Cvičenie č. 7
Regulárne jazyky

Peter Kostolányi

2. novembra 2022

Deterministické konečné automaty, nedeterministické konečné automaty a regulárne gramatiky sú – ako už dobre vieme – rovnako silné modely opisujúce rovnakú triedu jazykov. Jazyky z tejto triedy nazývame *regulárnymi*. Na definíciu regulárnych jazykov tak môžeme použiť ľubovoľný z uvedených troch ekvivalentných modelov.

Definícia 1. *Regulárny jazyk* je jazyk L, pre ktorý existuje deterministický konečný automat A taký, že $L(A) = L$. Triedu všetkých regulárnych jazykov označujeme $\mathscr{R}$.

Veta 1. *Nech L je jazyk. Nasledujúce tvrdenia sú ekvivalentné:*

(*i*) *Jazyk L je regulárny.*

(*ii*) *Existuje nedeterministický konečný automat A taký, že $L(A) = L$.*

(*iii*) *Existuje regulárna gramatika G taká, že $L(G) = L$.*

1 Pumpovacia lema pre regulárne jazyky

Nech $L \subseteq \Sigma^*$ je *ľubovoľný* regulárny jazyk. Určite potom *existuje* deterministický konečný automat $A = (K, \Sigma, \delta, q_0, F)$ taký, že $L(A) = L$. Ten má nejaký počet stavov $|K| =: p$. Nech $w \in L$ je *ľubovoľné* slovo z jazyka L také, že $|w| \geq p$. Z Dirichletovho princípu vyplýva, že počas výpočtu automatu A na tomto slove sa niektorý jeho stav q_{op} musí „vyskytnúť“ aspoň dvakrát. *Existujú* teda slová $u, v \in \Sigma^*$ a $x \in \Sigma^+$ také, že $w = uxv$ a

$$(q_0, uxv) \vdash^* (q_{op}, xv) \vdash^+ (q_{op}, v) \vdash^* (q, \varepsilon),$$

kde $q \in F$ je akceptačný stav. Nech teraz $i \in \mathbb{N}$ je ľubovoľné prirodzené číslo. Ľahko vidieť, že existuje akceptačný výpočet automatu A na slove ux^iv: stačí namiesto jedného „cyklu“ $(q_{op}, x) \vdash^+ (q_{op}, \varepsilon)$ vykonať takýchto „cyklov“ niekoľko. Inými slovami:

$$(q_0, ux^iv) \vdash^* (q_{op}, x^iv) \vdash^+ (q_{op}, x^{i-1}v) \vdash^+ \ldots \vdash^+ (q_{op}, v) \vdash^* (q, \varepsilon).$$

Pre $i = 0$ zodpovedá uvedený zápis „vynechaniu cyklu“. Je navyše zrejmé, že prvý opakovaný výskyt nejakého stavu musí prísť po najviac p krokoch výpočtu, takže možno pridať obmedzenie $|ux| \leq p$.

Pomocou takýchto jednoduchých úvah možno odvodiť pumpovaciu lemu pre regulárne jazyky, pričom správne poradie kvantifikátorov je dané poradím slov, ktoré sme zvýrazňovali italikou.

Veta 2. *Nech Σ je abeceda a $L \subseteq \Sigma^*$ regulárny jazyk. Potom existuje $p \in \mathbb{N}$ také, že pre všetky $w \in L$ s $|w| \geq p$ existujú slová $u, x, v \in \Sigma^*$ také, že:*

(*i*) $w = uxv$,

(*ii*) $|ux| \leq p$,

(*iii*) $|x| \geq 1$,

(*iv*) $\forall i \in \mathbb{N} : ux^iv \in L$.

Namiesto memorovania znenia pumpovacej lemy je lepšie túto vetu vždy „na počkanie“ odvodiť. Priestor na chyby (predovšetkým v poradí kvantifikátorov) je v takom prípade omnoho menší.

