110` e `1001` sono stringhe del linguaggio (rappresentano rispettivamente i numeri 6 e 9), mentre `10` e `111` no (rappresentano rispettivamente i numeri 2 e 7). **Suggerimento:** usare tre stati per rappresentare il resto della divisione per 3 del numero.

Esercizio 1.6. Progettare e implementare un DFA con l'alfabeto $\{/, \star, a\}$ che riconosca il linguaggio di stringhe che contengono almeno 4 caratteri che iniziano con `/*`, che finiscono con `*/`, e che contengono una solo occorrenza della sequenza `*/`, quella finale. Verificare che il DFA accetti le stringhe `/****/`, `/*a*a*/`, `/*a/**/`, `/**a///a/a**/` e `/**/` ma non `/*/` oppure `/**/***/`.

Esercizio 1.7 (opzionale). Modificare l'automa dell'esercizio precedente in modo che riconosca il linguaggio di stringhe in cui una occorrenza della sequenza `/*` deve essere seguita (anche non immediatamente) da una occorrenza della sequenza `*/` che, a sua volta, deve essere seguita da una sequenza che consiste solo da simboli da $\{a, \star\}$. Le stringhe del linguaggio possono non avere nessuna occorrenza della sequenza `/*`. Ad esempio, il DFA deve accettare le stringe `aaa/****/aa`, `aa/*a*a*/`, `aaaa`, `/****/`, `/*aa*/`, `*/a` e `a/**/***a`, ma non `aaa/*/aa`, `a/**/***/a` oppure `aa/*aa`. Implementare l'automa seguendo la construzione vista in Figura 2.

Esercizio 1.8. Costruire il DFA equivalente al ϵ-NFA in Figura 3, e implementare il DFA seguendo la construzione vista in Figura 2.

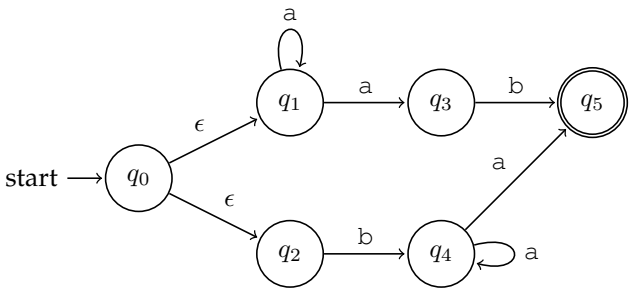

Figura 3: ϵ-NFA di Esercizio 1.8.

2 Analisi lessicale

Consideriamo un semplice linguaggio di programmazione, dove la sintassi è descritta dalla seguente grammatica:

$$
\begin{array}{rcl}
\langle prog \rangle & ::= & [\ \langle decl \rangle\ \texttt{;}\ ]^*\ \langle stmt \rangle\ \texttt{EOF} \\
\langle decl \rangle & ::= & \texttt{var}\ \texttt{ID}\ [\ \texttt{,}\ \texttt{ID}\ ]^*\ \texttt{:}\ \langle type \rangle \\
\langle type \rangle & ::= & \texttt{integer} \mid \texttt{boolean} \\
\langle expr \rangle & ::= & \langle andExpr \rangle\ [\ \texttt{||}\ \langle andExpr \rangle\ ]^* \\
\langle andExpr \rangle & ::= & \langle relExpr \rangle\ [\ \texttt{\&\&}\ \langle relExpr \rangle\ ]^* \\
\langle relExpr \rangle & ::= & \langle addExpr \rangle\ [\ \texttt{==}\ \langle addExpr \rangle \mid \texttt{<>}\ \langle addExpr \rangle \\
 & & \quad \mid \texttt{<=}\ \langle addExpr \rangle \mid \texttt{>=}\ \langle addExpr \rangle \\
 & & \quad \mid \texttt{<}\ \langle addExpr \rangle \mid \texttt{>}\ \langle addExpr \rangle\ ]^? \\
\langle addExpr \rangle & ::= & \langle mulExpr \rangle\ [\ \texttt{+}\ \langle mulExpr \rangle \mid \texttt{-}\ \langle mulExpr \rangle\ ]^* \\
\langle mulExpr \rangle & ::= & \langle unExpr \rangle\ [\ \texttt{*}\ \langle unExpr \rangle \mid \texttt{/}\ \langle unExpr \rangle\ ]^* \\
\langle unExpr \rangle & ::= & \texttt{not}\ \langle unExpr \rangle \mid \langle primary \rangle \\
\langle primary \rangle & ::= & \texttt{(}\ \langle expr \rangle\ \texttt{)} \mid \texttt{ID} \mid \texttt{NUM} \mid \texttt{true} \mid \texttt{false} \\
\langle stmt \rangle & ::= & \texttt{ID}\ \texttt{:=}\ \langle expr \rangle \\
 & \mid & \texttt{print}\ \texttt{(}\ \langle expr \rangle\ \texttt{)} \\
 & \mid & \langle stmt \rangle\ [\ \texttt{;}\ \langle stmt \rangle\ ]^*
\end{array}
$$

