Big Data & Data Science

Infraestrutura Computacional
Parte 1: Linux e Shell

DSBD

Filtros e Expressões Regulares

Filtros

Princípio Unix (e Linux): todo comando deve fazer apenas uma coisa, e pode-se facilmente concatenar resultados de diferentes comandos

Um filtro é um programa de linha de comando que recebe dados em forma de texto e transforma estes dados

- Concatenando comandos, podemos aumentar sua aplicabilidade (mais sobre isso na próxima aula)
- Veremos alguns filtros de maneira individual

Filtros

Comando	Significado
`cat`	mostra o conteúdo
`tac`	mostra o conteúdo do fim para o início
`head, tail`	mostra as primeiras/últimas linhas
`nl`	numera linhas
`diff`	mostra diferença entre dois arquivos
`wc`	conta linhas, palavras e caracteres (*word count*)
`cut`	separa colunas
`uniq`	remove linhas adjacentes duplicadas
`sort`	ordena os dados
`sed`	busca padrões e substitui (*stream editor*)

Exemplos

```
Fred apples 20
Susy oranges 5
Susy oranges 5
Mark watermellons 12
Robert pears 4
Terry oranges 9
Lisa peaches 7
Susy oranges 12
Mark grapes 39
Mark grapes 39
Anne mangoes 7
Greg pineapples 3
Oliver rockmellons 2
Betty limes 14
```

Arquivo **`dados.txt`**

Exemplos

Comando
`cat dados.txt`
`tac dados.txt`
`head -4 dados.txt`
`nl -s '. ' -w 5 dados.txt`
`diff dados.txt dados.txt.old`
`wc -l dados.txt`
`cut -f 1 -d ' ' dados.txt`
`cut -f 1,2 -d ' ' dados.txt`
`sort dados.txt`
`sed 's/oranges/bananas/g' dados.txt`

Expressões Regulares

Que '`^.{5}$`' é isso?

É uma linguagem para descrever padrões de dados

- Utilizadas por diversos comandos e linguagens de programação
- Os caracteres podem ter significado diferente dos *wildcards*
- Exemplos com o comando **`egrep`** (**`grep -E`**)
  - Expressão regular entre aspas simples (`' '`)

Expressões Regulares

Símbolo	Significado
`.`	(ponto) - um caractere simples
`?`	o caractere anterior é casa 0 ou 1 vez
`*`	o caractere anterior casa 0 ou mais vezes
`+`	o caractere anterior casa 1 ou mais vezes
`{n}`	o caractere anterior casa exatamente `n` vezes
`{n,m}`	o caractere anterior casa pelo menos `n` e não mais que `m` vezes

Expressões Regulares

Símbolo	Significado
`[agd]`	o caractere é um dos incluídos entre colchetes
`[^agd]`	o caractere não é um dos incluídos entre colchetes
`[c-f]`	o caractere é um no intervalo entre c e f
`()`	permite agrupar diversos caracteres como se fossem um
`|`	(pipe symbol) - operação lógica OU
`^`	casa com início de linha
`$`	casa com final de linha

Exemplos com egrep

```
egrep 'mellon' dados.txt
egrep -n 'mellon' dados.txt
egrep -c 'mellon' dados.txt
egrep '[aeiou]{2,}' dados.txt
egrep '2.+' dados.txt
egrep '2$' dados.txt
egrep 'or|is|go' dados.txt
egrep '^[A-K]' dados.txt
```

Referências

- Anatomy of the Linux kernel
- Linux OS Tutorial
- Introduction to UNIX
- Introduction to Linux
- Ryans Linux Tutorial
- Ryans Regular Expressions