4. Další vlastnosti regulárních jazyků a pumping lemma

Formální jazyky a automaty

Jiří Balun

Obsah

Motivace

Na minulé přednášce

- rozšířili jsme NFA o ε-přechody
- zavedli jsme nový formalismus, takzvané **regulární výrazy**
  - pomocí nich jsme ukázali, že regulární jazyky jsou uzavřené na zřetězení a Kleeneho uzávěr
- ukázali jsme ekvivalenci všech dosavadních modelů

Na této přednášce

- dokončíme uzávěrové operace regulárních jazyků
- ukážeme si příklady neregulárních jazyků a operací, které nejsou uzávěrové
- představíme si **pumping lemma**
  - nástroj, pomocí kterého lze elegantně ukázat, že některé jazyky nejsou regulární

Homomorfismus

Intuice: každý symbol jedné abecedy přepíšeme na nějaký řetězec nad druhou abecedou.

Definice

Homomorfismus na abecedě Σ_1 je funkce $h\colon \Sigma_1 \to \Sigma_2^*$, která přiřazuje každému symbolu ze Σ_1 řetězec nad abecedou Σ_2.

- pro homomorfismus $h\colon \Sigma_1 \to \Sigma_2^*$ definici rošíříme i pro řetězce a jazyky:
  - pro řetězec $w = w_1 \dots w_n \in \Sigma_1^*$ je $h(w) = h(w_1) \dots h(w_n)$ řetězec nad Σ_2
  - pro jazyk $L \subseteq \Sigma_1^*$ je $h(L) = \{y \in \Sigma_2^* \mid x \in L \wedge h(x) = y\}$ jazyk nad Σ_2

Příklad

Mějme abecedy $\Sigma_1 = \{a, b, c\}$ a $\Sigma_2 = \{0, 1\}$ a homomorfismus $h\colon \Sigma_1 \to \Sigma_2^*$, kde $h(a) = 00$, $h(b) = 1$ a $h(c) = 0$, pak:

- $h(abc) = h(a) \cdot h(b) \cdot h(c) = 00 \cdot 1 \cdot 0 = 0010$,
- pro $L = \{a^n b^n \mid n \in \mathbb{N}_0\}$ dostaneme $h(L) = \{0^{2n} 1^n \mid n \in \mathbb{N}_0\}$.

Uzavřenost na homomorfimus

Věta

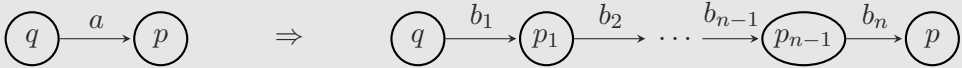
Nechť $L \subseteq \Sigma_1^*$ je regulární jazyk a $h\colon \Sigma_1 \to \Sigma_2^*$ je homomorfismus, pak jazyk $h(L) \subseteq \Sigma_2^*$ je také regulární.

Důkaz

Jazyk L je regulární, proto existuje nějaký DFA $A = (Q, \Sigma_1, \delta, q_0, F)$ takový, že $L(A) = L$.

Zkonstruujeme takový ε-NFA $A' = (Q \cup Q', \Sigma_2, \delta', q_0, F)$, že $L(A') = h(L)$:

- pro každý přechod $\delta(q, a) = p$, kde $h(a) = \varepsilon$, definujeme $\delta'(q, \varepsilon) = p$ v A',
- pro každý přechod $\delta(q, a) = p$, kde $h(a) = b \in \Sigma_2$, definujeme $\delta'(q, b) = p$ v A',
- pro každý přechod $\delta(q, a) = p$, kde $h(a) = b_1 \dots b_n \in \Sigma_2^*$, definujeme v A':
  - nové stavy $p_1, \dots, p_{n-1}$, které přidáme do Q',
  - přechody $\delta'(q, b_1) = p_1$, $\delta'(p_1, b_2) = p_2$, $\dots$, a nakonec $\delta'(p_{n-1}, b_n) = p$.



□

Inverzní homomorfismus

Intuice: obraz inverzního homomorfismu h^{-1} pro řetězec y je množina takových řetězců, které h zobrazí na y.

Definice

Pro homomorfismus $h\colon \Sigma_1 \to \Sigma_2^*$ definujeme **Inverzní homomorfismus** $h^{-1}\colon \Sigma_2^* \to 2^{\Sigma^*}$ definujeme jako $h^{-1}(y) = \{x \in \Sigma_1^* \mid h(x) = y\}$.

