5. Regularni izrazi

Branislava Šandrih
branislava.sandrih@fil.bg.ac.rs

Šta su regularni izrazi?

- Regularni izraz je šablon koji opisuje nekakav niz karaktera
  - Sav tekstualni sadržaj u računaru u stvari niz karaktera
- Terminologija:
  - Regular expressions
  - Regex
  - Regexp
- Niske koje odgovaraju šablonu, možemo nazvati „uparenim niskama“
  - eng. match

Prvi primer i online-alat

- Neka je tekst:

  Ovo je kratko predavanje o regularnim izrazima.

- Za regularni izraz

  predavanje

- Uparena niska je (istaknuta):

  Ovo je kratko predavanje o regularnim izrazima

- Dakle, sami literali su vrsta regularnih izraza
- Regularne izraze ćemo isprobavati direktno na http ://regexr.com/

Vrste karaktera

- Karakteri koji se javljaju u regularnim izrazima su:
  - literali
  - meta-karakteri
  - specijalni karakteri
- Literali su karakteri:
  - a b c d … z A B C D … Z 0 1 2 … 9
  - i svi ostali koji nisu meta-karakteri
- Meta-karakteri (ima ih 12) su:

  \ ^ $. | ? * + ( ) [{
  - retko se koriste samostalno
  - **upravo meta-karakteri pružaju mogućnost složenijeg pretraživanja teksta!**
- Specijalni karakteri su karakteri koji nemaju grafičku reprezentaciju
  - novi red, tab

Meta-karakteri kao literali

- Ukoliko je potrebno u tekstu pronaći baš neki od karaktera koji pripada skupu meta-karaktera, to je moguće
- Za to se koristi karakter \
  - eng. backslash
- Na primer, za regularni izraz

  1\+1 = 2

  u tekstu

  50–8*6+1+1=2+2

  pronalazi se podniska

  50–8*6+1+1=2+2

Meta-karakteri kao literali

- Ukoliko je potrebno u tekstu pronaći baš neki od karaktera koji pripada skupu meta-karaktera, to je moguće
- Za to se koristi karakter \
  - eng. backslash
- Na primer, za regularni izraz

  1\+1=2

  u tekstu

  50–8*6+1+1=2+2

  pronalazi se podniska

  50–8*6+**1+1=2**+2

Oprez!
Regularni izraz
1+1=2
bi pronašao, npr, nisku poput
111=2

Meta-karakteri kao literali

- Ukoliko je potrebno [illegible] od karaktera koji pripada skupu meta
- Za to se koristi kara
  - eng. backslash
- Na primer, za regularni izraz

  1\+1=2

  u tekstu

  50–8*6+1+1=2+2

  pronalazi se podniska

  50–8*6+**1+1=2**+2

Ostali meta-karakteri kao literali:

\\ \^ \| * \(\)
\. \$ \? \+ \[\{

Specijalni karakteri

- Najčešće u upotrebi su:
  - novi red **\n**
  - tab **\t**
- Na primer, za regularni izraz

  četvrtak\n

  i tekst koji se pretražuje

  Danas je četvrtak, ne volim četvrtak
  A ti?

  pronađena podniska je

  Danas je četvrtak, ne volim **četvrtak**
  A ti?

Specijalni karakteri

- Najčešće u upotrebi su:
  - novi red **\n**
  - tab **\t**
- Na primer, za regularni izraz

  četvrtak\n

  i tekst koji se pretražuje

  Danas je četvrtak, ne volim četvrtak
  A ti?

  pronađena podniska je

  Danas je četvrtak, ne volim **četvrtak**
  A ti?

Posle ove reči „četvrtak“ sledi zarez, a ne novi red, tako da ta podniska ne odgovara ovom regularnom izrazu!

Opseg pretraživanja regularnih izraza

- Ukoliko se to na određen način ne naglasi, regularni izrazi pronalaze samo prvo pojavljivanje uparene podniske
- Na primer, za regularni izraz

  dan

  i nisku koja se pretražuje

  Danas je dan D, dan za kojim sledi dan koji je taj dan!

  rezultat pretrage je

  Danas je **dan** D, dan za kojim sledi dan koji je taj dan!

Karakterske klase (karakterski skupovi)

- Karakterske klase se koriste upotrebom uglastih zagrada [], u okviru kojih se navedu karakteri
- Tada se ne traži jedan karakter, već bilo koji među tim navedenim između zagrada
- Na primer, regularnom izrazu

  gr[ae]y

  odgovaraju niske

  gray

  grey

  ali ne odgovaraju niske

  graey

  greay

Opseg karakterskih skupova

- Ukoliko se radi o nekom nizu karaktera, na primer svim malim slovima engleske abecede ili svim ciframa, moguće je izbeći navođenje svih karaktera
- Ovo se postiže upotrebom povlake, a navođenjem samo prvog i poslednjeg karaktera iz opsega
- Razni opsezi:

  [0-9] [a-z] [A-Z]

- Moguće su i razne kombinacije opsega i samostalnih karaktera:

  [a-zA-Z] [a-z0-9_] [A-Zabc0-9] [a-x] [0-3] [1-5A-M]

