LINGUAGENS FORMAIS E AUTÓMATOS

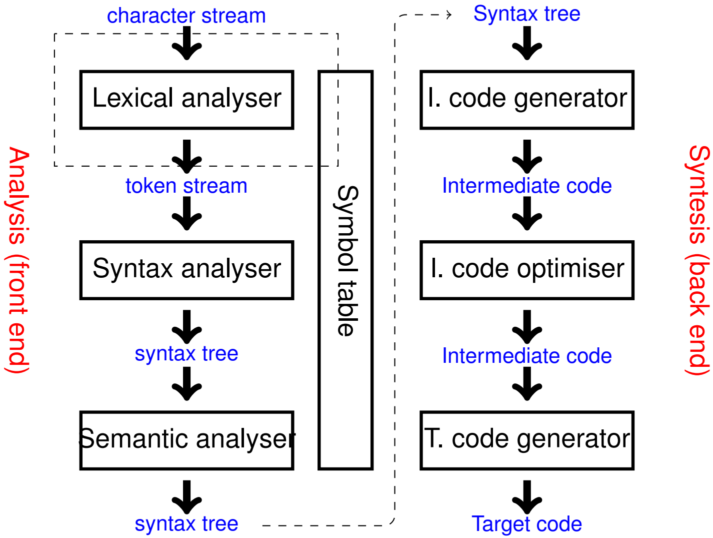
LINGUAGENS REGULARES E EXPRESSÕES REGULARES

Artur Pereira <artur@ua.pt>

DETI, Universidade de Aveiro

Sumário

Papel da análise lexical



Papel da análise lexical

- Lexical analysis
  - Convert the stream of characters into a sequence of lexemes/tokens
  - A lexeme/token is a tuple <token-name, attribute-value>
  - token-name is an abstract symbol representing a type of input
  - attribute-value represents the actual value of that element
  - Example:

    ```
    pos = pos + vel * 5;
    ```

    is converted to

    ```
    <id,pos> <=,> <id,pos> <+,> <id,vel>
    <*,> <int,5>
    ```

  - Typically, blanks are discard by the lexical analyser
  - Token patterns are represented by regular languages

Definição de linguagem regular

A classe das **linguagens regulares** sobre o alfabeto A define-se indutivamente da seguinte forma:

1. O conjunto vazio, $\emptyset$, é uma linguagem regular (LR).
2. Qualquer que seja o $a \in A$, o conjunto $\{a\}$ é uma LR.
3. Se L_1 e L_2 são linguagens regulares, então $L_1 \cup L_2$ é uma LR.
4. Se L_1 e L_2 são linguagens regulares, então $L_1.L_2$ é uma LR.
5. Se L_1 é uma linguagem regular, então $(L_1)^*$ é uma LR.
6. Nada mais é linguagem regular.

- Note que $\{\varepsilon\}$ é uma LR, uma vez que $\{\varepsilon\} = \emptyset^*$.

$\mathcal{Q}$ Qualquer linguagem finita é uma LR. *Mostre-o com base nesta definição*

$\mathcal{Q}$ *Com base nesta definição, mostre que o conjunto dos números binários começados em 1 e terminados em 0 é uma LR sobre o alfabeto* $A = \{0, 1\}$

Definição de expressão regular

O conjunto das **expressões regulares** sobre o alfabeto A define-se indutivamente da seguinte forma:

1. () é uma expressão regular (ER) que representa a LR $\{\}$.
2. Qualquer que seja o $a \in A$, a é uma ER que representa a LR $\{a\}$.
3. Se e_1 e e_2 são ER representando respectivamente as LR L_1 e L_2, então $(e_1|e_2)$ é uma ER representando a LR $L_1 \cup L_2$.
4. Se e_1 e e_2 são ER representando respectivamente as LR L_1 e L_2, então $(e_1 e_2)$ é uma ER representando a LR $L_1.L_2$.
5. Se e_1 é uma ER representando a LR L_1, então $e_1{}^*$ é uma ER representando a LR $(L_1)^*$.
6. Nada mais é expressão regular.

- É habitual representar-se por ε a ER $()^*$. Representa a linguagem $\{\varepsilon\}$.

PROJETO DE UMA EXPRESSÃO REGULAR

$\mathcal{Q}$ *Determine uma ER que representa o conjunto dos números binários começados em* `1` *e terminados em* `0`.