- pro homomorfismus $h\colon \Sigma_1 \to \Sigma_2^*$ definici opět rošíříme i pro jazyky:
  - pro jazyk $L \subseteq \Sigma_2^*$ je $h^{-1}(L) = \{x \in \Sigma_1^* \mid h(x) \in L\}$

Příklad

Mějme abecedy $\Sigma_1 = \{a, b, c\}$ a $\Sigma_2 = \{0, 1\}$ a homomorfismus $h\colon \Sigma_1 \to \Sigma_2^*$, kde $h(a) = 00$, $h(b) = 1$ a $h(c) = 0$, pak:

- $h^{-1}(0010) = L((\boldsymbol{a} + \boldsymbol{cc})\boldsymbol{bc}) = \{abc, ccbc\}$,
- pro $L = L(\mathbf{0}^*)$ dostaneme $h^{-1}(L) = L((\boldsymbol{a} + \boldsymbol{cc})^*)$.

Uzavřenost na inverzní homomorfimus

Věta

Nechť $L \subseteq \Sigma_2^*$ je regulární jazyk a $h\colon \Sigma_1 \to \Sigma_2^*$ je homomorfismus, pak jazyk $h^{-1}(L) \subseteq \Sigma_1^*$ je také regulární.

Důkaz

Jazyk L je regulární, proto existuje nějaký DFA $A = (Q, \Sigma_2, \delta, q_0, F)$ takový, že $L(A) = L$.

Zkonstruujeme takový DFA $A' = (Q, \Sigma_1, \delta', q_0, F)$, že $L(A') = h^{-1}(L)$:

- zde je stačí správně definovat δ' pro každý stav q a symbol a jako $\delta'(q, a) = \hat{\delta}(q, h(a))$.

Příklad pro $h(a) = b_1 b_2$ a přechody $\hat{\delta}(p, b_1 b_2) = \hat{\delta}(q, b_2) = r$:

□

Reverzní jazyk

Věta

Nechť L je regulární jazyk, pak L^R je také regulární jazyk.

Důkaz

Jazyk L je regulární, proto existuje nějaký DFA $A = (Q, \Sigma, \delta, q_0, F)$ takový, že $L(A) = L$.

Zkonstruujeme takový ε-NFA $A^R = (Q \cup \{q_0'\}, \Sigma, \delta^R, \{q_0'\}, \{q_0\})$, že $L(A') = L^R$:

- platí, že $L^R = \{w^R \mid \delta(q_0, w) \in F\}$,
- nejprve definujeme NFA $A' = (Q, \Sigma, \delta^R, F, \{q_0\})$ s množinou počátečních stavů F,
- obrátíme všechny přechody z δ, tj. definujeme $p \in \delta^R(q, a)$ pokud $\delta(p, a) = q$,
- z A' dostaneme A^R zredukováním počátečních stavů na jeden nový počáteční stav q_0', například pomocí ε-přechodů z q_0' do všech stavů v F. $\square$

Jazyk prefixů a sufixů

Věta

Nechť L je regulární jazyk, pak $L_{sfx} = \{x \mid x \in Sfx(y) \wedge y \in L\}$ je také regulární jazyk.

Důkaz

Jazyk L je regulární, proto existuje nějaký DFA $A = (Q, \Sigma, \delta, q_0, F)$ takový, že $L(A) = L$.

Zkonstruujeme takový ε-NFA $A' = (Q \cup \{q_0'\}, \Sigma, \delta', q_0', F)$, že $L(A') = Sfx(L)$:

- platí, že $Sfx(L) = \{w \mid \delta(q, w) \in F$, kde q je libovolný stav $Q\}$,
- nejprve definujeme NFA $(Q, \Sigma, \delta', Q, F)$ s množinou počátečních stavů Q,
- v tomto automatu redukujeme počáteční stavy na jeden nový počáteční stav q_0', například pomocí ε-přechodů z q_0' do všech stavů v Q. □

- uzavřenost na prefixy lze dokázat díky vztahu $Pfx(L) = (Sfx(L^R))^R$
- podobně lze ukázat i uzavřenost na infixy

Souhrn vlastností regulárních jazyků

Uzávěrové vlastnosti

- klasické množinové operace: sjednocení, průnik a doplněk
- operace regulárních výrazů: zřetězení a Kleeneho uzávěr
- zobrazení pomocí homoromfismu a inverzního homoromfismu
- reverzní jazyk
- jazyk prefixů, sufixů a infixů