1.1 Riešené úlohy

Keďže je pumpovacia lema sformulovaná ako tvrdenie platné pre každý regulárny jazyk, platnosť podmienok z nej vyplývajúcich je pre jazyk L *nutnou podmienkou* jeho regulárnosti. Ak teda nejaký jazyk nespĺňa podmienky vyplývajúce z pumpovacej lemy, nemôže byť regulárny. Táto skutočnosť je základom metódy dokazovania negatívnych výsledkov o regulárnosti jazykov, ktorú teraz demonštrujeme na príklade jazyka všetkých palindrómov nad abecedou $\{a, b\}$.

Úloha 1. Dokážte, že jazyk $L = \{w \in \{a,b\}^* \mid w = w^R\}$ nie je regulárny.

Riešenie. Sporom, nech $L \in \mathscr{R}$. Jazyku L potom podľa pumpovacej lemy prislúcha číslo $p \in \mathbb{N}$. Vezmime slovo $w = a^p b a^p$ – zjavne $w \in L$ a $|w| \geq p$. Preto existujú slová $u, x, v \in \{a,b\}^*$, pre ktoré sú splnené podmienky (i) až (iv) pumpovacej lemy.

Podľa (i) je $w = uxv$. Z podmienok (ii) a (iii) teda vyplýva, že existujú $r, s \in \mathbb{N}$ také, že $s \geq 1$, $u = a^r$, $x = a^s$, a $v = a^{p-s-r}ba^p$.

Z podmienky (iv) pumpovacej lemy pre $i = 2$ napokon vyplýva, že slovo

$$ux^2v = a^r a^{2s} a^{p-s-r} b a^p = a^{p+s} b a^p$$

patrí do jazyka L. To je ale spor, pretože $s \geq 1$ a slovo $a^{p+s}ba^p$ teda nie je palindróm. □

Pri nasledujúcej „písomkovej“ úlohe uvádzame okrem jej správneho riešenia aj nesprávne riešenie, s ktorého variáciami prichádzali menej úspešní študenti pomerne často.

Úloha 2. Zistite, či je jazyk $L = \{a^{2^n} \mid n \in \mathbb{N}\}$ regulárny. Svoje tvrdenie dokážte.

Nesprávne *riešenie.* Dokážeme, že jazyk L nie je regulárny. Sporom, nech $L \in \mathscr{R}$. Nech $p \in \mathbb{N}$ je konštanta zodpovedajúca jazyku L podľa pumpovacej lemy. Bez ujmy na všeobecnosti môžeme predpokladať, že $p \geq 1$. Vezmime slovo $w = a^{2^p}$. Potom $w = uxv$, kde $u = a$, $x = a$ a $v = a^{2^p-2}$. Nech $i = 2$. Potom z podmienky (iv) pumpovacej lemy vyplýva $ux^2v = a^{2^p+1} \in L$, čo je spor, pretože $2^p + 1$ nemôže byť mocninou čísla 2.

Toto riešenie **je chybné**, pretože pumpovacia lema nám nedovoľuje zvoliť si slová u, x, v úplne ľubovoľne. Jediné, čo nám dovoľuje o týchto slovách predpokladať, sú podmienky (i) až (iii). □

Správne *riešenie.* Dokážeme, že jazyk L nie je regulárny. Sporom, nech $L \in \mathscr{R}$. Nech $p \in \mathbb{N}$ je konštanta zodpovedajúca jazyku L podľa pumpovacej lemy. Vezmime teraz ľubovoľné slovo $w = a^{2^m}$ také, že $2^m > p$ – očividne $w \in L$ a $|w| \geq p$. Potom existujú slová u, x, v, pre ktoré sú splnené podmienky (i) až (iv) pumpovacej lemy.

