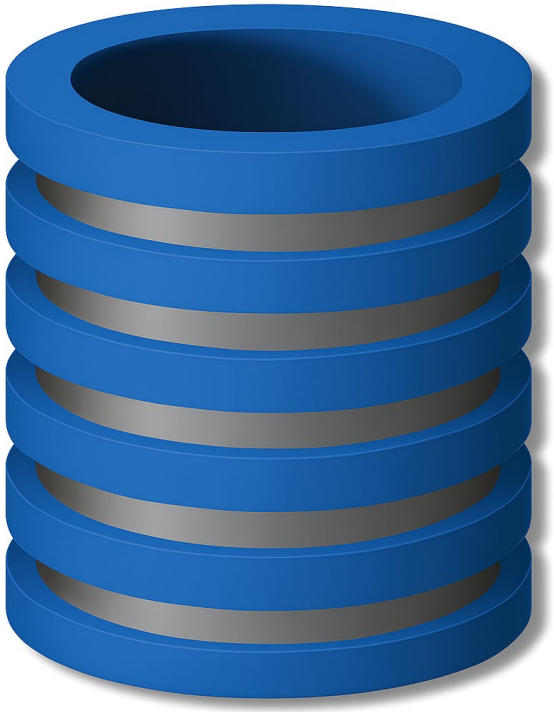


Regularni iskazi (RegExp)

Predmet: Administriranje Baze Podataka

Predavač: dr Dušan Stefanović

REGULARNI SKAZI

RegExp definicija

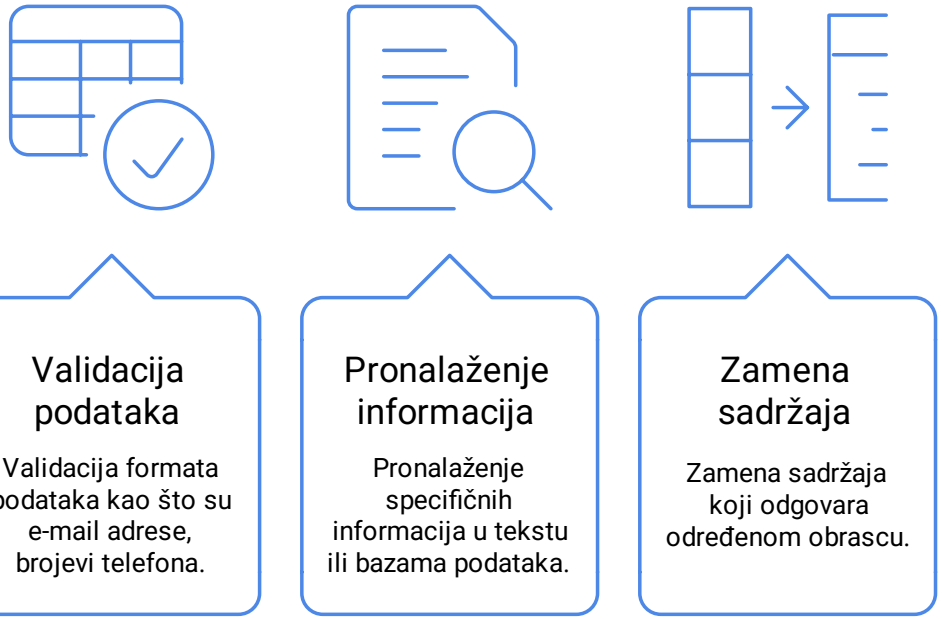
predstavljaju specijalan tekstualni niz koji definiše obrazac (šablon, uzorak) za pretragu ili validaciju podataka

Fleksibilnost

Za razliku od wildcard izraza, regularni izrazi su fleksibilniji i omogućavaju preciznije definisanje složenih obrazaca

Efikasnost

Pretraga pomoću regularnih izraza je efikasna jer se rezultati dobijaju već tokom jednog prolaza kroz podatke.



REGULARNI SKAZI - PRETRAGA

Početak procesa

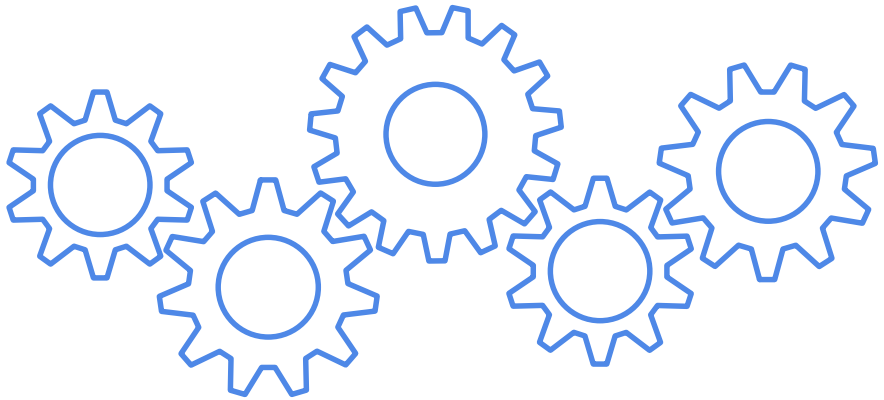
Funkcioniše kao automat koji pregleda tekst karakter po karakter, bez potrebe da se vraća na početak.

