REGULARNI IZRAZI

REGEX

REGULAR EXPRESSIONS

REGULARNI IZRAZ

- Regular Expression (RegEx) – jezik za opisivanje obrazaca
- Koristi se za naprednu pretragu teksta
- Podržan od strane mnogo programskih jezika
- Jako je nečitak, ali je jako koristan

JAVA REGULARNI IZRAZI

Konstrukcija	Opis
`[abc]`	a, b ili c (simple class)
`[^abc]`	Bilo koji karakter osim a, b ili c (negacija - negation)
`[a-zA-Z]`	a do z, ili A do Z, inkluzivno (opseg - range)
`[a-d[m-p]]`	a do d ili m do p: [a-dm-p] (unija - union)
`[a-z&&[def]]`	d, e ili f (presjek - intersection)
`[a-z&&[^bc]]`	a do z, osim b I c: `[ad-z]` (razlika - subtraction)
`[a-z&&[^m-p]]`	a do z bez slova od m do p: `[a-lq-z]` (razlika - subtraction)
.	Bilo koji karakter
`\d`	Cifra: [0-9]
`\D`	Nije clfra : `[^0-9]`
`\s`	Bjelina (whitespace) karakter: `[ \t\n\x0B\f\r]`
`\S`	non-whitespace karakter: `[^\s]`
`\w`	word karakter: `[a-zA-Z_0-9]`
`\W`	non-word karakter: `[^\w]`

JAVA REGULARNI IZRAZI

Konstrukcija	Opis
`[abc]`	a, b ili c (simple class)
`[^abc]`	Bilo koji karakter osim a, b ili c (negacija - negation)
`[a-zA-Z]`	a do z, ili A do Z, inkluzivno (opseg - range)
`[a-d[m-p]]`	a do d ili m do p: [a-dm-p] (unija - union)
`[a-z&&[def]]`	d, e ili f (presjek - intersection)
`[a-z&&[^bc]]`	a do z, osim b I c: `[ad-z]` (razlika - subtraction)
`[a-z&&[^m-p]]`	a do z bez slova od m do p: `[a-lq-z]` (razlika - subtraction)
.	Bilo koji karakter
`\d`	Cifra: [0-9]
`\D`	Nije clfra : `[^0-9]`
`\s`	Bjelina (whitespace) karakter: `[ \t\n\x0B\f\r]`
`\S`	non-whitespace karakter: `[^\s]`
`\w`	word karakter: `[a-zA-Z_0-9]`
`\W`	non-word karakter: `[^\w]`

JAVA REGULARNI IZRAZI

`X? X`	jednom ili nijednom
`X* X`	nula ili više puta
`X+ X`	jedan ili više puta
`X{n} X`	tačno n puta
`X{n,} X`	najmanje n puta
`X{n,m} X`	najmanje n puta ali ne više od m puta

METODE KLASE STRING

Metod	Opis
`s.matches("regex")`	Izračunava da li s odgovara "regex". Vraća true ako CIO string odgovara.
`s.split("regex")`	Kreira niz podstringova stringa s koji se dobija dijeljenjem stringa `s` na osnovu separatora "regex". "regex" se ne uključuje u rezultat.
`s.replaceFirst("regex","replacement")`	Mijenja prvo pojavljivanje "regex" sa "replacement”.
`s.replaceAll("regex"),"replacement"`	Mijenja sva pojavljivanja "regex" sa replacement.