$\mathcal{R}$ $\mathtt{1(0|1)^{*}0}$

$\mathcal{Q}$ *Determine uma ER que represente as sequências definidas sobre o alfabeto* $A = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ *que satisfazem o requisito de qualquer* `b` *ter um* `a` *imediatamente à sua esquerda e um* `c` *imediatamente à sua direita.*

$\mathcal{R}$ $\mathtt{(a|abc|c)}*$

$\mathcal{Q}$ *Determine uma ER que represente as sequências binárias com um número par de zeros.*

$\mathcal{R}$ $\mathtt{1^{*}(01^{*}01^{*})^{*}}$

Propriedades das expressões regulares (1)

Operação de escolha — $|$

- comutativa: $e_1 \mid e_2 = e_2 \mid e_1$
- associativa: $e_1 \mid (e_2 \mid e_3) = (e_1 \mid e_2) \mid e_3 = e_1 \mid e_2 \mid e_3$
- existência de elemento neutro: $e_1 \mid () = () \mid e_1 = e_1$
- idempotência: $e_1 \mid e_1 = e_1$

Operação de concatenação — .

- associativa: $e_1(e_2 e_3) = (e_1 e_2) e_3 = e_1 e_2 e_3$
- existência de elemento neutro: $e_1 \varepsilon = \varepsilon e_1 = e_1$
- existência de elemento absorvente: $e_1() = ()e_1 = ()$
- não goza da propriedade comutativa

Propriedades das expressões regulares (2)

Operações de escolha e concatenação

- distributiva à esquerda da concatenação em relação à escolha:
  $e_1(e_2 \mid e_3) = e_1 e_2 \mid e_1 e_3$
- distributiva à direita da concatenação em relação à escolha:
  $(e_1 \mid e_2)e_3 = e_1 e_3 \mid e_2 e_3$

Operação de fecho

- $(e^*)^* = e^*$
- $(e_1^* \mid e_2^*)^* = (e_1 \mid e_2)^*$
- $(e_1 \mid e_2^*)^* = (e_1 \mid e_2)^*$

- $(e_1 \mid e_2)^* \neq e_1^* \mid e_2^*$
- $(e_1 e_2)^* \neq e_1^* e_2^*$

Simplificação notacional

- Na escrita de expressões regulares assume-se que a operação de fecho (*) tem precedência em relação à operação de concatenação e que esta tem precedência em relação à operação de escolha ($|$).
- O uso destas precedências em conjunto com as propriedades associativas da concatenação e da escolha permite a queda de alguns parêntesis e consequentemente uma notação simplificada.

Exemplo:

$$e_1|e_2.e_3^* = e_1|(e_2.(e_3^*))$$

SIMPLIFICAÇÃO NOTACIONAL

$\mathcal{Q}$ *Determine uma ER que representa o conjunto dos números binários começados em* `1` *e terminados em* `0`*.*

$\mathcal{R}$ $\mathtt{1}(\mathtt{0}|\mathtt{1})^*\mathtt{0} = (\mathtt{1}((\mathtt{0}|\mathtt{1})^*))\mathtt{0}$

$\mathcal{Q}$ *Determine uma ER que represente as sequências definidas sobre o alfabeto* $A = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ *que satisfazem o requisito de qualquer* `b` *ter um* `a` *imediatamente à sua esquerda e um* `c` *imediatamente à sua direita.*

$\mathcal{R}$ $(\mathtt{a}|\mathtt{abc}|\mathtt{c})* = ((\mathtt{a}|((\mathtt{ab})\mathtt{c}))|\mathtt{c})*$

$\mathcal{Q}$ *Determine uma ER que represente as sequências binárias com um número par de zeros.*

$\mathcal{R}$ $\mathtt{1}^*(\mathtt{0}\mathtt{1}^*\mathtt{0}\mathtt{1}^*)^* = (\mathtt{1}^*)(((((\mathtt{0}(\mathtt{1}^*))\mathtt{0})(\mathtt{1}^*)))^*)$

