

# Moderné regulárne výrazy

Tatiana Tóthová

školiťel: RNDr. Michal Forišek PhD.

Fakulta matematiky, fyziky a informatiky  
Univerzita Komenského

22.apríla 2015

## Motivácia:

- teoretický model regulárnych výrazov bol v praxi rozšírený o nové konštrukcie

## Cieľ :

- preskúmať z hľadiska formálnych jazykov – Chomského hierarchia, vlastnosti triedy jazykov, zložitosť

Regex tvoria **znaky** a **metaznaky**

Základné operácie:

- $(e_1)(e_2)$  zreťazenie (bez znaku)
- $(e_1)|(e_2)$  alternácia
- $(e_1)^*$  Kleeneho uzáver

Navyše: ľubovoľný znak  $.$ , začiatok slova  $^$ , koniec slova  $\$$

Ďalšie operácie:

- množiny znakov  $[abc], [a - z]$
- komplementy množín znakov  $^{\wedge}abc, ^{\wedge}a - z]$
- opakovania  $?, \{n\}, \{n, m\}, +$

**Regex má konečnú dĺžku!**

# Zložitejšie konštrukcie (1)

**Číslovanie zátvoriek** zľava doprava, podľa otváracej zátvorky.  
Nečíslujú sa konštrukcie ( $? \dots$ )

**Spätné referencie**  $\backslash k$  – odkazujú sa na  $k$ -te zátvorky

$$\alpha \underset{k}{(} \underset{k}{\beta)} \gamma \backslash k \delta$$

$$w = \underbrace{x_1 \dots x_{i-1}}_{\alpha} \underbrace{\overbrace{x_i \dots x_{j-1}}^{w_k}}_{\underset{k}{(} \underset{k}{\beta)}} \underbrace{x_j \dots x_{l-1}}_{\gamma} \underbrace{\overbrace{x_l \dots x_{m-1}}^{w_k}}_{\backslash k} \underbrace{x_m \dots x_n}_{\delta}$$

$$w_k = x_i \dots x_{j-1} = x_l \dots x_{m-1}$$

## Zložitejšie konštrukcie (2)

**Lookahead** ( $? = \dots$ ) – nazeranie dopredu:  $\alpha(? = \beta)\gamma$

$$w = \underbrace{x_1 \dots x_{i-1}}_{\alpha} \overbrace{x_i \dots x_j}^{\beta} \underbrace{x_{j+1} \dots x_n}_{\gamma}$$

**Lookbehind** ( $? \leq \dots$ ) – nazeranie dozadu:  $\alpha(? \leq \beta)\gamma$

$$w = \underbrace{x_1 \dots x_{i-1}}_{\alpha} \overbrace{x_i \dots x_j}^{\beta} \underbrace{x_{j+1} \dots x_n}_{\gamma}$$

Ich **negatívne formy** ( $?! \dots$ ), ( $?<! \dots$ )  $\rightarrow$  opačná akceptácia (slová, ktoré sa v danom jazyku nenachádzajú)

**Regex** základné operácie a znaky  $\cdot \wedge \$ ( = \mathcal{R}$ )

**Eregex** + spätné referencie

**LEregex** + lookahead, lookbehind

**nLEregex** + negatívny lookahead, negatívny lookbehind

Triedy jazykov:  $\mathcal{R}$ ,  $\mathcal{L}_{ERE}$ ,  $\mathcal{L}_{LERE}$ ,  $\mathcal{L}_{nLERE}$

## Veta

Trieda nad *Regex* s pozitívnym a negatívnym lookaroundom je  $\mathcal{R}$ .

## Hierarchia

$$\mathcal{R} \subsetneq \mathcal{L}_{ERE} \subsetneq \mathcal{L}_{LERE} \subseteq \mathcal{L}_{nLERE} \subsetneq \mathcal{L}_{CS}$$

$\mathcal{L}_{LERE}$  a  $\mathcal{L}_{nLERE}$  sú neporovnateľné s  $\mathcal{L}_{CF}$

## Veta

Jazyk všetkých platných výpočtov Turingovho stroja patrí do  $\mathcal{L}_{LRE}$ .

## Veta

Nech  $\alpha \in LRegex$  nad unárnou abecedou  $\Sigma = \{a\}$  taký, že neobsahuje lookahead s  $\$$  ani lookbehind s  $\wedge$  vnútri iterácie  $(*)$ . Existuje konštanta  $N$  taká, že ak  $w \in L(\alpha)$  a  $|w| > N$ , potom existuje dekompozícia  $w = xy$  s nasledujúcimi vlastnosťami:

- (i)  $|y| \geq 1$
- (ii)  $\exists k \in \mathbb{N}, k \neq 0; \forall j = 1, 2, \dots : xy^{kj} \in L(\alpha)$



## Veta

$\mathcal{L}_{LRE}$  je uzavretá na zreťazenie.