Z podmienky (i) máme $w = uxv$. Z podmienok (ii) a (iii) vyplýva, že existujú čísla $r, s \in \mathbb{N}$ také, že $s \geq 1$, $r + s \leq p$, $u = a^r$, $x = a^s$ a $v = a^{2^m-r-s}$. Z podmienky (iv) pumpovacej lemy potom vyplýva $ux^2v = a^{2^m+s} \in L$. Keďže ale $s \geq 1$ a $s \leq r + s \leq p < 2^m$, je $2^m < 2^m + s < 2^{m+1}$, a teda $2^m + s$ nie je mocnina dvoch, čo je spor. □

Poznámka 1. Vyvarovať sa chýb podobných tej vyššie je možné len pri správnom pochopení toho, čo sa pri dôkazoch s použitím pumpovacej lemy deje.

Typický takýto dôkaz je dôkazom sporom – za účelom sporu predpokladáme, že jazyk $L \subseteq \Sigma^*$, o ktorom chceme ukázať, že nie je regulárny, regulárny je. Pumpovacia lema je zárukou, že za tohto predpokladu platí pre L určité tvrdenie. K sporu prídeme tak, že dokážeme negáciu tohto tvrdenia – teda, že *pre všetky* $p \in \mathbb{N}$ *existuje* $w \in L$ s $|w| \geq p$ také, že *pre žiadne* $u, x, v \in \Sigma^*$ nemôže súčasne platiť (i) až (iv).

Kým teda s konštantou p musíme pracovať *vo všeobecnosti*, slovo w si môžeme zvoliť ako *ľubovoľné* slovo z jazyka L dĺžky aspoň p. Slová u, x, v ale opäť musíme uvažovať vo všeobecnosti – potrebujeme totiž ukázať, že pre *žiadne* prípustné u, x, v nemôžu byť súčasne splnené podmienky (i) až (iv). To väčšinou dokazujeme tak, že predpokladáme platnosť podmienok (i) až (iii) a dokazujeme, že neplatí (iv). Slová u, x, v teda predsa len nemusíme uvažovať úplne ľubovoľné – môžeme predpokladať, že sú pre ne splnené podmienky (i) až (iii). Nič viac ale predpokladať nemôžeme.

2 Uzáverové vlastnosti triedy regulárnych jazykov

Trieda jazykov $\mathscr{L}$ je uzavretá na nejakú k-árnu operáciu Φ, ak $\Phi(L_1, \ldots, L_k) \in \mathscr{L}$ kedykoľvek $L_1, \ldots, L_k \in \mathscr{L}$. Napríklad trieda regulárnych jazykov $\mathscr{R}$ je teda uzavretá na zjednotenie, pretože pre všetky $L_1, L_2 \in \mathscr{R}$ je $L_1 \cup L_2 \in \mathscr{R}$.

Dokázať uzavretosť triedy regulárnych jazykov na k-árnu operáciu Φ teda znamená pre všetky regulárne jazyky $L_1, \ldots, L_k$ ukázať, že aj jazyk $\Phi(L_1, \ldots, L_k)$ musí byť regulárny. To možno urobiť napríklad konštrukciou konečného automatu alebo regulárnej gramatiky pre jazyk $\Phi(L_1, \ldots, L_k)$, pričom pri konštrukcii vychádzame z konečných automatov alebo regulárnych gramatík, ktorých existenciu predpokladáme pre jazyky $L_1, \ldots, L_k$.

Vyvrátiť uzavretosť triedy $\mathscr{R}$ na operáciu Φ naopak znamená nájsť konkrétny príklad jazykov $L_1, \ldots, L_k \in \mathscr{R}$ takých, že $\Phi(L_1, \ldots, L_k) \notin \mathscr{R}$.

2.1 Riešené úlohy

Pripomeňme si najprv definíciu operácie „*shuffle*“ na jazykoch – ďalšie podrobnosti možno nájsť v sade úloh na tretie cvičenie.