Odlučivanje o poklapanju

Odlučuje da li trenutno pročitani niz odgovara šablonu.

Završetak procesa

Zahvaljujući tzv. finite state machine modelu iza regularnih izraza, svaki karakter se procesira najviše jednom



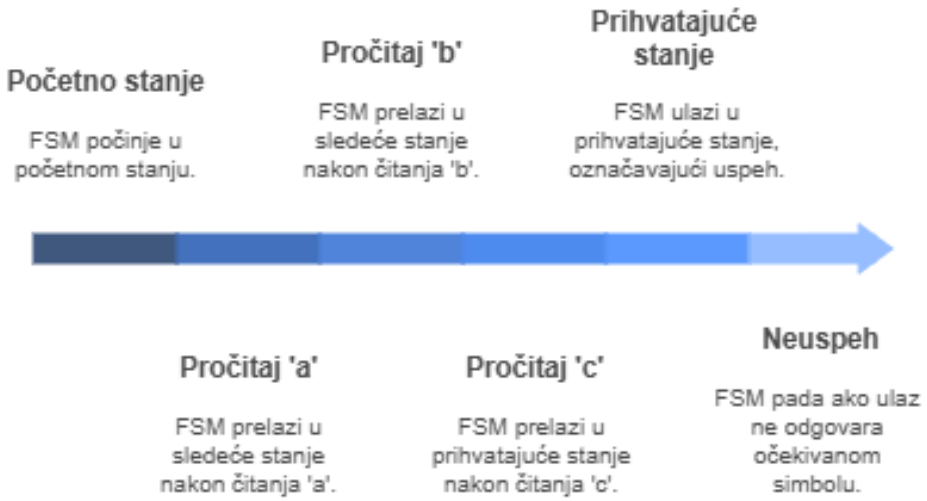
Provera karaktera

Svaki karakter se proverava redom.

Napredovanje kroz tekst

Proces se kreće napred kroz tekst.

FSM – PRIMER PREPOZNAVANJA REGULARNOG IZRAZA abc



Karakteristika	Opis
Konačan broj stanja	FSM ima tačno određena stanja kroz koja može da se kreće
Deterministički prelazi	Za svaki ulaz i trenutno stanje postoji tačno jedno naredno stanje
Ulaz se čita sekvencijalno	FSM čita jedan simbol po jedan — **ne vraća se unazad**
Prihvatajuća stanja	FSM označava da li ulazni niz odgovara pravilima
Jednostavna logika	Sve odluke se zasnivaju na trenutnom stanju + sledeći simbol

OSOBINE REGEXP Patern (^)

REGEXP **nije case sensitive**, osim ukoliko se koristi **binary** string.

MySQL upit prikazuje imena studenata čija imena počinju sa ‘k’.

‘^’ se koristi za mečovanje početka imena.

```
SELECT *
FROM Student
WHERE Ime REGEXP '^k';
```

MySql iskaz prikazuje imena studenata koja počinju sa ‘k’ vodeći računa o malim tj. velikim slovima jer se koristi BINARY operator.

```
SELECT *
FROM Student
WHERE Ime REGEXP BINARY ’^k';
```

OSOBINE REGEXP
Patern ($)

MySql iskaz pronalazi imena studenata čija se imena završavaju na ‘on’.
‘$’ se koristi za mečovanje završetka stringa.

```
SELECT *
FROM Student
WHERE Ime REGEXP "on$" ;
```

MySql iskaz pronalazi autore čija imena sadrže ‘t’ u imenu.

```
SELECT *
FROM Student
WHERE Ime REGEXP "t";
```

OSOBINE REGEXP
Patern ([...])