Simplificação notacional

$\mathcal{Q}$ *Sobre o alfabeto* $A = \{\texttt{0}, \texttt{1}\}$ *construa uma expressão regular que reconheça a linguagem*

$$L = \{\omega \in A^* \ : \ \#(\texttt{0}, \omega) = 2\}$$

$\mathcal{R}$ $\texttt{1}^*\texttt{0}\texttt{1}^*\texttt{0}\texttt{1}^*$

$\mathcal{Q}$ *Sobre o alfabeto* $A = \{\texttt{a}, \texttt{b}, \cdots, \texttt{z}\}$ *construa uma expressão regular que reconheça a linguagem*

$$L = \{\omega \in A^* \ : \ \#(\texttt{a}, \omega) = 3\}$$

$\mathcal{R}$ $(\texttt{b}|\texttt{c}|\cdots|\texttt{z})^*\texttt{a}(\texttt{b}|\texttt{c}|\cdots|\texttt{z})^*\texttt{a}(\texttt{b}|\texttt{c}|\cdots|\texttt{z})^*\texttt{a}(\texttt{b}|\texttt{c}|\cdots|\texttt{z})^*$

Extensões notacionais (1)

- uma ou mais ocorrências:

  $e^+ = e.e^*$

- uma ou nenhuma ocorrência:

  $e? = (e|\varepsilon)$

- um símbolo do sub-alfabeto dado:

  $[a_1 a_2 a_3 \cdots a_n] = (a_1 \mid a_2 \mid a_3 \mid \cdots \mid a_n)$

- um símbolo do sub-alfabeto dado:

  $[a_1 - a_n] = (a_1 \mid \cdots \mid a_n)$

- um símbolo do alfabeto fora do conjunto dado:

  $[\hat{}\, a_1 a_2 a_3 \cdots a_n]$

- um símbolo do alfabeto fora do conjunto dado:

  $[\hat{}\, a_1 - a_n]$

Extensões notacionais (2)

- n ocorrências de:

$$e\{n\} = \underbrace{e.e.\cdots.e}_{n}$$

- de n_1 a n_2 ocorrências:

$$e\{n_1, n_2\} = \underbrace{e.e.\cdots.e}_{n_1, n_2}$$

- n ou mais ocorrências:

$$e\{n,\} = \underbrace{e.e.\cdots.e}_{n,}$$

Extensões notacional

$\mathcal{Q}$ *Sobre o alfabeto* $A = \{\mathtt{0}, \mathtt{1}\}$ *construa uma expressão regular que reconheça a linguagem*

$$L = \{\omega \in A^* \ : \ \#(\mathtt{0}, \omega) = 2\}$$

$\mathcal{R}$ $\mathtt{1}^*\mathtt{0}\mathtt{1}^*\mathtt{0}\mathtt{1}^* = (\mathtt{1}^*\mathtt{0})\{2\}\mathtt{1}^*$

$\mathcal{Q}$ *Sobre o alfabeto* $A = \{\mathtt{a}, \mathtt{b}, \cdots, \mathtt{z}\}$ *construa uma expressão regular que reconheça a linguagem*

$$L = \{\omega \in A^* \ : \ \#(\mathtt{a}, \omega) = 3\}$$

$\mathcal{R}$ $(\mathtt{b}|\mathtt{c}|\cdots|\mathtt{z})^*\mathtt{a}(\mathtt{b}|\mathtt{c}|\cdots|\mathtt{z})^*\mathtt{a}(\mathtt{b}|\mathtt{c}|\cdots|\mathtt{z})^*\mathtt{a}(\mathtt{b}|\mathtt{c}|\cdots|\mathtt{z})^* =$
$([\mathtt{b}\text{–}\mathtt{z}]^*\mathtt{a})\{3\}[\mathtt{b}\text{–}\mathtt{z}]^*$

Extensões notacionais (3)

Expressões regulares especiais frequentes

- $.$ um símbolo qualquer diferente de \n
- $\hat{}$ palavra vazia no início de linha
- $\$$ palavra vazia no fim de linha
- $\backslash <$ palavra vazia no início de palavra
- $\backslash >$ palavra vazia no fim de palavra

Definição de gramática

Uma gramática é um quádruplo $G = (T, N, P, S)$, onde

- T é um conjunto finito não vazio de símbolos terminais;
- N, sendo $N \cap T = \emptyset$, é um conjunto finito não vazio de símbolos não terminais;
- P é um conjunto de produções (ou regras de rescrita), cada uma da forma $\alpha \to \beta$;
- $S \in N$ é o símbolo inicial.