I terminali `ID` corrispondono all'espressione regolare `a..z|A..Z(a..z|A..Z|0..9)*`, e i terminali `NUM` corrispondono all'espressione regolare `0..9(0..9)*`.

Il linguaggio permette di scrivere programmi costituiti da due sezioni (la prima delle quali può essere omessa):

1. Dichiarazioni di variabili, dove l'elenco di variabili è preceduto dalla parola chiave `var` e seguito da `:` e un tipo, `integer` oppure `boolean`. Esempi:
   - `var x,y:integer;`
   - `var z:boolean;`

2. Sequenze di comandi, con due tipologie distinte di comando.

   - Assegnamento di un valore valori ad una variabile, usando la sintassi `:=`. Esempi: `x:=5;`, `y:=x+7;` e `z:=x&&y;`.
   - Scrittura sul terminale, usando la parola chiave `print`. Esempi: `print(x)`, `print(x||y)` e `print((5+2)*3)`.

Le espressioni che possiamo scrivere in commandi di assegnamento e `print` sono composte soltanto da numeri non negativi (ovvero sequenze di cifre decimali), operatori di somma e sottrazione `+` e `-`, operatori di moltiplicazione e divisione `*` e `/`, simboli di parentesi `(` e `)`, operatori relazionali `<`, `<=`, `==`, `>=`, `>`, e `<>`, operatori logici `&&`, `||` e `not`. Esempi:

- `34+26-5`
- `(34+26)-5`
- `x == (y+26)*5`
- `x && y == false`
- `x <= y+10`
- `not x || y == true`

Ad esempio, il seguente programma calcola la velocita media di un viaggio dati il numero di metri percorsi `s` e i secondi impiegati `t`:

```
var s, t: integer;
var datiOK: boolean;

s:=300000000;
t:=1;

datiOK:= (s >=0) && (t > 0);
print(datiOK);

print(s/t)
```

Esercizio 2.1. Si scriva in Java un analizzatore lessicale che legga da tastiera comandi scritti in questo linguaggio e per ciascuna comando stampi una sequenza di token.

- I token che corrispondono a numeri prenderanno la forma $\langle \text{NUM}, \text{valore} \rangle$, dove valore è un intero non negativo. Ad esempio, il token che corrisponde al numero 12 sarà $\langle \text{NUM}, 12 \rangle$
- I token che corrispondono agli identificatori prenderanno la forma $\langle \text{ID}, \text{“lessema”} \rangle$. Ad esempio, il token che corrisponde a `x` sarà $\langle \text{ID}, \text{“x”} \rangle$, e il token che corrisponde a `temp` sarà $\langle \text{ID}, \text{“temp”} \rangle$.
- I token che corrispondono agli elementi della sintassi che consistono di un solo carattere (ad esempio, `'('`, `'+'` e `':'`) prenderanno la forma $\langle \text{nome} \rangle$, dove il nome è il codice ASCII del carattere. Ad esempio, il token che corrisponde a `'('` sarà $\langle ' (' \rangle$, e il token che corrisponde a `':'` sarà $\langle ' : ' \rangle$.
- I token che corrispondono agli elementi della sintassi che consistono di più caratteri (ad esempio, `&&`, `<>`, `print`, `integer` e `:=`) prenderanno la forma $\langle \text{nome}, \text{“lessema”} \rangle$. Ad esempio, il token che corrisponde a `:=` sarà $\langle \text{ASSIGN}, \text{“:=”} \rangle$, e il token che corrisponde a `print` sarà $\langle \text{PRINT}, \text{“print”} \rangle$.