Rozhodnutelné vlastnosti

- neprázdnost jazyka
- rovnost a inkluze regulárních jazyků
- minimalita reprezentace
  - lze rozhodnout, zda daný DFA je nejmenší možný (příště)

Neregulární jazyky

- už jsme si ukázali příklady neregulárních jazyků (ale zatím bez důkazu):
  - $\{a^n b^n \mid n \in \mathbb{N}_0\}$,
  - $\{a^p \mid p$ je prvočíslo$\}$,
  - a tak dále...
- nejprve formálně dokážeme, že $\{a^n b^n \mid n \in \mathbb{N}_0\}$ není regulární
  - to nám pomůže odhalit vlastnosti, pro které regulární jazyky nejsou uzavřené
- představíme si postup, jak odhalit většinu neregulárních jazyků
  - protože mít speciální důkaz pro každý neregulární jazyk je nepraktické
  - využijeme toho, že regulární jazyky jsou rozpoznávány automaty, a tedy mají v sobě ukrytou nějakou strukturu
  - toho využijeme k formulaci pumping lemma

Neregulární jazyk $a^n b^n$

Věta

Jazyk $L = \{a^n b^n \mid n \in \mathbb{N}_0\}$ není regulární.

Důkaz

Důkaz vedeme sporem. Nechť je L regulární, pak existuje DFA $A = (Q, \Sigma, \delta, q_0, F)$, který jej rozpoznává, tj. $L(A) = L$.

- Mějme stavy q_i a f_i takové, že $\delta(q_0, a^i) = q_i$ a zároveň $\delta(q_i, b^i) = f_i \in F$, pro nějaké $i \in \mathbb{N}_0$ (to že existují vyplývá z $a^i b^i \in L$)
- nechť $j \in \mathbb{N}_0$ takové, že $j \neq i$, pak existují stavy $\delta(q_0, a^j) = q_j$ a $\delta(q_j, b^j) = f_j \in F$,
- nejprve uvažme, že q_i a q_j jsou jeden a tentýž stav, tj. $q_i = q_j$, pak platí:

$$\delta(q_0, a^j b^j) = \delta(q_j, b^j) = f_j = \delta(q_0, a^i b^j) = \delta(q_j, b^j)$$

  z čehož by vyplývalo $a^i b^j \in L(A)$, což je v rozporu s předpokladem $L(A) = L$,
- proto pro libovolné různé i a j platí: $\delta(q_0, a^i) = q_i \neq q_j = \delta(q_0, a^j)$,
- potom by A musel mít nekonečně mnoho stavů, což je v rozporu s definicí DFA. □

Nekonečné sjednocení jazyků

Intuice: sjednocení konečného počtu regulárních jazyků lze rozpoznat pomocí součinového automatu. Ten by ale pro nekonečný počet sjednocení musel mít nekonečný počet stavů.

Věta

Třída regulárních jazyků není uzavřená na sjednocení nekonečného počtu regulárních jazyků.

Důkaz

Sjednocením nekonečného počtu regulárních jazyků sestrojíme neregulární jazyk L:

$$L = \{a^0b^0\} \cup \{a^1b^1\} \cup \{a^2b^2\} \cup \ldots = \bigcup_{i=0}^{\infty} \{a^ib^i\} = \{a^nb^n \mid n \in \mathbb{N}_0\}$$

kde každý podjazyk $\{a^ib^i\}$ je regulární (obsahuje jen jedno slovo, a proto je konečný), ale jejich sjednocením dostaneme jazyk $L = \{a^nb^n \mid n \in \mathbb{N}_0\}$, který není regulární. $\square$

Nekonečný průnik jazyků

Věta

Třída regulárních jazyků není uzavřená na průnik nekonečného počtu regulárních jazyků.

Důkaz

Průnikem nekonečného počtu regulárních jazyků sestrojíme neregulární jazyk L:

$$L = \overline{\{a^0b^0\}} \cap \overline{\{a^1b^1\}} \cap \ldots = \bigcap_{i=0}^{\infty} \overline{\{a^ib^i\}} = \overline{\{a^nb^n \mid n \in \mathbb{N}_0\}} = \Sigma^* \setminus \{a^nb^n \mid n \in \mathbb{N}_0\}$$

kde každý podjazyk $\{a^ib^i\}$ je regulární a tedy i jeho doplněk $\overline{\{a^ib^i\}}$ je regulární, ale:

- jejich průnikem dostaneme jazyk $L = \overline{\{a^nb^n \mid n \in \mathbb{N}_0\}}$, který regulární není,
- což vyplývá z toho, že není regulární jeho doplněk $\overline{L} = \{a^nb^n \mid n \in \mathbb{N}_0\}$. □

- dále lze ukázat, že regulární jazyky nejsou uzavřené na podmožinovost
  - například $\{a^nb^n \mid n \in \mathbb{N}_0\} \subseteq L(\boldsymbol{a^*b^*})$

Pumping lemma (PL)

Intuice: každé dostatečně dlouhé slovo regulárního jazyka obsahuje nějaký vzor, který lze z tohoto slova odstranit, nebo jej můžeme na daném místě libovolně opakovat, a pořád dostaneme slovo téhož jazyka.

Lemma

Nechť L je regulární jazyk, pak existuje konstanta $n \in \mathbb{N}$ taková, že každý řetězec $w \in L$ délky aspoň n lze rozdělit na tři části $w = xyz$, které splňují:

1. $|y| > 0$,
2. $|xy| \leq n$,
3. $xy^iz \in L$ pro všechna $i \geq 0$.

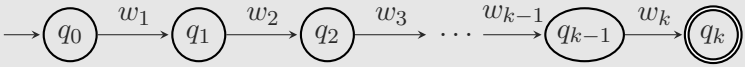
- pokud je jazyk L konečný (a tedy i regulární), pak lemma platí triviálně
  - konstantu n zvolíme větší než je délka nejdelšího řetězce v L
- tvrzení je tvaru implikace: pokud je L regulární pak musí něco splňovat...
  - pokud jazyk není regulární, pak z pro něj z lemma nic konkrétního nevyplývá
  - kontrapozicí dostaneme: pokud L nesplňuje podmínky z PL, pak L není regulární
  - existuje neregulární jazyk, který splňuje všechny podmínky z lemma

Důkaz PL

Důkaz

Jazyk L je regulární, proto existuje nějaký DFA $A = (Q, \Sigma, \delta, q_0, F)$ takový, že $L(A) = L$.

- Zvolíme konstantu $n = |Q|$ jako počet stavů A,
- pro každé $w = w_1 \dots w_k \in L$ takové, že $|w| \geq n$ najdeme rozdělení $w = xyz$ tak, aby platily podmínky 1, 2 a 3 z PL,
- nechť $\langle q_0, w_1 \dots w_k \rangle$, $\langle q_1, w_2 \dots w_k \rangle, \dots, \langle q_k, \varepsilon \rangle$ je posloupnost konfigurací reprezentující výpočet A na slově w, který postupně prochází stavy $q_0, \dots, q_k$:

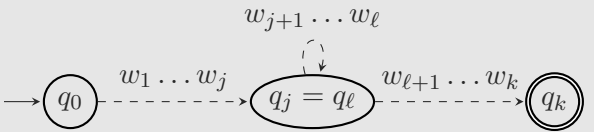


- v této posloupnosti se alespoň jeden stav musí vyskytovat vícekrát:
  - w má nejméně tolik znaků, kolik je stavů A a q_0 je navštívený hned na začátku výpočtu,
  - používáme tzv. *pigeonhole principle* – pokud máme více holubů než přihrádek, pak alespoň v jedné přihrádce musí být více jak jeden holub,

Důkaz – pokračování

- nechť v této posloupnosti $q_j = q_\ell$, kde $j < \ell$, je první stav, který navštívíme podruhé:
  - dále platí, že $\ell \leq n$ (opět z *pigeonhole principle* vyplývá, že tento stav navštívíme nejpozději po přečtení n-tého znaku w),
- nyní rozdělíme w na tři části $w = xyz$ takto:
  - $x = w_1 \dots w_j$
  - $y = w_{j+1} \dots w_\ell$
  - $z = w_{\ell+1} \dots w_k$

1. $|y| > 0$ platí, protože $j < \ell$, a tedy z q_j do q_ℓ provedeme alespoň jeden přechod,
2. $|xy| \leq n$ platí, protože $|xy| = \ell \leq n$,
3. $xy^iz \in L$ pro každé $i \geq 0$ vyplývá z rovnosti $q_j = q_\ell$, neboli A čte slovo y ve smyčce, kterou lze vynechat nebo libovolně opakovat.