- $\alpha \in (N \cup T)^* N (N \cup T)^*$
- $\beta \in (N \cup T)^*$
- As restrições a α e β definem uma taxonomia das linguagens (gramáticas)

Gramáticas regulares

$\mathcal{D}$ Uma gramática $G = (T, N, P, S)$ diz-se **regular** se, para qualquer produção $(\alpha \to \beta) \in P$, as duas condições seguintes são satisfeitas

$$\begin{aligned} \alpha &\in N \\ \beta &\in T^* \cup T^* N \end{aligned}$$

- A linguagem gerada por uma gramática regular é regular
  - Logo, é possível converter uma gramática regular numa expressão regular que represente a mesma linguagem e vice-versa
- As gramáticas regulares são fechadas sob as operações de reunião, concatenação, fecho, intersecção e complementação.

Operações sobre gramáticas regulares

Reunião de gramáticas regulares

$\mathcal{D}$ Sejam $G_1 = (T_1, N_1, P_1, S_1)$ e $G_2 = (T_2, N_2, P_2, S_2)$ duas gramáticas regulares quaisquer, com $N_1 \cap N_2 = \emptyset$. A gramática $G = (T, N, P, S)$ onde

$$\begin{array}{rcl} T & = & T_1 \cup T_2 \\ N & = & N_1 \cup N_2 \cup \{S\} \quad \text{com} \quad S \notin (N_1 \cup N_2) \\ P & = & \{S \to S_1, S \to S_2\} \cup P_1 \cup P_2 \end{array}$$

é regular e gera a linguagem $L = L(G_1) \cup L(G_2)$.

- Para $i = 1, 2$, a nova produção $S \to S_i$ permite que G gere a linguagem $L(G_i)$

Operações sobre gramáticas regulares

$\mathcal{Q}$ Sobre o conjunto de terminais $T = \{\texttt{a}, \texttt{b}, \texttt{c}\}$, determine uma gramática regular que represente a linguagem

$$L = L1 \cup L_2$$

sabendo que

$$L_1 = \{\texttt{a}\omega \ : \ \omega \in T^*\}$$
$$L_2 = \{\omega\texttt{a} \ : \ \omega \in T^*\}$$

Comece por obter as gramáticas regulares que representam L_1 e L_2.

$\mathcal{R}$

$$\begin{aligned} S_1 &\to \texttt{a} \ X_1 \\ X_1 &\to \texttt{a} \ X_1 \\ X_1 &\to \texttt{b} \ X_1 \\ X_1 &\to \texttt{c} \ X_1 \\ X_1 &\to \varepsilon \end{aligned}$$

$$\begin{aligned} S_2 &\to \texttt{a} \ S_2 \\ S_2 &\to \texttt{b} \ S_2 \\ S_2 &\to \texttt{c} \ S_2 \\ S_2 &\to \texttt{a} \end{aligned}$$

$$\begin{aligned} S &\to S_1 \ | \ S_2 \\ S_1 &\to \texttt{a} \ X_1 \\ X_1 &\to \texttt{a} \ X_1 \ | \ \texttt{b} \ X_1 \ | \ \texttt{c} \ X_1 \\ X_1 &\to \varepsilon \\ S_2 &\to \texttt{a} \ S_2 \ | \ \texttt{b} \ S_2 \ | \ \texttt{c} \ S_2 \\ S_2 &\to \texttt{a} \end{aligned}$$

Operações sobre gramáticas regulares

Concatenação de gramáticas regulares

$\mathcal{D}$ Sejam $G_1 = (T_1, N_1, P_1, S_1)$ e $G_2 = (T_2, N_2, P_2, S_2)$ duas gramáticas regulares quaisquer, com $N_1 \cap N_2 = \emptyset$. A gramática $G = (T, N, P, S)$ onde

$$\begin{aligned}
T &= T_1 \cup T_2 \\
N &= N_1 \cup N_2 \\
P &= \{A \to \omega S_2 \;:\; (A \to \omega) \in P_1 \;\wedge\; \omega \in {T_1}^*\} \\
&\quad \cup \{A \to \omega \;:\; (A \to \omega) \in P_1 \;\wedge\; \omega \in {T_1}^* N_1\} \\
&\quad \cup P_2 \\
S &= S_1
\end{aligned}$$

é regular e gera a linguagem $L = L(G_1) \cdot L(G_2)$.