MySql iskaz pronalazi imena studenata koja sadrže sledeća slova u imenu ‘z’ ili ‘v’ ili ‘m’.

```
SELECT *
FROM Student
WHERE Ime REGEXP "[zvm]";
```

MySql iskaz pronalazi imena studenata koja sadrže karaktere od b do g

```
SELECT *
FROM Student
WHERE Ime REGEXP "[b-g]" ;
```

OSOBINE REGEXP Patern (.) i ({...})

MySql iskaz pronalazi imena studenata koja sadrže tačno 12 karaktera.
Koristi se **‘^’** i **‘$’** podudaranje za početak i kraj imena i **12 instanci ‘.’** za podudaranje sa 12 karaktera.
Tačka (**.**) je patern koji odgovara jednom bilo kom karakteru

```
SELECT *
FROM Student
WHERE Ime REGEXP '^.....$';
```

MySql iskaz pronalazi imena studenata koja sadrže tačno 12 karaktera.
Koristi se ‘^’ i ‘$’ podudaranje za početak i kraj imena i instanca ‘.’ .
Srednje zagrade **{}** za definisanje koliko puta se tačka ponavlja.

```
SELECT * FROM Student
WHERE Ime REGEXP '^.{12}$';
```

OSOBINE REGEXP Paterni

Patern	Šta patern mečuje
^	Početak stringa
$	Kraj stringa
.	Jedan bilo koji karakter
[...]	Bilo koji karakter koji se nalazi u uglastim zagradama
[^...]	Bilo koji karakter koji se ne nalazi u uglastim zagradama
p1|p2|p3	Alternativa; mečuje ili p1 ili p2 ili p3 patern
*	Nula ili više instanci prethodnog elementa
+	Jedna ili više instanci prethodnog elementa
{n}	n instanci prethodnog elementa
{m,n}	m do n instanci prethodnog elementa

Podudaranje po tipu klase

POSIX Klasa	Opis mečovanja
[:alnum:]	Alfanumerički znakovi (slova ili brojevi). Odgovara [a-zA-Z0-9]
[:alpha:]	Bilo koje slovo. Odgovara [a-zA-Z]
[:blank:]	Space i tabulator. Odgovara [\t]
[:cntrl:]	ASCII kontrolni karakteri (npr. \n, \r, \0)
[:digit:]	Cifre. Odgovara [0-9]
[:graph:]	Bilo koji karakter osim whitespace (space, tab)
[:lower:]	Mala slova. Odgovara [a-z]
[:punct:]	Interpunkcijski znakovi (npr. . , ! ? ; : - _)
[:space:]	Bilo koji razmak (space, tab, newline, carriage return...)
[:upper:]	Velika slova. Odgovara [A-Z]
[:xdigit:]	Hexadecimalne vrednosti. Odgovara [0-9A-Fa-f]

Napomena:
POSIX klase ([:alpha:], [:digit:],...) funkcionišu u MySQL, grep, awk, ali ne u JavaScript, Python, ...

SELECT * FROM Zemlje where Ime regexp '[[:blank:]]'

Escape Sekvenca

Želimo da pretražimo oznaku CD modela [7543], problem je što se uglaste zagrade koriste i za podudaranje karaktera koji se nalaze u uglastim zagradama.

Escape sekvencu koristimo ukoliko želimo da mečujemo same uglaste zagrade a ne njenu REGEXP sintaksu.

```
SELECT *

FROM Proizvod

WHERE Ime_proivoda REGEXP 'CD-RW Model \[7543\]';
```

Repetition Match

Koristi se kada želimo da mečujemo ponavljanje određene istance više puta

Patern	Opis
*	Nula ili više instanci prethodnog elementa
+	Jedna ili više instanci prethodnog elementa
{n}	n instanci prethodnog elementa
{n,}	minimum n instanci prethodnog elementa
{n1,n2}	Minimum n1 a maksimum n2 instance
?	Nula ili jedna instanca prethodnog elementa

```
SELECT *
FROM product
WHERE opis REGEXP 'Drives?';
```

```
|      kod      |       ime                    | opis                  |

|       3       | WildTech 250Gb 1700          | SATA Disk Drive       |
|       20      | MasterSlave Multi-pack       | 5 SATA Disk Drives    |
```

Podudaranje na osnovu pozicije u tekstu

Patern	Opis
^	Početak teksta
$	Kraj teksta
[[:<:]]	Početak reči
[[:>:]]	Kraj reči

```
SELECT ime
FROM Proizvod
WHERE ime REGEXP '[[:<:]]One[[:>:]]';
```

OSOBINE REGEXP

Za mečovanje kopleksnih iskaza tj paterna potrebno je razumeti tri simbola koja se koriste kod REGEXP (zagrade, $(mečuje kraj stringa) i ^(mečuje početak stringa))

