Týden 4

Přednáška

Regulární výrazy

Nejprve jsme si připomněli přesnou definici regulárních výrazů.
$RV(\Sigma)$ … množina *regulárních výrazů* nad abecedou Σ;
je to množina řetězců v abecedě $\Sigma \cup \{\, \emptyset, \varepsilon, +, \cdot, {}^*, (,) \,\}$, kterou induktivně definujeme takto:

i) $\emptyset \in RV(\Sigma), \quad \epsilon \in RV(\Sigma), \quad a \in RV(\Sigma)$ pro vš. $a \in \Sigma$;

ii) když $\alpha, \beta \in RV(\Sigma)$, pak také $(\alpha + \beta) \in RV(\Sigma)$, $(\alpha \cdot \beta) \in RV(\Sigma)$, $(\alpha^*) \in RV(\Sigma)$.

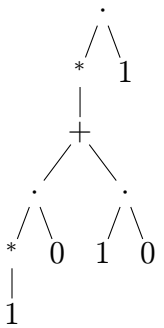
Reg. výraz $\alpha \in RV(\Sigma)$ reprezentuje jazyk v abecedě Σ, který označíme $[\alpha]$ a induktivně definujeme takto: $[\emptyset] = \emptyset$, $[\epsilon] = \{\varepsilon\}$, $[a] = \{a\}$, $[(\alpha + \beta)] = [\alpha] \cup [\beta]$, $[(\alpha \cdot \beta)] = [\alpha] \cdot [\beta]$, $[(\alpha^*)] = [\alpha]^*$.
Konvence při psaní regulárních výrazů. Příkladem regulárního výrazu je

$$((((0 \cdot 1)^* \cdot 1) \cdot (1 \cdot 1)) + ((0 \cdot 0) + 1)^*).$$

Domluvíme se na následujících možných zjednodušeních zápisů:

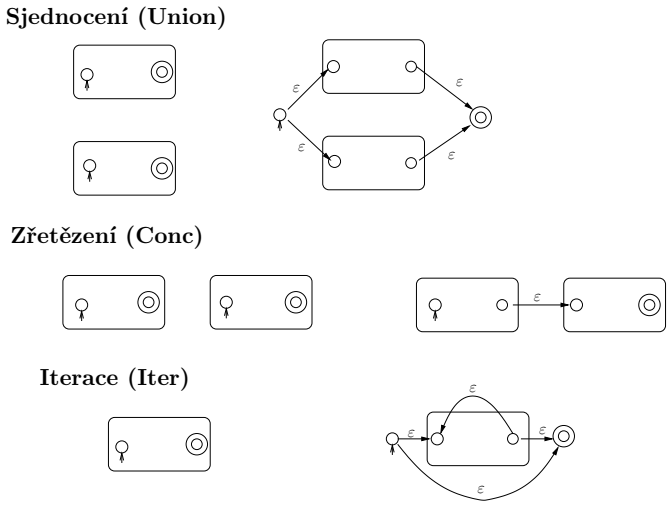
- vynecháváme vnější závorky;
- můžeme vynechat znak $\cdot$ pro zřetězení;
- využíváme asociativitu operace zřetězení pro vynechání závorek;
- další vynechání závorek umožňuje dohodnutá priorita operátorů:
  operátor * má vyšší prioritu než $\cdot$ a dále $\cdot$ má vyšší prioritu než $+$.

Uvedený příklad regulárního výrazu lze tedy ekvivalentně psát $(01)^*111 + (00 + 1)^*$.
Syntaktický strom. Závorky nám slouží k tomu, aby každému reg. výrazu byl jednoznačně přiřazen jeho syntaktický strom; naše zjednodušení zápisu tuto vlastnost zachovává. Definici syntaktického stromu necháme na cvičení; zde jako příklad uvedeme syntatický strom výrazu $(1^*0 + 10)^*1$.



Konstrukce ZNKA k regulárnímu výrazu

Na základě synt. stromu výrazu α lze snadno zkonstruovat ZNKA přijímající jazyk $[\alpha]$. Podstatu konstrukcí zachycuje následující obrázek.



Problém algoritmické konstrukce syntaktického stromu k danému regulárnímu výrazu je jednou z motivací pojmů, ke kterým se postupně dostaneme (speciálně jde o bezkontextové gramatiky a zásobníkové automaty).