- As produções da primeira gramática do tipo $\beta \in T^*$ ganham o símbolo inicial da segunda gramática no fim
- As produções da primeira gramática do tipo $\beta \in T^* N$ mantêm-se inalteradas
- As produções da segunda gramática mantêm-se inalteradas

Operações sobre gramáticas regulares

$\mathcal{Q}$ Sobre o conjunto de terminais $T = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, determine uma gramática regular que represente a linguagem

$$L = L_1 \cdot L_2$$

sabendo que

$$L_1 = \{\mathtt{a}\omega \;:\; \omega \in T^*\}$$

$$L_2 = \{\omega\mathtt{a} \;:\; \omega \in T^*\}$$

$\mathcal{R}$

$$\begin{aligned} S_1 &\to \mathtt{a}\ X_1 \\ X_1 &\to \mathtt{a}\ X_1 \\ X_1 &\to \mathtt{b}\ X_1 \\ X_1 &\to \mathtt{c}\ X_1 \\ X_1 &\to \varepsilon \end{aligned}$$

$$\begin{aligned} S_2 &\to \mathtt{a}\ S_2 \\ S_2 &\to \mathtt{b}\ S_2 \\ S_2 &\to \mathtt{c}\ S_2 \\ S_2 &\to \mathtt{a} \end{aligned}$$

$$\begin{aligned} S_1 &\to \mathtt{a}\ X_1 \\ X_1 &\to \mathtt{a}\ X_1 \mid \mathtt{b}\ X_1 \mid \mathtt{c}\ X_1 \\ X_1 &\to S_2 \\ S_2 &\to \mathtt{a}\ S_2 \mid \mathtt{b}\ S_2 \mid \mathtt{c}\ S_2 \\ S_2 &\to \mathtt{a} \end{aligned}$$

Operações sobre gramáticas regulares

Fecho de Kleene de gramáticas regulares

$\mathcal{D}$ Seja $G_1 = (T_1, N_1, P_1, S_1)$ uma gramática regular qualquer. A gramática $G = (T, N, P, S)$ onde

$$\begin{aligned} T &= T_1 \\ N &= N_1 \cup \{S\} \quad \text{com} \quad S \notin N_1 \\ P &= \{S \to \varepsilon, S \to S_1\} \\ &\quad \cup \{A \to \omega S \,:\, (A \to \omega) \in P_1 \,\wedge\, \omega \in {T_1}^*\} \\ &\quad \cup \{A \to \omega \,:\, (A \to \omega) \in P_1 \,\wedge\, \omega \in {T_1}^* N_1\} \end{aligned}$$

é regular e gera a linguagem $L = (L(G_1))^*$.

- As produções que terminam num não terminal mantêm-se inalteradas
- As produções que só têm terminais ganham o símbolo inicial no fim
- As novas produções $S \to \varepsilon$ e $S \to S_1$ garante que $(L(G_1))^n \subseteq L(G)$, para qualquer $n \geq 0$

Operações sobre gramáticas regulares

$\mathcal{Q}$ Sobre o conjunto de terminais $T = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, determine uma gramática regular que represente a linguagem

$$L = {L_1}^*$$

sabendo que

$$L_1 = \{\mathtt{a}\omega \,:\, \omega \in T^*\}$$

Comece por obter a gramática regular que representa L_1.

$\mathcal{R}$

$$\begin{aligned} S_1 &\to \mathtt{a}\ X_1 \\ X_1 &\to \mathtt{a}\ X_1 \\ X_1 &\to \mathtt{b}\ X_1 \\ X_1 &\to \mathtt{c}\ X_1 \\ X_1 &\to \varepsilon \end{aligned}$$

$$\begin{aligned} S &\to \varepsilon \mid S_1 \\ S_1 &\to \mathtt{a}\ X_1 \\ X_1 &\to \mathtt{a}\ X_1 \mid \mathtt{b}\ X_1 \mid \mathtt{c}\ X_1 \\ X_1 &\to S \end{aligned}$$