Definiamo una classe `Tag` in Listing 1, utilizzando un insieme opportuno di costanti intere per rappresentare il nome dei token. (Si nota che non è assolutemente necessario definire tali costanti per tutti i token: per quelli che corrispondono a un solo carattere, si può utilizzare il codice ASCII del carattere.)

Listing 1: Classe `Tag`

```
public class Tag {
    public final static int
            EOF = -1,
            NUM = 256,
            ID = 257,
            AND = 258,
            OR = 259,
            VAR = 260,
            INTEGER = 261,
            BOOLEAN = 262,
            ASSIGN = 263,
            EQ = 264,
            GE = 265,
            LE = 266,
            NE = 267,
            TRUE = 268,
            FALSE = 269,
            NOT = 270,
            PRINT = 271;
}
```

Questa scelta ha la conseguenza che l'output del nostro programma sarà della forma $\langle 271, \text{“print”}\rangle$ $\langle 40\rangle$ $\langle 257, \text{“x”}\rangle$ $\langle 259, \text{“||”}\rangle$ $\langle 257, \text{“y”}\rangle$ $\langle 41\rangle$ $\langle 59\rangle$ per il comando `print(x||y);` anzichè $\langle \text{PRINT}, \text{“print”}\rangle$ $\langle '('\rangle$ $\langle \text{ID}, \text{“x”}\rangle$ $\langle \text{OR}, \text{“||”}\rangle$ $\langle \text{ID}, \text{“y”}\rangle$ $\langle ')'\rangle$ $\langle ';'\rangle$

Nota: l'analizzatore lessicale non è preposto al riconoscimento della *struttura* dei comandi. Pertanto, esso accetterà anche comandi “errati” quali ad esempio:

- `5+)`
- `(34+26( - (2+15-( 27`
- `var 5 := print < boolean`

L'analizzatore lessicale dovrà ignorare tutti i caratteri riconosciuti come “spazi” (incluse le tabulazioni e i ritorni a capo), ma dovrà segnalare la presenza di caratteri illeciti, quali ad esempio `#` o `@`. Per semplicità, si può utilizzare un carattere particolare, come `$`, per segnalare la fine del input (e quindi produrre un token con nome EOF che corrisponde alla fine del input).

Definiamo una classe `Token` per rappresentare i token (una possibile implementazione della classe `Token` è in Listing 2). Definiamo inoltre le classe `Word` e `Number` derivate da `Token`, dove la classe `Word` rappresenta i token che corrispondono agli identificatori, alle parole chiave e agli elementi del sintassi che consistono di più caratteri (ad esempio `<=` e `&&`), e dove la classe `Number` rappresenta i token che corrispondono ai numeri. Una possibile implementazione della classe `Word` è in Listing 3.

Listing 2: Classe `Token`

```
public class Token {
    public final int tag;
    public Token(int t) { tag = t;  }
    public String toString() {return "<" + tag + ">";}
    public static final Token
            comma = new Token(','),
            colon = new Token(':'),
```

```
            semicolon = new Token(';'),
            lpar = new Token('('),
            rpar = new Token(')'),
            plus = new Token('+'),
            minus = new Token('-'),
            mult = new Token('*'),
            div = new Token('/'),
            lt = new Token('<'),
            gt = new Token('>');
}
```

Listing 3: Classe `Word`

```
public class Word extends Token {
    public String lexeme = "";
    public Word(int tag, String s) { super(tag); lexeme=s; }
    public String toString() { return "<" + tag + ", " + lexeme + ">"; }
    public static final Word
            and = new Word(Tag.AND, "&&"),
            or = new Word(Tag.OR, "||"),
            eq = new Word(Tag.EQ, "=="),
            le = new Word(Tag.LE, "<="),
            ne = new Word(Tag.NE, "<>"),
            ge = new Word(Tag.GE, ">="),
            assign = new Word(Tag.ASSIGN, ":=");
}
```

