

Programovanie

3. ročník – odborná prax

Ing. Daniela Kravcová

REGULÁRNE VÝRAZY

Regulárny výraz - (regexp, regex, RE (*regular expression*))

- **je textový reťazec, obsahujúci vyhľadávací vzor, ktorý popisuje celú množinu reťazcov, na základe určitých syntaktických pravidiel.**
- regulárne výrazy sú používané v mnohých textových editoroch a nástrojoch na hľadanie a manipuláciu s textom podľa vzorov.
- mnoho programovacích jazykov podporuje regulárne výrazy na prácu s reťazcami (napr. Perl).

Využitie regulárnych výrazov

Medzi základné spôsoby využitia regulárnych výrazov patrí:

- 1. kontrola, či vstupný text vyhovuje danému regulárnemu výrazu**
- 2. rozdelenie vstupného textu na časti, oddelené separátorom, vyhovujúcim danému regulárnemu výrazu**
- 3. vyhľadávanie vo vstupnom texte**
 - zistenie pozícií vo vstupnom texte, kde sa nachádzajú výskyty, vyhovujúce danému regulárnemu výrazu
 - výber všetkých výskytov, vyhovujúcich danému regulárnemu výrazu
- 4. manipulácia s textom - náhrada výskytov**, pričom nahrádzať je v praxi možné:
 - fixným textom
 - nahrádzacím výrazom, kombinujúcim fixný text s odvolávkami na fragmenty,

Jednoznakové atómy

- x Literálny znak (ľubovoľný znak, mimo vyhradených metaznakov: . \ + * ? [^] \$ () { } | a prípadných ďalších, implementačne špecifických) popisuje práve len doslovný výskyt tohoto konkrétneho znaku. Napr. výrazu A tak vyhovuje jedine reťazec „A“ (v režime bez rozlišovania malých/veľkých písmen aj „a“).
- \x Opačné lomítko umožňuje zapísať doslovný výskyt vyhradeného metaznaku. Napr. výrazu * tak vyhovuje jedine reťazec „*“ a výrazu \\ reťazec „\“.
- Bodka indikuje ľubovoľný jeden znak (pričom podľa konkrétneho nastavenia sa môže alebo nemusí vzťahovať aj na riadiaci kód prechodu na nový riadok). Napr. výrazu Ad.m vyhovujú reťazce „Adam“, „AdXm“, „Ad9m“, „Ad!m“ a ďalšie.

Jednoznakové atómy

[trieda] Hranaté zátvorky indikujú jeden ľubovoľný znak z triedy, špecifikovanej obsahom zátvoriek. Trieda môže byť špecifikovaná vymenovaním konkrétnych znakov, uvedením rozsahu znakov (so symbolom spojovníka, „-“) a kombináciou týchto možností (jednotlivé implementácie potom ponúkajú aj ďalšie praktické možnosti špecifikácie). Napr. výrazu [a-dxy -] vyhovujú reťazce „a“, „b“, „c“, „d“, „x“, „y“, „“, „-“.

[^trieda] Hranaté zátvorky s obsahom, začínajúcim znakom striešky (^) indikujú jeden ľubovoľný znak, nepatriaci do triedy, špecifikovanej obsahom zátvoriek. Napr. výrazu [^a-dxy -] vyhovuje ľubovoľný jednoznakový reťazec, okrem reťazcov „a“, „b“, „c“, „d“, „x“, „y“, „“, „-“.

Zreťazenie

AB Ak A a B sú regulárne výrazy, tak výrazu AB vyhovujú reťazce, získané spojením (zreťazením) ľubovoľného reťazca, vyhovujúceho výrazu A s ľubovoľným reťazcom, vyhovujúcim výrazu B. Napr. výrazu Adam (zreťazenie jednoznakových literálnych atómov) tak vyhovuje jedine reťazec „Adam“. Výrazu $[a-c][01]$ vyhovujú reťazce „a0“, „a1“, „b0“, „b1“, „c0“ a „c1“.

Alternatívy (vetvenie)

| Zvislá čiara oddeľuje alternatívy (vetvy), predstavuje ekvivalent logického „alebo“. Výrazu Adam|Eva vyhovujú reťazce „Adam“ aj „Eva“.

Zoskupovanie

- () Jednoduché zátvorky vymedzujú rámec (scope), umožňujú tvorbu podvýrazov (a následné aplikovanie kvantifikátorov na celý podvýraz) a definujú skupiny, zachytávajúce podreťazce v nájdených výskytoch. Napr. výrazu Sol(ivar|ana|aris) vyhovujú reťazce „Solivar“, „Solana“ a „Solaris“.

Kotviace (pozičné) znaky

- ^ Strieška kotví na začiatok reťazca (prípadne riadku, pri riadkovom spracovaní). Bez použitia tohoto kotviaceho znaku vyhovie výrazu výskyt, začínajúci od ľubovoľnej pozície vo vstupnom texte.
- \$ Dolár kotví na koniec reťazca (prípadne riadku, pri riadkovom spracovaní). Bez použitia tohoto kotviaceho znaku vyhovie výrazu výskyt, končiaci na ľubovoľnej pozícii vo vstupnom texte.

Kvantifikácia

- ? Otáznik indikuje žiaden alebo jeden výskyt predchádzajúceho atómu. Napr. $ab?c$ popisuje reťazce „ac“ a „abc“.
- * Hviezdička (Kleeneho operátor) indikuje nula alebo viac výskytov predchádzajúceho atómu. Napr. ab^*c popisuje „ac“, „abc“, „abbc“, „abbbc“, atď.
- + Plus indikuje jeden alebo viac výskytov predchádzajúceho atómu. Napr. $ab+c$ popisuje „abc“, „abbc“, „abbbc“, atď.; nie však reťazec „ac“.

Kvantifikácia

- {n}** Exaktne n výskytov predchádzajúceho atómu. Napr. $\text{mu}(\text{ha})\{3\}$ popisuje reťazec „muhahaha“.
- {min,}** Najmenej min výskytov predchádzajúceho atómu. Napr. $\text{mu}(\text{ha})\{2, \}$ popisuje reťazce „muhaha“, „muhahaha“, „muhahahaha“, atď.
- {min,max}** Najmenej min , najviac max výskytov predchádzajúceho atómu. Napr. $\text{mu}(\text{ha})\{2,4\}$ popisuje reťazce „muhaha“, „muhahaha“ a „muhahahaha“.

Príklady

- **So(ň|v)a** popisuje reťazce „Soňa“ a „Sova“.
- **Ba*f** popisuje reťazce „Bf“, „Baf“, „Baaf“, „Baaaf“, atď.
- **\d{3} ?\d{2}** popisuje formát [PŠČ](#) – postupnosť troch číslíc, nepovinnú medzeru a dve číslice, napr. „845 08“, „97644“ a pod.
- **<[^>]*>** popisuje (zjednodušene) tag v jazyku [HTML](#) – ľubovoľný text, uzavretý medzi ostré zátvorky.
- **[0-9a-fA-F]+(, *[0-9a-fA-F]+)*** popisuje zoznam [hexadecimálnych](#) čísel, oddelených čiarkami a nepovinnými medzerami, napr. „10C, 61, 75, 263A, 0D, 0A“.
- **((a*b)(a*b))*a*** popisuje slová z abecedy {a, b}, obsahujúce párny počet b, napr. „baaabbaaaaaaba“.