Bezkontextové gramatiky

Připomeňme, že

$$(((a \cdot a) + b)^*)$$

je příklad (úplně uzávorkovaného) regulárního výrazu, který reprezentuje jazyk v abecedě $\{a, b\}$. Obecný regulární výraz nad abecedou $\{a, b\}$ je prostě řetězcem symbolů abecedy

$$\Sigma = \{a, b, +, \cdot, ^*, (,)\}\,.$$

(Pro úplnost bychom měli přidat znaky $\emptyset, \epsilon$, ale zde se bez nich obejdeme.)
Ne každý řetězec v Σ^* je ovšem regulárním výrazem; např. řetězec "$a) + +($" regulárním výrazem není.
Snadno jsme vyvodili, že množina těchto regulárních výrazů, označovaná $RV(\{a, b\})$, není regulárním jazykem (tedy není rozpoznatelná konečným automatem). Dá se ale generovat bezkontextovou gramatikou, např.

$$R \longrightarrow\ a \mid b \mid (R + R) \mid (R \cdot R) \mid (R^*)\,.$$

Demonstrovali jsme si základní pojmy teorie bezkontextových gramatik. Ukázali jsme si *(levou) derivaci* slova $(((a \cdot a) + b)^*)$, příslušný *derivační strom*, apod.

Připomněli jsme definici *bezkontextové gramatiky* jako struktury

$$G = (\Pi, \Sigma, S, P)$$

a *jazyka generovaného gramatikou*

$$L(G) = \{w \in \Sigma^* \mid S \Rightarrow^* w\}\,.$$

Pak jsme se vrátili k jazyku $RV(\{a, b\})$ a upravili jej tak, že v příslušných řetězcích (regulárních výrazech) můžeme vynechávat tečku pro zřetězení a některé závorky podle naší výše uvedené dohody. Např. výraz $(((a \cdot a) + b)^*)$ můžeme zjednodušit na $(aa + b)^*$. Takto upravený jazyk regulárních výrazů generuje např. gramatika

$$R \longrightarrow a \mid b \mid R + R \mid R \cdot R \mid RR \mid R^* \mid (R)\,.$$

Všimli jsme si ovšem, že např. slovo $aa + b$ má v této gramatice dva různé derivační stromy; tedy tato gramatika *není jednoznačná*. Příčinou je fakt, že naše dohodnutá priorita operátorů není v uvedené gramatice reflektována.
Po jistém zamyšlení se nám podařilo navrhnout ekvivalentní gramatiku (tedy generující tentýž jazyk), která jednoznačná je; konkrétně šlo o následující gramatiku

$$\begin{aligned} R &\longrightarrow T + R \mid T \\ T &\longrightarrow FT \mid F \\ F &\longrightarrow F^* \mid (R) \mid C \\ C &\longrightarrow a \mid b \end{aligned}$$

Uvedli jsme pojem *(vnitřně) jednoznačný bezkontextový jazyk* a několik souvisejících poznámek.

Partie textu k prostudování

Jedná se zejména o části 4.1., 4.2., 4.3. (bezkontextové gramatiky, jednoznačné gramatiky).
(Máte si udělat přinejmenším dobrou první představu a zamyslet se nad příklady, speciálně těmi plánovanými na cvičení, ať se můžete na cvičení aktivně účastnit a případné problémy si tam objasnit.)