Gestione delle parole chiave. Abbiamo bisogno di un meccanismo per memorizzare le parole chiave (`var`, `integer`, `boolean`, `not`, `true`, `false` e `print`), in modo tale che possiamo distinguirle da eventuali identificatori. Un modo è di utilizzare una *tabella hash*. La classe `Hashtable` è utilizzata per immagazzinare coppie che consistono di una *chiave* (*key*) e un *valore*. La classe `Hashtable` mette a disposizione due metodi, `put` e `get`, che servono rispettivamente per:

1. inserire associazioni (chiave, valore) nella tabella;
2. recuperare il valore associato a una chiave dalla tabella.

Nel contesto di nostra implementazione di un'analizzatore lessicale, possiamo utilizzare la classe `Hashtable` nel modo seguente.

```
Hashtable<String,Word> words = new Hashtable<String,Word>();
void reserve(Word w) { words.put(w.lexeme, w); }
...
reserve(new Word(Tag.VAR, "var"));
...
Word w = (Word)words.get(s);
...
```

In questo frammento di codice, l'operazione `words.put(w.lexeme, w)` è utilizzata per associare la chiave `w.lexeme` al valore `w`. Quindi la chiamata del metodo `reserve` corrisponde all'associazione della chiave `"var"` al oggetto creato da `new Word(Tag.VAR, "var")`.

Dato una stringa `s`, `(Word)words.get(s)` corrisponde al oggetto del tipo `Word` che è associato alla chiave `s` nella tabella hash `words`. Se la chiave `s` non è stata inserita nella tabella, si otterrà come risultato `null`.

La tabella hash può essere utilizzata anche per memorizzare gli identificatori già visti durante l'analisi lessicale.

Classe `Lexer`. Una possibile struttura del programma (ispirato al testo [1, Appendice A.3]) è la seguente:

Listing 4: Analizzatore lessicale di comandi semplici

```
import java.io.*;
import java.util.*;

public class Lexer {

    public static int line = 1;
    private char peek = ' ';

    Hashtable<String,Word> words = new Hashtable<String,Word>();
    void reserve(Word w) { words.put(w.lexeme, w); }
    public Lexer() {
        reserve( new Word(Tag.VAR, "var"));

                        // ... gestire le altre parole chiave ... //

    }

    private void readch() {
        try {
            peek = (char) System.in.read();
        } catch (IOException exc) {
            peek = (char) -1; // ERROR
        }
    }

    public Token lexical_scan() {
        while (peek == ' ' || peek == '\t' || peek == '\n'  || peek == '\r') {
            if (peek == '\n') line++;
            readch();
        }

        switch (peek) {
            case ',':
                peek = ' ';
                return Token.comma;

                        // ... gestire gli altri casi ... //

            case '&':
                readch();
                if (peek == '&') {
                    peek = ' ';
                    return Word.and;
                } else {
                    System.err.println("Erroneous character"
                            + " after & : "  + peek );
                    return null;
                }

                        // ... gestire gli altri casi ... //

            default:
                if (Character.isLetter(peek)) {
                    String s = "";
```

```
                    do {
                        s+= peek;
                        readch();
                    } while (Character.isDigit(peek) ||
                            Character.isLetter(peek));
                    if ((Word)words.get(s) != null)
                        return (Word)words.get(s);
                    else {
                    Word w = new Word(Tag.ID,s);
                    words.put(s, w);
                    return w;
                    }
                } else {

                        // ... gestire il caso dei numeri ... //

                        if (peek == '$') {
                            return new Token(Tag.EOF);
                        } else {
                        System.err.println("Erroneous character: "
                                + peek );
                        return null;
                        }
                }
            }
    }

    public static void main(String[] args) {
        Lexer lex = new Lexer();

        Token tok;
        do {
            tok = lex.lexical_scan();
            System.out.println("Scan: " + tok);
        } while (tok.tag != Tag.EOF);
    }

}
```