Definícia 2. Nech Σ je abeceda a $L_1, L_2 \subseteq \Sigma^*$ sú jazyky. Jazyk $L_1 \sqcup\!\sqcup L_2$ potom definujeme ako

$$L_1 \sqcup\!\sqcup L_2 = \{u_1 v_1 u_2 v_2 \ldots u_n v_n \mid n \in \mathbb{N};\ u_1, \ldots, u_n, v_1, \ldots, v_n \in \Sigma^*;\ u_1 \ldots u_n \in L_1;\ v_1 \ldots v_n \in L_2\}.$$

Do jazyka $L_1 \sqcup\!\sqcup L_2$ teda patria všetky slová, ktoré vzniknú z nejakého slova $u \in L_1$ a slova $v \in L_2$ ich „premiešaním“ zachovávajúcim relatívne poradie písmen v oboch slovách. Nejakým ľubovoľným spôsobom tieto dve slová vyjadríme ako zreťazenia $n \in \mathbb{N}$ faktorov – položíme teda $u = u_1 u_2 \ldots u_n$ a $v = v_1 v_2 \ldots v_n$, kde $u_1, \ldots, u_n \in \Sigma^*$ a $v_1, \ldots, v_n \in \Sigma^*$ sú *slová*. Do jazyka $L_1 \sqcup\!\sqcup L_2$ potom bude patriť slovo $u_1 v_1 u_2 v_2 \ldots u_n v_n$. Faktory $u_1, \ldots, u_n$ a $v_1, \ldots, v_n$ tu môžu byť aj prázdne, čo okrem iného znamená, že slovo z jazyka $L_1 \sqcup\!\sqcup L_2$ sa môže začínať aj faktorom slova z L_2.

Príklad 1. Nech $L_1 = \{aa, ba\}$ a $L_2 = \{bb\}$. Potom

$$L_1 = \{aabb, abab, abba, baab, baba, bbaa, babb, bbab, bbba\}.$$

Úloha 3. Zistite, či je trieda $\mathscr{R}$ uzavretá na operáciu „shuffle“. Svoje tvrdenie dokážte.

Riešenie. Dokážeme, že trieda $\mathscr{R}$ *je uzavretá* na túto operáciu. Nech L_1, L_2 sú regulárne jazyky a nech Σ je ľubovoľná abeceda taká, že $L_1 \subseteq \Sigma^*$ a zároveň $L_2 \subseteq \Sigma^*$.[1] Nech $A_1 = (K_1, \Sigma, \delta_1, q_{0,1}, F_1)$ a $A_2 = (K_2, \Sigma, \delta_2, q_{0,2}, F_2)$ sú deterministické konečné automaty také, že $L(A_1) = L_1$ a $L(A_2) = L_2$. Zostrojíme *nedeterministický* konečný automat $A = (K, \Sigma, \delta, q_0, F)$ taký, že $L(A) = L_1 \sqcup\!\sqcup L_2$.

Automat A dostane slovo $w = a_1 a_2 \ldots a_m$ pre nejaké $m \in \mathbb{N}$ a písmená $a_1, \ldots, a_m \in \Sigma$. Toto slovo má akceptovať práve vtedy, keď $w \in L_1 \sqcup\!\sqcup L_2$ – to znamená, keď možno indexovú množinu $[m]$ rozložiť na dve disjunktné podmnožiny $I = \{i_1, \ldots, i_s\}$ a $J = \{j_1, \ldots, j_t\}$ s $i_1 < \ldots < i_s$ a $j_1 < \ldots < j_t$ tak, že $a_{i_1} \ldots a_{i_s} \in L_1$ a $a_{j_1} \ldots a_{j_t} \in L_2$. Každý výpočet automatu A na slove w bude zodpovedať nejakému takémuto „rozdeleniu písmen“ tvoriacich slovo w podľa toho, či prislúchajú slovu z L_1 alebo slovu z L_2; naopak pre každé takéto prípustné „rozdelenie písmen“ bude v automate A existovať jeden výpočet na slove w. Na písmenách „priradených“ jazyku L_1 bude automat A simulovať príslušný výpočet automatu A_1 a na zvyšných písmenách bude simulovať výpočet automatu A_2. Automat A napokon slovo w akceptuje práve vtedy, keď po jeho dočítaní budú pre aspoň jedno „rozdelenie písmen“ obidva simulované automaty A_1 a A_2 v akceptačnom stave.