Equivalência com as expressões regulares

Conversão de uma ER em uma GR

- Basta obter GR para as expressões regulares primitivas e aplicar as operações regulares sobre GR
- A GR para a ER ε é dada por

$$S \to \varepsilon$$

- A GR para a ER $\mathtt{a}$, qualquer que seja o $\mathtt{a}$, é dada por

$$S \to \mathtt{a}$$

$\mathcal{Q}$ Obtenha uma GR equivalente à ER $e = (\mathtt{a}|\mathtt{b}|\mathtt{c})^*(\mathtt{bb}|\mathtt{cc})(\mathtt{a}|\mathtt{b}|\mathtt{c})^*$

Equivalência com as expressões regulares

Conversão de uma GR em uma ER

Seja $G = (T, N, P, S)$ uma gramática regular qualquer. Uma ER que represente a mesma linguagem que a gramática G pode ser obtida por um processo de transformação de equivalência.

Algoritmo de conversão

1. Converte-se a gramática $G = (T, N, P, S)$ no conjunto de triplos seguinte:

$$\begin{array}{rcl} \mathcal{E} & = & \{(E, \varepsilon, S)\} \\ & \cup & \{(A, \omega, B) \;:\; (A \rightarrow \omega B) \in P \;\wedge\; B \in N\} \\ & \cup & \{(A, \omega, \varepsilon) \;:\; (A \rightarrow \omega) \in P \wedge \omega \in T^*\} \end{array}$$

   com $E \notin N$.

2. Removem-se, por transformações de equivalência, um a um, todos os símbolos de N, até se obter um único triplo da forma (E, e, ε).

Equivalência com as expressões regulares

Remoção dos símbolos de N

- Para cada símbolo $B \in N$
  - (a) Substituir todos os triplos da forma (A, β_i, B) por um único (A, ω_1, B), onde $\omega_1 = \beta_1 \mid \beta_2 \mid \cdots \mid \beta_n$
  - (b) Substituir todos os triplos da forma (B, α_i, B) por um único (B, ω_2, B), onde $\omega_2 = \alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_m$
  - (c) Substituir todos os triplos da forma (B, γ_i, C) por um único (B, ω_3, C), onde $\omega_3 = \gamma_1 \mid \gamma_2 \mid \cdots \mid \gamma_k$
  - (d) Substituir o triplo de triplos $((A, \omega_1, B), (B, \omega_2, B), (B, \omega_3, C))$ pelo triplo $(A, \omega_1 {\omega_2}^* \omega_3, C)$

Equivalência com as expressões regulares

$\mathcal{Q}$ Obtenha uma ER equivalente à gramática regular seguinte

$$S \to \texttt{a}\ S \mid \texttt{b}\ S \mid \texttt{c}\ S \mid \texttt{aba}\ X$$
$$X \to \texttt{a}\ X \mid \texttt{b}\ X \mid \texttt{c}\ X \mid \varepsilon$$

$\mathcal{R}$

$$\begin{aligned}
\mathcal{E} &= \{(E,\varepsilon,S),(S,\texttt{a},S),(S,\texttt{b},S),(S,\texttt{c},S),(S,\texttt{aba},X),\\
&\quad (X,\texttt{a},X),(X,\texttt{b},X),(X,\texttt{c},X),(X,\varepsilon,\varepsilon)\}\\
&= \{(E,\varepsilon,S),(S,\texttt{a}|\texttt{b}|\texttt{c},S),(S,\texttt{aba},X),\\
&\quad (X,\texttt{a},X),(X,\texttt{b},X),(X,\texttt{c},X),(X,\varepsilon,\varepsilon)\}\\
&= \{(E,(\texttt{a}|\texttt{b}|\texttt{c})^*\texttt{aba}),\\
&\quad (X,\texttt{a},X),(X,\texttt{b},X),(X,\texttt{c},X),(X,\varepsilon,\varepsilon)\}\\
&= \{(E,(\texttt{a}|\texttt{b}|\texttt{c})^*\texttt{aba},X),(X,(\texttt{a}|\texttt{b}|\texttt{c}),X),(X,\varepsilon,\varepsilon)\}\\
&= \{(E,(\texttt{a}|\texttt{b}|\texttt{c})^*\texttt{aba}(\texttt{a}|\texttt{b}|\texttt{c})^*,\varepsilon)\}
\end{aligned}$$