Algebra
visoko učilište

Izdvajanje sadržaja i regularni izrazi

www.racunarstvo.hr

Izdvajanje sadržaja

- može biti vrlo čest zadatak sa kojim čak i krajnji korisnici "zabavljaju" sistem administratora
- izdvajanje podataka koji administratora zanimaju, ma kakav ulazni format bio - tekstualna, PDF, CSV, XLS, DOC datoteka, postaje sve češći zadatak

PDF sadržaji

- konverzija u slike

[root@OOS2 ~]# convert datoteka.pdf datoteka.jpg

- konverzija u PS datoteke

[root@OOS2 ~]# pdf2ps datoteka.pdf

- u tekstualne datoteke

[root@OOS2 ~]# pdftotext datoteka.pdf

- u HTML dokumente

[root@OOS2 ~]# pdftohtml datoteka.pdf

XLS(X) sadržaji

- dio paketa *gnumeric*, komanda *ssconvert*, može raditi konverziju XLS(X) datoteka u CSV:

```
[root@OOS2 ~]# ssconvert datoteka.xlsx datoteka.csv
```

- komanda *ssconvert* prihvaća četiri izlazna tipa - Excel95-2000 format, Excel 2007 format, PDF i CSV

Algebra
visoka škola
za primijenjeno
računarstvo

DOC(X) sadržaji

- za konverziju .doc datoteka, vrlo je koristan alat *antiword*:

```
[root@OOS2 ~]# antiword datoteka.doc > datoteka.txt
```

- za konverziju .docx datoteka, vrlo dobra aplikacija je *docx2txt*, za koju se treba prvo snimiti izvorni kod, zakompajlirati pa onda koristiti:

```
[root@OOS2 ~]# docx2txt.sh datoteka.docx
```

Algebra
visoka škola
za primijenjeno
računarstvo

Konverzija više slika u PDF datoteku

- korištenjem komande *convert*, možemo konvertirati i proizvoljan broj slika u jednu PDF datoteku
- pripaziti na opcije (različiti ulazni formati slika imaju različite opcije)

```
[root@OOS2 ~]# convert -quality 100 -density 100 *.jpg datoteka.pdf
```

- zlazna datoteka će biti prihvatljive kvalitete za čitanje uz napomenu da primarna ideja PDF datoteka nije pohrana samo slika

Regularni izrazi

- Osnovni principi:

1. Kada radimo sa rasponom brojeva, pretražujemo mjesto po mjesto (jedinice po jedinice, desetice po desetice, stotice po stotice itd.).
2. Ako tražimo znakove (*stringove*) koji u sebi uključuju specijalne znakove, potrebno je primjeniti *escaping*, proceduru koju ćemo objasniti nekim od primjera koji slijede. Primjeri za takve znakove su [\ ^ $. | ? * + ( )
3. Korištenje znakova kao što su ., ? i * mogu (i trebaju) davati različite rezultate u regularnim izrazima, stoga ih trebamo koristiti sa eksplicitnom svrhom
4. Velika prednost je mogućnost korištenja logičkog operatora kao što je ILI (|), pošto nam to jako skraćuje konačni regularni izraz

Algebra
visoka škola
za primijenjeno
računarstvo

Dodatni znakovi

- Možemo kroz regularne izraze tretirati i znakove kao što su:

\n - nova linija

\t - tab

\w - bilo koji alfanumerički znak, kao [a-zA-Z0-9_]

\W - bilo koji non-alfanumerički znak, kao [^a-zA-Z0-9_]

\d - bilo koji broj, kao [0-9]

\D - bilo koji non-broj, kao [^0-9]

\s - bilo koji *whitespace* znak - space, tab, newline, itd.

\| - traženje *pipe* znaka

\[- traženje lijeve uglate zagrade

Algebra
visoka škola
za primijenjeno
računarstvo

Primjeri, I

- IP adresa

^((([2][5][0-5]|([2][0-4]|[1][0-9]|[0-9])?[0-9])\.){3})([2][5][0-5]|([2][0-4]|[1][0-9]|[0-9])?[0-9])$

- e-mail adrese

^[A-Za-z0-9._-]+@[[A-Za-z0-9.-]+$

- US telefonski brojevi u 1-(XYZ)-XYZ-XYZW formatu

/^(1[-\s.])?(\()?\d{3}(?(2)\))[-\s.]?\d{3}[-\s.]?\d{4}$/

Algebra
visoka škola
za primijenjeno
računarstvo

Primjeri, II

- e-mail adresa

{^[A-Za-z0-9._-]+@[[A-Za-z0-9.-]+$}

- traženje datuma formata DD-MM-YYYY

\d{2}-\d{2}-\d{4}

- u US formatu (YYYY-MM-DD)

\d{4}-\d{2}-\d{2}

Korisna opcija

- Uključivanje opcije *highlight* kod komande *grep* (da nam sa bojom označi rezultate naše pretrage)

grep --color regular_expression sadržaj

Algebra
visoka škola
za primijenjeno
računarstvo