Stavmi automatu A teda budú dvojice $[p, q]$, kde p je stav automatu A_1 a q je stav automatu A_2. Na každé písmeno c sa automat A bude môcť zo stavu $[p, q]$ pohnúť dvoma spôsobmi: odsimulovaním kroku výpočtu automatu A_1 prechodom do stavu $[\delta_1(p, c), q]$ – to zodpovedá prípadu, keď je príslušný výskyt písmena c „priradený“ jazyku L_1 – a odsimulovaním kroku výpočtu automatu A_2 prechodom do stavu $[p, \delta_2(q, c)]$ – čo zodpovedá prípadu, keď je výskyt písmena c „priradený“ jazyku L_2.

[1] Možno napríklad vziať $\Sigma = \Sigma_{L_1} \cup \Sigma_{L_2}$.

Formálne teda automat $A = (K, \Sigma, \delta, q_0, F)$ skonštruujeme nasledovne: $K = K_1 \times K_2$,

$$\delta([p,q],c) = \{[\delta_1(p,c),q],[p,\delta_2(q,c)]\} \text{ pre všetky } p \in K_1,\ q_2 \in K_2 \text{ a } c \in \Sigma,$$

$\delta([p,q],\varepsilon) = \emptyset$ pre všetky $p \in K_1$ a $q \in K_2$, $q_0 = [q_{0,1}, q_{0,2}]$ a $F = F_1 \times F_2$.

Indukciou by sme ľahko dokázali, že pre $p \in K_1$, $q \in K_2$ a $w \in \Sigma^*$ je $([q_{0,1}, q_{0,2}], w) \vdash_A^* ([p,q],\varepsilon)$ práve vtedy, keď existuje $n \in \mathbb{N}$ a slová $u_1, \ldots, u_n, v_1, \ldots, v_n \in \Sigma^*$ také, že $w = u_1 v_1 u_2 v_2 \ldots u_n v_n$, pričom $(q_{0,1}, u_1 \ldots u_n) \vdash_{A_1}^* (p,\varepsilon)$ a $(q_{0,2}, v_1 \ldots v_n) \vdash_{A_2}^* (q,\varepsilon)$. Jednoduchým dôsledkom tohto pozorovania je, že $w \in L(A)$ – čiže $([q_{0,1}, q_{0,2}], w) \vdash_A^* ([p,q],\varepsilon)$ pre $[p,q] \in F = F_1 \times F_2$ – práve vtedy, keď $w \in L(A_1) \sqcup L(A_2) = L_1 \sqcup L_2$, z čoho vyplýva správnosť našej konštrukcie. □

Definícia 3. Nech Σ je abeceda a $L \subseteq \Sigma^*$ je jazyk. *Rotáciou* jazyka L nazveme jazyk

$$\mathrm{rot}(L) = \{vu \mid u, v \in \Sigma^*;\ uv \in L\}.$$

Úloha 4. Zistite, či je trieda $\mathscr{R}$ uzavretá na rotáciu. Svoje tvrdenie dokážte.

Riešenie. Dokážeme, že trieda $\mathscr{R}$ *je uzavretá* na rotáciu. Nech $L \in \mathscr{R}$ je regulárny jazyk. Potom existuje deterministický konečný automat A taký, že $L(A) = L$. Zostrojíme *nedeterministický* konečný automat A' taký, že $L(A') = \mathrm{rot}(L)$.

Automat A' dostane vstup w, ktorý má akceptovať v prípade, že existujú slová u, v, pre ktoré platí $w = vu$ a zároveň $uv \in L$. Ak $uv \in L$, musí existovať akceptačný výpočet automatu A na slove uv – tento výpočet bude „základom“ pre výpočet automatu A'. Automat A' teda najprv „nedeterministicky uhádne“ stav q, v ktorom automat A dočíta slovo u – správnosť tohto „tipu“ overí na konci výpočtu. Následne spustí simuláciu automatu A na vstupe w – ktoré má v prípade akceptácie zodpovedať slovu vu – so začiatkom simulácie v stave q. Automat A' bude simulovať automat A, až kým sa „nedeterministicky rozhodne“, že prečítal slovo v.