POSIX I PERL EKSTENZIJE

.	Tačka pronalazi bilo koji znak.
[ ]	Pronalazi bilo koji znak koji se nalazi u uglatim zagradama. Na primjer [abc] pronalazi a, b i c, ali ne i ostale znakove.
[-]	Pronalazi bilo koji znak koji se nalazi u uglatim zagradama između ova dva znaka. Na primjer [a-c] pronalazi a, b i c, ali ne i ostale znakove.
[^]	Pronalazi bilo koji znak osim onih koji se nalaze u uglatim zagradama. Na primjer [^abc] pronalazi sve znakove osim a, b i c.
^	Ukoliko je prvi znak regularnog izraza označava da izraz mora biti pronađen na početku linije (u Microsoft Wordu odlomka).
$	Znak dolar na kraju regularnog izraza označava da izraz mora biti pronađen na kraju linije (u Microsoft Wordu odlomka).
*	Znak * pronalazi 0 ili više puta prethodni znak. Na primjer bo* će pronaći b, bo i boo.
+	Znak + pronalazi jednom ili više puta prethodni znak. Na primjer bo* će pronaći bo i boo, ali neće b.
?	Znak ? pronalazi niti jednom ili jednom prethodni znak. Na primjer bo? će pronaći b i bo, ali neće boo.
{n}	Gdje je n broj, pronalazi prethodni znak točno n puta. Na primjer o{2} pronalazi točno oo.
{n,}	Gdje je n broj, pronalazi prethodni znak najmanje n puta. Na primjer o{2,} pronalazi oo, ooo, oooo, itd.
{n,m}	Gdje su n i m brojevi, pronalazi prethodni znak najmanje n, a najviše m puta. Na primjer o{2,3} pronalazi oo i ooo.
( )	Pronalazi bilo koju sekvencu unutar zagrada. Na primjer (ab)+ pronalazi ab, abab, ababab. Zagrade nam služe za grupiranje znakova, tako da kvantifikatori '*', '+', '?' i {} na njih gledaju kao na jedan znak.
|	Logički ILI operator. Na primjer prvo|drugo pronalazi prvo ili drugo, dok R(1|2) pronalazi R1 ili R2.
\	Obrnuta kosa crta tretira specijalne znakove literalno. Na primjer znak + označava pronađi prethodni znak jednom ili više puta, dok \+. Pronalazi baš +.
\n	Pronalazi znak kraja linije.
\r	Pronalazi carriage-return znak. U Windowsima se kraj linije označava s \n\r, ali će i sami '\n' u većini programa pronaći kraj linije.
\t	Pronalazi tabulator.
\a,\e,\f,\v	Osim \n, \r i \t ostali kontrolni znakovi su praktično izumrli, pa ih ovdje nema potrebe opisivati.

POSIX I PERL EKSTENZIJE

\C	Pronalazi bilo koji znak. Ekvivalentno tački ('.').
\d	Pronalazi broj (digit). Ekvivalentno izrazu [0-9].
\D	Pronalazi bilo koji znak osim brojeva. Ekvivalentno izrazu [^0-9].
\l	Pronalazi bilo koje malo slovo (lower case) uključujući i internacionalna slova. Za hrvatski jezik ekvivalentno izrazu [a-zčćđšž].
\L	Pronalazi bilo koji znak osim malog slova. Za hrvatski jezik ekvivalentno izrazu [^a-zčćđšž].
\s	Pronalazi bilo koji razmak (space) uključujući i tabulatore. Ekvivalentno izrazu [\t\n\r\f\v].
\S	Pronalazi bilo koji znak koji nije razmak (tabulator). Ekvivalentno izrazu [^ \t\n\r\f\v].
\u	Pronalazi bilo koje veliko slovo (upper case)uključujući i internacionalna slova. Za hrvatski jezik ekvivalentno izrazu [A-ZČĆĐŠŽ].
\U	Pronalazi bilo koji znak osim velikih slova. Za hrvatski jezik ekvivalentno izrazu [^A-ZČĆĐŠŽ].
\w	Pronalazi bilo koji word znak. Word znakovi su slova, brojevi i podvučeno. Uglavnom irelevantno za obične korisnike, naime radi se o znakovima koji se u C/C++ jeziku mogu koristiti za varijable, funkcije itd.
\W	Pronalazi bilo koji osim word znakova.
\0dd	Pronalazi znak s ovim oktalnim ASCII/ANSI kodom. Gdje je dd jedan ili više oktalnih brojeva.
\xXX	Pronalazi znak s ovim heksadecimalnim ASCII/ANSI kodom. Gdje je xx jedan ili heksadecimalni broj.
\Q	Svi znakovi koji slijede nakon ovog znaka sve do \E znaka se tretiraju literalno.
\Q	Zatvara slijed literalnih znakova započet s \Q znakom.