Postoje tri vrste zagrada

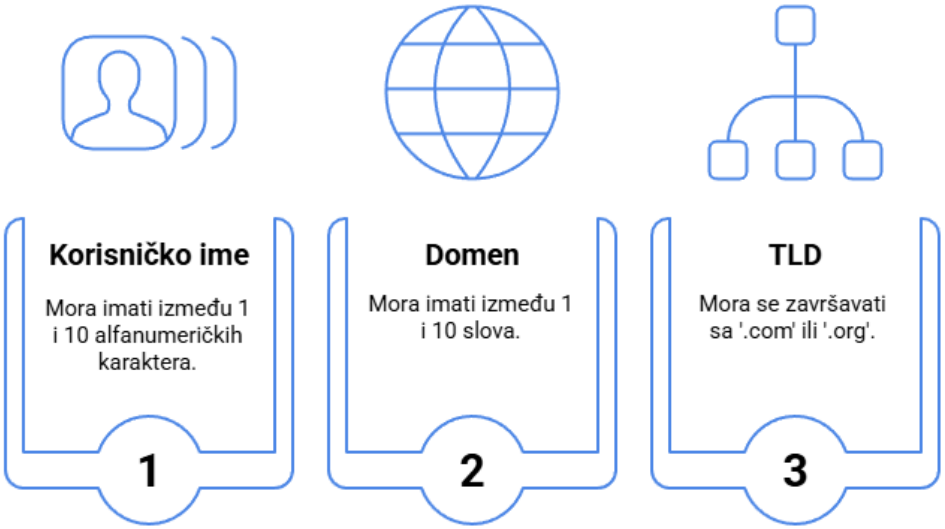
[] -> definišu karaktere koje treba mečovati

{} -> definišu koliko karaktera treba mečovati

() -> koriste se za grupisanje

Šta odgovara sledećem regexp-u

```
^[a-zA-Z0-9]{1,10}@[a-zA-Z]{1,10}\.(com|org)$
```



Primeri koji odgovaraju izrazu:

pera123@domen.com

ANA9@abcORG.org

x@a.com

Primeri koji ne odgovaraju:

pera@domen.rs (ne završava se sa com ili org)

pera1234567890@abc.com (korisničko ime duže od 10 znakova)

ime@123.com (domen sadrži cifre – nije dozvoljeno ovim izrazom)

ime@a.b.com (nema podršku za poddomene)

PRIMERI

1. Mečuj karakter između a-g bilo gde u iskazu

   [a-g]

2. Mečuj karaktere između a-g sa dužinom 3. Potrebno je definisati dužinu validacije

   [a-g]{3}

   ^[a-g]{3}$ -> dozvoljavamo samo tri bilo koja slova iz opsega a-g u iskazu

3. Mečuj karaktere između a-g sa minimum 1 karakterom a maximum 3

   ^[a-g]{1,3}$

4. Mečovanje 8 digita fiksne dužine (npr 12345678)

   ^[0-9]{8}$

5. Mečovanje minimum 3 a maksimalno 7 digita

   ^[0-9]{3,7}$

6. Mečovanje računa gde su prva tri karaktera tekstualna a naredna 8 su brojevi

   ^[a-zA-Z]{3}[0-9]{8}$

7. Prosta validacija web adrese

   ^[w]{3}\.[A-Za-z0-9]{1,12}\.(com|edu|rs)

PRIMERI

1. Napisati REGEXP koji pronalazi mala slova, cifre, donju i srednju crtu. Dozvoljen broj karaktera je između 3 i 16.

   ^[a-z0-9_-]{3,16}$

2. Prikaz heksa vrednosti sa '#' opcionim karakterom i sa mogućnošću da se prikažu 6 ili 3 heksa vrednosti.

   ^#?([a-f0-9]{6}|[a-f0-9]{3})$

PRIMERI

Pretraga email adrese

^([a-z0-9_\.-]+)@([\da-z\.-]+)\.([a-z\.]{2,6})$

Redni broj	Opis regularnog izraza
^	Početak stringa
([a-z0-9_\.-]+)	Korisničko ime: mala slova, cifre, donja crta, tačka, crtica (jedan ili više puta)
@	Obavezni znak '@'
([\da-z\.-]+)	Domen: cifre, mala slova, tačka, crtica (jedan ili više puta)
\.	Obavezna tačka između domena i TLD-a
([a-z\.]{2,6})	TLD: mala slova i tačka, dužina 2 do 6 karaktera
$	Kraj stringa

PRIMERI

Pretraga web adrese

```
^(https?:\/\/)?([\da-z\.-]+)\.([a-z\.]{2,6})([\/\w \.-]*)*\/?$
```

Deo izraza	Značenje
^	Početak stringa
(https?:\/\/)?	Opcioni protokol http:// ili https://
([\da-z\.-]+)	Domen (npr. www, google, example123) — dozvoljene su cifre, mala slova, tačka i crtica
\.	Obavezna tačka pre TLD-a
([a-z\.]{2,6})	TLD (npr. com, org, co.uk) — 2 do 6 malih slova ili tačaka
([\/\w \.-]*)*	Putanja: opcioni deo koji sadrži /, reči, razmake, tačke i crtice, i može da se ponavlja
\/?	Opcioni završni /
$	Kraj stringa