Ak táto simulácia skončila v neakceptačnom stave automatu A, buď na začiatku nešlo o správny „nedeterministický tip“ stavu q, alebo automat A slovo uv neakceptuje. V oboch prípadoch automat A' svoj vstup zamietne. Ak naopak simulácia skončila v akceptačnom stave, automat A' overí svoj „nedeterministický tip“ stavu q. To znamená, že na zvyšku svojho vstupu – ten má zodpovedať slovu u – spustí simuláciu automatu A so začiatkom simulácie v stave q_0 a overí, či po jeho dočítaní bude automat A naozaj v stave q. Ak áno, automat A' svoj vstup akceptuje.

Nech teda $A = (K, \Sigma, \delta, q_0, F)$. Potom $A' = (K', \Sigma', \delta', q_0', F')$, kde $K' = \{q_0'\} \cup K \times \{1,2\} \times K$, $\Sigma' = \Sigma$,

$$\begin{aligned}
\delta'(q_0,\varepsilon) &= \{[q,1,q] \mid q \in K\},\\
\forall p,q \in K\ \forall c \in \Sigma : \delta'([p,1,q],c) &= \{[\delta(p,c),1,q]\},\\
\forall p \in F\ \forall q \in K : \delta'([p,1,q],\varepsilon) &= \{[q_0,2,q]\},\\
\forall p,q \in K\ \forall c \in \Sigma : \delta'([p,2,q],c) &= \{[\delta(p,c),2,q]\},
\end{aligned}$$

ostatné výstupy prechodovej funkcie sú prázdne množiny a množina akceptačných stavov F' je

$$F' = \{[q,2,q] \mid q \in K\}.$$

Dôkaz správnosti tejto konštrukcie je možné založiť na nasledujúcich invariantoch: pre všetky $p, q \in K$ a $v \in \Sigma^*$ je $(q_0', v) \vdash_{A'}^* ([p,1,q],\varepsilon)$ práve vtedy, keď $(q,v) \vdash_A^* (p,\varepsilon)$; ďalej $(q_0', w) \vdash_{A'}^* ([p,2,q],\varepsilon)$ práve vtedy, keď existujú slová $u, v \in \Sigma^*$ také, že $w = vu$ a pre nejaký akceptačný stav $q_F \in F$ je $(q,v) \vdash_A^* (q_F,\varepsilon)$ a $(q_0,u) \vdash_A^* (p,\varepsilon)$. Z toho vyplýva, že $w \in L(A')$ – čiže $(q_0', w) \vdash_{A'}^* ([q,2,q],\varepsilon)$ pre nejaké $q \in K$ – práve vtedy, keď existujú slová $u, v \in \Sigma^*$ také, že $w = vu$ a pre nejaký akceptačný stav $q_F \in F$ je $(q_0,u) \vdash_A^* (q,\varepsilon)$ a $(q,v) \vdash_A^* (q_F,\varepsilon)$, t. j. $uv \in L(A) = L$. To znamená, že naozaj $w \in L(A')$ práve vtedy, keď $w \in \mathrm{rot}(L)$. □

Úloha 5. Nech Σ je abeceda a $L \subseteq \Sigma^*$ je jazyk. Položme

$$\mathrm{pal}(L) := \{w \in L \mid w = w^R\}.$$

Zistite, či je trieda $\mathscr{R}$ uzavretá na operáciu pal. Svoje tvrdenie dokážte.

Riešenie. Dokážeme, že trieda $\mathscr{R}$ *nie je uzavretá* na túto operáciu. Uvažujme napríklad regulárny jazyk $L = \{a, b\}^*$. Potom zrejme $\mathrm{pal}(L) = \{w \in \{a, b\}^* \mid w = w^R\}$, čo je jazyk, o ktorom z úlohy 1 vieme, že nie je regulárny. □