□

Použití PL při důkazu sporem

PL je nástroj, kterým můžeme v některých případech ukázat, že jazyk **není** regulární:

1. Předpokládáme, že daný jazyk je regulární a tedy musí platit PL.
2. Dle PL musí pro daný jazyk existovat nějaká konstanta $n \in \mathbb{N}$,
   - v důkazu sporem nikdy nevolíme konkrétní n, jen předpokládáme jeho existenci,
   - chceme ukázat, že L není regulární, a v takovém případě konstanta pro daný jazyk nemusí vůbec existovat.
3. Zvolíme vhodný tvar řetězce $w \in L$:
   - w musí být vyjádřeno vzhledem ke konstantě n tak, aby platilo $|w| \geq n$,
   - opět nikdy nevolíme konkrétní řetězec!
4. Snažíme se najít spor s PL
   - ukážeme, že pro **každé** rozdělení $xyz = w$ nemohou být zároveň splněny všechny tři podmínky 1, 2 a 3 z PL.

Příklad: jazyk $a^n b^n$ a PL

Věta

Jazyk $L = \{a^n b^n \mid n \in \mathbb{N}_0\}$ není regulární.

Důkaz

Využijeme PL v důkazu sporem: předpokládáme, že L je regulární a najdeme spor s PL.

- Dle PL tedy existuje konstanta, pojmenujme ji třeba $m \in \mathbb{N}$,
- zvolíme řetězec $w = a^m b^m$, na kterém ukážeme spor,
- platí $|w| \geq m$, proto dle PL by mělo jít rozdělit na $w = xyz$,
- každé rozdělení w splnující 1 a 2 z PL má tento tvar:
  - $x = a^j$
  - $y = a^\ell$
  - $z = a^{m-j-\ell} b^m$

  kde platí, že $\ell > 0$ (z 1) a zároveň $j + \ell \leq m$ (z 2), proto x a y obsahují jen znaky a,
- po napumpování w dle podmínky 3 dostáváme spor:
  - $xy^0z = a^j (a^\ell)^0 a^{m-j-\ell} b^m = a^{m-\ell} b^m \notin L$, což je spor ($xy^0z$ by měl patřit do L). □

Příklad: jazyk $\#_a(w) = \#_b(w)$ a PL

Věta

Jazyk $L = \{w \mid \#_a(w) = \#_b(w)\}$ není regulární.

Důkaz

Využijeme PL v důkazu sporem: předpokládáme, že L je regulární a najdeme spor s PL.

- Dle PL existuje konstanta $m \in \mathbb{N}$, a nejprve zvolíme řetězec $w' = (ab)^m$,
- w' sice splňuje podmínku $|w'| \geq m$, ale **spor na něm ukázat nejde**,
- takový řetězec w' lze totiž rozdělit následově:
  - $x = \varepsilon$
  - $y = ab$
  - $z = (ab)^{m-1}$
- přičemž jsou splněny všechny podmínky z PL, 1 a 2 platí triviálně, a pro 3 dostaneme, že pro každé $i \geq 0$ platí: $(ab)^i(ab)^{m-1} = (ab)^{m-1+i} \in L$,
- nyní zvolíme řetězec $w = a^m b^m$, na kterém už lze ukázat spor (postupujeme stejně jako u přechozího příkladu – po napumpování nebude mít xy^0z stejný počet a a b). □

Příklad: jazyk a^{n^2} a PL

Věta

Jazyk $L = \{a^{n^2} \mid n \in \mathbb{N}_0\} = \{\varepsilon, a^1, a^4, a^9, a^{25}, \ldots\}$ není regulární.

Důkaz

Využijeme PL v důkazu sporem: předpokládáme, že L je regulární a najdeme spor s PL.

- Dle PL existuje konstanta $m \in \mathbb{N}$,
- zvolíme řetězec $w = a^{m^2}$, který patří do jazyka L a $|w| \geq m$,
- z podmínky 2 dostáváme, že $|xy| \leq m$, a tedy i $|y| \leq m$,
- další řetězec délky čtverce následujícím po w je řetězec $v = a^{(n+1)^2}$,
- dále platí, že $|w| = n^2$ a $|v| = (n+1)^2$,
- po napumpování dle 3 má řetězec xy^2z délku $|xy^2z| \leq n^2 + n$,
- xy^2z nemá délku čtverce, protože $|w| < xy^2z < |v|$, to jest $n^2 < n^2 + n < n^2 + 2n + 1$ (pro $n > 0$), a tedy $xy^2z \notin L$ leží mezí dvěma po sobě následujícími čtverci. □