

Regularni izrazi

1 Osnovni regularni izrazi

Regularni izrazi predstavljaju mehanizam za prepoznavanje određenih šablona u tekstu. Osnovni regularni izrazi su:

- a : Ako je a karakter iz azbuke, tada je a regularni izraz koji prepoznaje samo taj karakter.
- r_1r_2 : Ako su r_1 i r_2 regularni izrazi, tada je r_1r_2 regularni izraz koji prepoznaje niz gde se r_2 nalazi nakon r_1 .
- $r_1 \mid r_2$: Ako su r_1 i r_2 regularni izrazi, tada je $r_1 \mid r_2$ regularni izraz koji prepoznaje sve što prepoznaje r_1 ili r_2 .
- r^* : Za regularni izraz r , izraz r^* prepoznaje 0 ili više pojavljivanja izraza r .

2 Prošireni regularni izrazi

Veličina zapisa regularnog izraza brzo raste kako raste složenost regularnog jezika. Radi skraćivanja zapisa i povećanja čitljivosti, uvode se tzv. *prošireni regularni izrazi* koji predstavljaju kraće zapise nekih osnovnih regularnih izraza.

Prošireni regularni izraz	Regularni izraz
$[a_1a_2 \dots a_n]$	$a_1 \mid a_2 \mid \dots \mid a_n$
$[\sim a_1a_2 \dots a_n]$	Svi karakteri iz azbuke bez $\{a_1, a_2, \dots, a_n\}$
$[a_1 - a_2]$	Disjunkcija karaktera od a_1 do a_2 iz ASCII tabele
$\backslash w$	$[A-Za-z0-9_]$
$\backslash W$	$[\sim A-Za-z0-9_]$
$\backslash d$	$[0-9]$
$\backslash D$	$[\sim 0-9]$
$\backslash s$	$[\backslash t \backslash r \backslash n \backslash f]$ (sve bjeline)
$\backslash S$	$[\sim \backslash t \backslash r \backslash n \backslash f]$
\cdot	$[\sim \backslash n]$ (sve osim novog reda)
a^+	aa^* (jedno ili više pojavljivanja)
$a^?$	$\varepsilon \mid a$ (nula ili jedno pojavljivanje)
$a\{n\}$	a^n
$a\{n,m\}$	$a^k, n \leq k \leq m$
$a\{n,\}$	$a^k, n \leq k$

Karakteristični sa specijalnim značenjem **u karakterskoj klasi**:

`-`, `\`, `]`, `^`

Karakteristični sa specijalnim značenjem **van karakterske klase**:

`*`, `+`, `?`, `|`, `.`, `\`, `(`, `)`, `[`, `...`

Za oduzimanje specijalnog značenja koristi se `\` (backslash).

3 Gramzivost i lenjo izračunavanje

Gramzivost: Poklapanje izraza vrši se u poslednjem mogućem trenutku. Gramzivi operatori su: `*`, `+`, `?`, `{ }`

Gramzivost se uklanja dodatnim operatorom `?` ukoliko je to podržano. Na primer: `a+?` je negramzivo poklapanje.

Lenjo izračunavanje: Ako se u disjunkciji prepozna prvi disjunkt, drugi se više ne razmatra.

4 Sidra

Sidra određuju položaj prepoznavanja regularnog izraza u tekstu, ne prepoznaju karaktere.

- `^` — početak teksta ili linije (u `multiline` režimu)
- `$` — kraj teksta ili linije (u `multiline` režimu)
- `\b` — granica reči (`\w+` naspram `\W`)

5 Preduvidi i postuvidi

5.1 Pozitivni preduvid

$r_1(? = r_2)$ izrazi r_1 iza kojih stoji r_2

$(? = r_1)r_2$ izrazi r_2 koji počinju izrazom r_1

5.2 Negativni preduvid

$r_1(!r_2)$ izrazi r_1 iza kojih se *ne* nalazi r_2

$(?!r_1)r_2$ izrazi r_2 koji *ne* počinju r_1

5.3 Pozitivni postuvid

$(? \leq r_1) r_2$ izrazi r_2 ispred kojih stoji r_1

$r_1 (? \leq r_2)$ izrazi r_1 koji se završavaju izrazom r_2

5.4 Negativni postuvid

$(? < !r_1) r_2$ izrazi r_2 ispred kojih *ne* stoji r_1

$r_1 (? < !r_2)$ izrazi r_1 koji se *ne* završavaju izrazom r_2

6 Grupisanje i referenciranje (Backreferences)

Grupisanje se vrši stavljanjem dijela izraza u zagrade (). Grupama se pristupa pomoću reference \k, gde je k redni broj grupe. Numeracija odgovara redosledu otvorenih zagrada.