**Dôkaz.** Nech  $\alpha, \beta \in LRE_{regex}$ . Regex pre  $L(\alpha)L(\beta)$  je

$$(\alpha \mid \beta) \mid \alpha \mid \beta$$

# Nový formalizmus

- konfigurácia:  $(r_1 \dots \lceil r_i \dots r_n, w_1 \dots \lceil w_j \dots w_m)$
- indexovateľné zátvorky
- krok výpočtu

$$(r_1 \dots (\dots) \lceil^k * \dots r_n, w_1 \dots w_a^k \dots w_b^{k'} \dots \lceil w_j \dots w_m)$$

$$(1) \vdash (r_1 \dots (\dots) * \lceil^k \dots r_n, w_1 \dots w_a^k \dots w_b^{k'} \dots \lceil w_j \dots w_m)$$

$$(2) \vdash (r_1 \dots (\lceil \dots) \lceil^k * \dots r_n, w_1 \dots w_a \dots w_b \dots \lceil^k w_j \dots w_m)$$

- akceptačný výpočet  $(\lceil r, \lceil w) \vdash^* (r \lceil, w \lceil)$
- jazyk

## Lema

Nech  $\alpha \in Eregex$  a  $w \in L(\alpha)$ . Potom existuje akceptačný výpočet, ktorý má najviac  $5 \cdot |\alpha| \cdot |w|$  konfigurácií.

Počet všetkých možných konfigurácií  $\alpha \in LEregex$  je

$$O(|\alpha| \cdot |w|^{|\alpha|+2})$$

Kvôli vlastnostiam lookaroundu lepší odhad nemáme.

Veta.

$$\mathcal{L}_{LRE} \subseteq NSPACE(\log n)$$

**Dôkaz.** Na páske si pamätáme informáciu z poschodových symbolov – vo forme adries

Zo Savitchovej vety vyplýva:  $\mathcal{L}_{LRE} \subseteq DSPACE(\log^2 n)$

Veta

$$\mathcal{L}_{nLRE} \subseteq DSPACE(\log^2 n)$$

**Dôkaz.** Idea Savitchovej vety, ale s konfiguráciami z formalizmu. Pre každý negatívny lookaround spúšťame nový Turingov stroj s opačnou akceptáciou.

$$r = |regex|, w = |word|$$

## Veta

$L(regex\#word) \in NSPACE(r \log w)$ ,  $regex \in LERegex$

**Dôkaz.** Vyplýva z  $\mathcal{L}_{LERE} \subseteq NSPACE(\log n)$ , dĺžka regexu už viac nie je konštanta.

## Veta

Nech počet konfigurácií v akceptačnom výpočte je polynomiálny od  $r$  a  $w$ , potom  $L(regex\#word) \in DSPACE(r \log^2 w)$ ,  $regex \in LERegex$

**Dôkaz.** Idea Savitchovej vety, ale s konfiguráciami z formalizmu.

Ďakujem za pozornosť!

## Lema 2

Nech  $\alpha \in \mathcal{L}_{LRE}$ ,  $s \in L(\alpha)$ ,  $r = |\alpha|$  a  $w = |s|$ . Potom existuje akceptačný výpočet, ktorý má nanajvýš  $O(r^2 w^3)$  konfigurácií.

Pripomeňme lemu 1

## Lema 1

Nech  $\alpha \in \mathcal{L}_{ERE}$  a  $w \in L(\alpha)$ . Potom existuje akceptačný výpočet, ktorý má najviac  $5 \cdot |\alpha| \cdot |w|$  konfigurácií.

Doplnenie na  $\mathcal{L}_{LRE}$ ? Funguje len pre regexy s konštantnou hĺbkou vnorenia lookaroundov.

Počet všetkých možných konfigurácií je  $O(|\alpha| \cdot |w|^{|\alpha|+2})$

Napríklad regex

$$((?= \underbrace{(?=(a^m) * \$)}_{a^{km}, k \in \mathbb{N}}) \underbrace{(a^{m+1}) * a\{1, m-1\} \$ \mid a^m \$)}_{\text{vie } a^* \text{ okrem } a^{m+1}, a^{m(m+1)l}, l \in \mathbb{N}}) a^m)^+ \\ \underbrace{\hspace{10em}}_{a^{km} \text{ také, že nevie } a^{m(m+1)l}, l \in \mathbb{N}}$$

generuje konečný jazyk obsahujúci slová

- $a^m$
- $a^{2m}$
- ...
- $a^{(m-1)(m+1)}$

Hlavný lookahead je spúšťaný každú iteráciu, teda pre slovo  $a^{zm}$  musí matchovať všetky  $a^{im}$  pre  $i \in \{1, \dots, z\}$ .