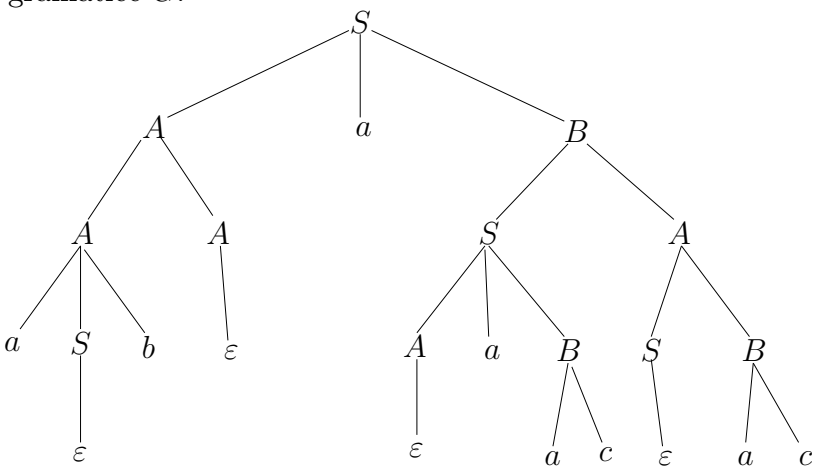
Cvičení

Příklad 4.1

Definujte (induktivně), co to je syntaktický strom regulárního výrazu.
(Pozor! Nepleťme si pojmy “syntaktický strom reg. výrazu” a “derivační strom podle gramatiky”.)

Příklad 4.2

Na obrázku je derivační strom pro slovo $w = abaaacac$ odpovídající jisté bezkontextové gramatice G.



- Vypište všechna pravidla G, jejichž existenci můžete vyvodit z daného derivačního stromu.
- Napište levé odvození (levou derivaci) slova w podle gramatiky G.
- Najděte *menší* derivační strom pro slovo $abaaacac$ a zakreslete jej tak, že všechny listy budou na stejné úrovni (tedy odvozené slovo bude celé na „jednom řádku“).
- Najděte nejlevější větev (v onom menším stromě), která obsahuje dva výskyty neterminálu B. Využijte to k důkazu, že gramatika generuje také slovo $abaac$. Pak ukažte, že gramatika také generuje slova $aba(a)ac(ac)$, $aba(a)^2ac(ac)^2$, $aba(a)^3ac(ac)^3$, ...
- Lze z dostupné informace zjistit něco ohledně jednoznačnosti gramatiky G ?

Příklad 4.3

Jaký jazyk generuje následující gramatika ? Porovnejte s gramatikou na přednášce. Je gramatika jednoznačná ?

$R \longrightarrow L \mid (RBR) \mid (RU)$
$L \longrightarrow a \mid b$
$B \longrightarrow + \mid \cdot$
$U \longrightarrow {}^*$

Příklad 4.4

Připomeňte si následující gramatiku z přednášky. Alespoň neformálně argumentujte, proč je gramatika jednoznačná.

$$\begin{aligned} R &\longrightarrow T + R \mid T \\ T &\longrightarrow FT \mid F \\ F &\longrightarrow F^* \mid (R) \mid C \\ C &\longrightarrow \ a \mid b \end{aligned}$$

(Nápověda. T (Term) reprezentuje ty regulární výrazy, které nejsou ve tvaru $R_1 + R_2$ (pro dva regulární výrazy R_1, R_2), a F (Factor) reprezentuje ty výrazy, které nejsou ve tvaru $R_1 + R_2$ ani $R_1 R_2$.)

Příklad 4.5

Uvažujme jazyk $L = \{ w \in \{a, b\}^* \mid |w| \geq 1 \text{ a } |w|_a = |w|_b \}$.
Charakterizujte slova z $L^2 = L \cdot L$. Je pravda, že $L = L^2$? Platí případně alespoň jedna z inkluzí $L \subseteq L^2$, $L^2 \subseteq L$?
Charakterizujte slova z $L - L^2$.
Na základě předešlých úvah navrhněte bezkontextovou gramatiku generující L.

Příklad 4.6

Snažte se co nejvýstižněji charakterizovat jazyk generovaný gramatikou $S \longrightarrow bSS \mid a$. (Možná vám pomůže chápat a jako "atomický výraz" a b jako "binární operátor".)

Příklad 4.7

(V případě nedostatku času dokončit příště.)
Navrhněte bezkontextové gramatiky generující následující jazyky:

- $L_1 = \{ w \in \{a, b\}^* \mid w \text{ obsahuje podslovo } baab \}$
- $L_2 = \{ w \in \{a, b\}^* \mid |w|_b \bmod 3 = 0 \}$
- $L_3 = \{ ww^R \mid w \in \{a, b\}^* \}$
- $L_4 = \{ 0^n 1^m 0^n \mid m, n \geq 0 \}$
- $L_5 = \{ 0^n 1^m \mid 1 \leq n \leq m \leq 2n \}$