JAVASCRIPT - MODIFIKATORI

- U javascriptu se regularni izraz sastoji od 2 dijela
  - Izraz (unutar kosih crta)
  - Modifikatori (opciono)

```
/izraz/modifikatori
```

- Modifikatori mogu biti:
  - i - *insensitive* – ne razlikuju mala i velika slova
  - g - *global* – odnosi se na sva podudaranja (ne samo na prvo)
  - m - *multiline* – pretraga u više redova teksta

JAVASCRIPT - MODIFIKATORI

- U primjeru pronalaska određenih riječi u tekstu:

```
const izjava = „Noćas ne k'o lubenica pun mjesec iznad Bosne!"
const izraz = /mjesec/i
if (izraz.test(izjava))
      console.log(„Pjesma sadrži riječ mjesec.")
else
      console.log("Pjesma ne sadrži riječ mjesec.")
```

```
Pjesma sadrži riječ mjesec.
```

JAVASCRIPT - MODIFIKATORI

- Znak ^ na početku izraza koji se traži očekuje da izraz bude na početku teksta
- Znak $ na kraju izraza koji se traži očekuje da izraz bude na kraju teksta

```
const izjava = „Mjesec, mjesec je pobjegao jutro je moramo i mi"
const izraz = /^Mjesec/
if (izraz.test(izjava))
      console.log(„Pjesma počinje traženim izrazom.")
else
      console.log("Pjesma ne počinje traženim izrazom.")
```

Pjesma počinje traženim izrazom.")

JAVASCRIPT - MODIFIKATORI

- Grupisanjem više znakova u uglaste zagrade `[]`, traži se bilo koji od njih

```
const izjava = „O, dobra noć Banjaluko!"
const izraz = /[BJ]an/
const ne = izraz.test(izjava) ? "" : "ne "
console.log(`Obrazac se ${ne}nalazi u izjavi.`)
```

```
Obrazac se nalazi u izjavi.`)
Traži se Jan ili Ban
```

JAVASCRIPT - MODIFIKATORI

- Znak ^ unutar uglaste zagrade [] ima značenje negacije

```
const izjava = „Banjaluka"
const izraz = /[^Bljak]/
const ne = izraz.test(izjava) ? "" : "ne "
console.log(`Obrazac se ${ne}nalazi u izjavi.`)
```

```
Traži bilo koji znak a da nije B,l,j,a,k:
Obrazac se ne nalazi u izjavi.
```

JAVASCRIPT - MODIFIKATORI

- Znak – unutar uglaste zagrade `[]` označava raspon

```
[a-z]     traži sva mala slova od a do z
[A-Z]     traži sva velika slova od A do Z
[a-Z]     traži sva slova
[0-9]     traži sve brojeve od 0 do 9
[^5-9]    traži sve brojeve koji nisu u rasponu od 5 do 9
```

MODIFIKATORI – BROJ PONAVLJANJA (?, +, *, {})

- Znak pitanja (`?`) znači da se prethodni znak javlja opciono (0 ili jedanput).
- Zvjezdica (`*`) znači da se prethodni znak javlja opciono, 0 ili više puta.
- Plus (`+`) znači da se prethodni znak javlja obavezno, `1` ili više puta zaredom.
- Broj `n` unutar vitičastih zagrada `{}` znači da se prethodni obrazac javlja `n` puta
- Kvantifikator `{n1,n2}` znači da se prethodni obrazac javlja najmanje `n1` puta, ali ne više od `n2` puta
- Kvantifikator `{n,}` znači da se prethodni znak javlja najmanje `n` puta.

MODIFIKATORI – BROJ PONAVLJANJA (?, +, *, {})

- `rj?ečnik traži rečnik i rječnik`
- `[0-9]{4} traži bilo koji četvorocifreni broj`
- `[0-9]{3, 6} traži sve brojeve između trocifrenih i šestocifrenih`

POSEBNI KARAKTERI (\D, \D, \W, \W, \S, \S)

- Znaci sa specijalnim značenjem

.	bilo koji znak
\w	slovo
\W	ne-slovo
\d	broj
\D	ne-broj
\s	praznina
\S	ne-praznina
\n	nova linija
\t	tab