Esercizio 2.2. Consideriamo la seguente versione modificata della produzione per $\langle stmt \rangle$:

$$
\begin{array}{rcl}
\langle stmt \rangle & ::= & \texttt{ID} \texttt{ :=} \langle expr \rangle \\
& | & \texttt{print (} \langle expr \rangle \texttt{)} \\
& | & \texttt{if} \langle expr \rangle \texttt{then} \langle stmt \rangle [\texttt{else} \langle stmt \rangle]^? \\
& | & \texttt{while} \langle expr \rangle \texttt{do} \langle stmt \rangle \\
& | & \texttt{begin} \langle stmt \rangle [\texttt{;} \langle stmt \rangle]^* \texttt{end}
\end{array}
$$

Estendere l'implementazione del analizzatore lessicale per gestire comandi di programmi scritti nel linguaggio modificato (cioè nel linguaggio con la versione modificata della produzione per $\langle stmt \rangle$).

Esercizio 2.3. Consideriamo la seguente nuova definizione di identificatori (`ID`): un identificatore è composto da una sequenza non vuota di lettere, numeri, ed il simbolo di sottolineatura _ che non comincia con un numero e che non può essere composto solo da _. Più precisamente, i terminali `ID` corrispondono all'espressione regolare $((\texttt{a..z}|\texttt{A..Z})|(\_(\_)^*(\texttt{a..z}|\texttt{A..Z}|\texttt{0..9})))(\texttt{a..z}|\texttt{A..Z}|\texttt{0..9}|\_)^*$. Estendere il metodo `lexer_scan` per gestire identificatori che corrispondono alla nuova definizione. **Suggerimento:** utilizzare la soluzione del Esercizio 1.4 come punto d'inizio.

Lettura da un file. La lettura di un programma da un file, anzichè dalla tastiera come in Listing 4, può essere realizzata nel modo illustrato in Listing 5. Il metodo `main` crea un oggetto della classe `BufferedReader`, che poi è passato come parametro al metodo `lexical_scan`, e a sua volta a `readch`.

Listing 5: Lettura da un file

```
private void readch(BufferedReader br) {
    try {
        peek = (char) br.read();
    } catch (IOException exc) {
        peek = (char) -1; // ERROR
    }
}

public Token lexical_scan(BufferedReader br) {
            // ... sostituire readch() con readch(br) ... //
            // ... //
            if (peek == (char)-1) {   // sostituisce if (peek == '$')
                  return new Token(Tag.EOF);
            }
            // ... //
}

public static void main(String[] args) {
    Lexer lex = new Lexer();
    String path = "...path..."; // il percorso del file da leggere
    try {
        BufferedReader br = new BufferedReader(new FileReader(path));
        Token tok;
        do {
            tok = lex.lexical_scan(br);
            System.out.println("Scan: " + tok);
        } while (tok.tag != Tag.EOF);
        br.close();
    } catch (IOException e) {e.printStackTrace();}
}
```

3 Analisi sintattica

Esercizio 3.1. Si scriva un analizzatore sintattico a discesa ricorsiva che parsifichi espressioni aritmetiche molto semplici, composte soltanto da numeri non negativi (ovvero sequenze di cifre decimali), operatori di somma e sottrazione + e −, operatori di moltiplicazione e divisione $*$ e /, simboli di parentesi (e). In particolare, l'analizzatore deve riconoscere le espressioni generate dalla grammatica che segue:

$$
\begin{array}{rcl}
\langle start \rangle & ::= & \langle expr \rangle \ \texttt{EOF} \\
\\
\langle expr \rangle & ::= & \langle term \rangle \ \langle exprp \rangle \\
\\
\langle exprp \rangle & ::= & + \ \langle term \rangle \ \langle exprp \rangle \\
& | & - \ \langle term \rangle \ \langle exprp \rangle \\
& | & \varepsilon \\
\\
\langle term \rangle & ::= & \langle fact \rangle \ \langle termp \rangle \\
\\
\langle termp \rangle & ::= & \star \ \langle fact \rangle \ \langle termp \rangle \\
& | & / \ \langle fact \rangle \ \langle termp \rangle \\
& | & \varepsilon \\
\\
\langle fact \rangle & ::= & (\ \langle expr \rangle\ ) \mid \texttt{NUM}
\end{array}
$$

Il programma deve fare uso dell'analizzatore lessicale sviluppato in precedenza. Si nota che l'insieme di token corrispondente alla grammatica di questa sezione è un sottoinsieme dell'insieme di token corrispondente alla grammatica della Sezione 2. Nei casi in cui l'input non corrisponde alla grammatica, l'output del programma deve consistere di un messaggio di errore (come illustrato nelle lezioni in aula) indicando la procedura in esecuzione quando l'errore è stato individuato.