Negacija opsega karakterskih skupova

- Ponekad je potrebno pronaći karakter(e) koji NISU iz određenog skupa
- To je moguće upotrebom simbola ^ neposredno iza otvorene uglaste zagrade
- Na primer,

  [^0-9] [^a-zA-Z0-9] [^aeiouAEIOU] [^\n\t]

Meta-karakteri unutar karakterskih skupova

- Meta-karakteri unutar karakterskih skupova koriste se na već objašnjen način

  [\?] [*\\\.] [^\\] [^\(\)]

- Ukoliko se ne navede jedan od granica opsega, iako meta-karakter, povlaka se može koristiti kao literal

  [-x] [x-] [^-x] [^x-]

Operatori ponavljanja (1)

- Često je potrebno navesti koliko puta se neki karakter, odnosno podniska, ponavlja
- Za tu svrhu postoje operatori
  - `?` nula ili jedno pojavljivanje
  - `*` nula ili više uzastopnih pojavljivanja
  - `+` jedno ili više uzastopnih pojavljivanja

Operatori ponavljanja (2)

- Na primer, regularnom izrazu

  [Ss]?koncentrisati

  odgovaraju niske

  Skoncentrisati skoncentrisati koncentrisati

- Regularnom izrazu

  ab*c

  odgovaraju niske

  ac abc abbbc abbbbbbbbbbc

- Regularnom izrazu

  ab+c

  odgovaraju niske

  abc abbbc abbbbbbbbbc

Operatori ponavljanja (2)

- Na primer, regularnom izrazu

  [Ss]?koncentrisati

  odgovaraju niske

  Skoncentrisati skoncentrisati konc

- Regularnom izrazu

  ab*c

  odgovaraju niske

  ac abc abbbc abbbbbbbbbbc

- Regularnom izrazu

  ab+c

  odgovaraju niske

  abc abbbc abbbbbbbbbc

Zvezda je „pohlepna“ (eng. greedy), što znači da pokušava da obuhvati što dužu sekvencu koja odgovara regularnom izrazu

Određen broj ponavljanja

- Postoji i operator {} koji omogućava da se navede tačan broj ponavljanja, ili minimalan i maksimalan broj ponavljanja
- Na primer,

[a-z]{0,1}	[a-z]?
[a-z]{0,}	[a-z]*
[a-z]{1,}	[a-z]+
[1-9][0-9]{3}	brojevi od 1000 do 9999
[1-9][0-9]{2,4}	brojevi od 100 do 99999

Prečice karakterskih klasa

- Da se ne bi pisali celi opsezi, moguće je upotrebiti skraćene verzije onih koji su najčešće u upotrebi
  - \d [0-9]
  - \w [A-Za-z0-9_]
  - \s [\t\n]
  - \D [^\d]
  - \W [^\w]
  - \S [^\s]

Tačka

- Tačka **.** je jedan od najčešće korišćenih meta-karaktera
- Tačka predstavlja TAČNO JEDAN karakter i to bilo koji karakter
  - osim znaka za novi red
- Treba posebno obratiti pažnju prilikom upotrebe tačke, zato što je ona često uzrok grešaka
- Na primer, regularni izraz

  \d\d.\d\d.\d\d

  pronalazi niske

  05/04/17 05/04/17 05.04.17

  ali i

  12345!78 33*45*45 12345678

Početak niske i kraj niske

- Ponekad je potrebno naći karaktere koje se nalaze baš na početku, odnosno na kraju niske
- Za tu svrhu koriste se karakteri „sidra“
  - **^** početak
  - **$** kraj
- Na primer, regularni izraz

  ^a+

  u tekstu

  aaaa baaac bbba abb

  pronalazi

  aaaab baaac bbba **a**bb

Granice niski

- Granice niski su **\b**
  - pozicija pre prvog karaktera u nisci (ako počinje sa [a-zA-Z])
  - pozicija posle poslednjeg karaktera u nisci (ako počinje sa [a-zA-Z])
  - između slovnog i numeričkog karaktera u nisci
- Karakteri unutar niske **\B**
- Granice niske omogućavaju pretragu celovitih reči
- Na primer, za regularni izraz

  \bbanana\b

  i tekst

  bananamen banana bananabanana

  pronalazi se

  bananamen **banana** bananabanana

Alternacija

- Alternacija je veoma slična karakterskim klasama
- Navedu se sve opcije (ali u ovom slučaju ne nužno karakteri), koje se pritom razdvoje simbolom **|**
- Na primer, regularni izraz

  \b(pas|mačka|miš|veverica)\b

  traži bilo koju od četiri navedene reči
- Alternacijom se mogu izraziti i karakterske klase

  [abcd] a|b|c|d

  [02468]+ 0+|2+|4+|6+|8+

Ponovna upotreba uparene niske

- Moguće je pronaći nisku, a onda tu nisku nekako upotrebiti u ostatku regularnog izraza
- Ovo se postiže tako što se regularni izraz stavi u oble zagrade i on se na taj način „pamti“
- Ako se pet delova regularnog izraza nalazi u zasebnim zagradama, onda se oni koriste kao \1 \2 \3 \4 i \5
- Na primer, regularni izraz

  ([a-c])x\1x\1

  pronalazi niske

  axaxa bxbxb cxcxc