Segue una possibile struttura del programma (ispirato al testo [1, Appendice A.8]). Consideriamo due varianti: nella prima (in Listing 6), l'input è letto dalla tastiera; invece nella seconda (in Listing 7), l'input è letto da un file.

Listing 6: Analizzatore sintattico di espressioni semplici: lettura dalla tastiera

```
import java.io.*;

public class Parser {
    private Lexer lex;
    private Token look;

    public Parser(Lexer l) {
        lex = l;
        move();
    }

    void move() {
        look = lex.lexical_scan();
        System.err.println("token = " + look);
    }

    void error(String s) {
        throw new Error("near line " + lex.line + ": " + s);
    }

    void match(int t) {
        if (look.tag == t) {
            if (look.tag != Tag.EOF) move();
        } else error("syntax error");
    }

    public void start() {  // la procedura start puo' essere estesa (opzionale)
        expr();
        match(Tag.EOF);
```

```
    }

    private void expr() {  // la procedura expr puo' essere estesa (opzionale)
        term();
        exprp();
    }

    private void exprp() {
        switch (look.tag) {
        case '+':
            match('+');
            term();
            exprp();
            break;

        case '-':
            // ... gestire gli altri casi ... //
        }
    }

    private void term() {
        // ... riempire ... //
    }

    private void termp() {
        switch (look.tag) {
        case '*':
            match('*');
            fact();
            termp();
            break;

        // ... gestire gli altri casi ... //
    }

    private void fact() {
        switch (look.tag) {
            // ... gestire tutti i casi ... //
        }
    }

    public static void main(String[] args) {
        Lexer lex = new Lexer();
        Parser parser = new Parser(lex);
        parser.start();
    }
}
```

Listing 7: Analizzatore sintattico di espressioni semplici: lettura da un file (`Lexer_extended` è la versione estesa di `Lexer` per gestire lettura da un file, descritto in Sezione 2)

```
import java.io.*;

public class Parser {
    private Lexer_extended lex;
    private BufferedReader pbr;
    private Token look;

    public Parser(Lexer_extended l, BufferedReader br) {
```

```
        lex = l;
        pbr = br;
        move();
    }

    void move() {
        look = lex.lexical_scan(pbr);
        System.err.println("token = " + look);
    }

            // ... come per il caso della lettura dalla tastiera ... //

    public static void main(String[] args) {
        Lexer_extended lex = new Lexer_extended();
        String path = "...path..."; // il percorso del file da leggere
        try {
            BufferedReader br = new BufferedReader(new FileReader(path));
            Parser parser = new Parser(lex, br);
            parser.start();
            br.close();
        } catch (IOException e) {e.printStackTrace();}
    }
}
```

4 Valutatore di espressioni semplici

Esercizio 4.1. Modificare l'analizzatore sintattico di Esercizio 3.1 in modo da valutare le espressioni aritmetiche semplici, facendo riferimento alla definizione guidata dalla sintassi seguente:

$$
\begin{array}{rcl}
\langle start \rangle & ::= & \langle expr \rangle \ \texttt{EOF} \ \{ \ print(expr.val) \ \} \\
\\
\langle expr \rangle & ::= & \langle term \rangle \ \{ \ exprp.i = term.val \ \} \ \langle exprp \rangle \ \{ \ expr.val = exprp.val \ \} \\
\\
\langle exprp \rangle & ::= & + \ \langle term \rangle \ \{ \ exprp_1.i = exprp.i + term.val \ \} \ \langle exprp_1 \rangle \ \{ \ exprp.val = exprp_1.val \ \} \\
& | & - \ \langle term \rangle \ \{ \ exprp_1.i = exprp.i - term.val \ \} \ \langle exprp_1 \rangle \ \{ \ exprp.val = exprp_1.val \ \} \\
& | & \varepsilon \ \{ \ exprp.val = exprp.i \ \} \\
\\
\langle term \rangle & ::= & \langle fact \rangle \ \{ \ termp.i = fact.val \ \} \ \langle termp \rangle \ \{ \ term.val = termp.val \ \} \\
\\
\langle termp \rangle & ::= & * \ \langle fact \rangle \ \{ \ termp_1.i = termp.i * fact.val \ \} \ \langle termp_1 \rangle \ \{ \ termp.val = termp_1.val \ \} \\
& | & / \ \langle fact \rangle \ \{ \ termp_1.i = termp.i / fact.val \ \} \ \langle termp_1 \rangle \ \{ \ termp.val = termp_1.val \ \} \\
& | & \varepsilon \ \{ \ termp.val = termp.i \ \} \\
\\
\langle fact \rangle & ::= & \texttt{(} \ \langle expr \rangle \ \texttt{)} \ \{ \ fact.val = expr.val \ \} \\
& | & \texttt{NUM} \ \{ \ fact.val = \texttt{NUM}.value \ \}
\end{array}
$$

Si nota che un indice (cioè 1) è usato per distinguere due diverse occorrenze dello stesso non-terminale (ad esempio, $\langle exprp \rangle$) nella stessa produzione. Inoltre, si nota che il terminale `NUM` ha l'attributo *value*, che è il valore numerico del terminale, fornito dall'analizzatore lessicale.

Una possibile struttura del programma è la seguente. **Nota:** come indicato, è fortemente consigliato la creazione di una nuova classe (chiamata `Valutatore` in Listing 8).

Listing 8: Valutazione di espressioni semplici

```
import java.io.*;

public class Valutatore {
    private Lexer lex;
    private Token look;
    // se l'input e' letto da un file, creare un campo del tipo BufferedReader,
    // come in Esercizio 3.1

    public Valutatore(Lexer l) {
        lex = l;
        move();
        // se l'input e' letto da un file, creare un oggetto del tipo
        // BufferedReader, come in Esercizio 3.1
    }

    void move() {
        // come in Esercizio 3.1
    }

    void error(String s) {
        // come in Esercizio 3.1
    }

    void match(int t) {
        // come in Esercizio 3.1
    }

    public void start() {
        int expr_val;

        expr_val = expr();
        match(Tag.EOF);

        System.out.println(expr_val);
    }
        // la procedura start puo' essere estesa
        // come in Esercizio 3.1 (opzionale)

    private int expr() {
        int term_val, exprp_val;

        term_val = term();
        exprp_val = exprp(term_val);

        return exprp_val;
    }
        // la procedura expr puo' essere estesa
        // come in Esercizio 3.1 (opzionale)

    private int exprp(int exprp_i) {
        int term_val, exprp_val;

        switch (look.tag) {
        case Tag.PLUS:
            match(Tag.PLUS);
            term_val = term();
            exprp_val = exprp(exprp_i + term_val);
```

```
            break;

        case Tag.MINUS:
            // ... completare ... //
    }

    private int term() {
        // ... completare ... //
    }

    private int termp(int termp_i) {
        int fact_val, termp_val;

        switch (look.tag) {
        case Tag.TIMES:
            match(Tag.TIMES);
            fact_val = fact();
            termp_val = termp(termp_i * fact_val);
            break;

            // ... completare ... //
        }
    }

    private int fact() {
        int fact_val;

        switch (look.tag) {
        case Tag.NUM:
            fact_val = Integer.parseInt(look.text);
            // ... completare ... //
        }

        return fact_val;
    }
}
```

Il main per eseguire il tutto è simile a quelli dei Listing 6 e Listing 7.

Riferimenti bibliografici

[1] Aho, Alfred V., Lam, Monica S., Sethi, Ravi, and Ullman, Jeffrey D. Compilatori: Principi, tecniche e strumenti. *Pearson Paravia Bruno Mondadori S.p.A.*, 2009.