正则表达式

-- 鸟哥第 12 章 347-362 页

授课教师：李正华

http://ir.hit.edu.cn/~lzh

什么是正则表达式

- regular expression
- 字符串的一种表示方法（语言）。
- 用于字符串的处理：查找、删除、删除特定模式（ patern ）的字符串
- 编译语言中的正则语言（ regular language ）

和 bash 下的通配符完全不同。通配符（wildcard）是 bash 提供的一个功能，而正则表达式是字符串处理的一种表示方式

正则表达式的用途

- 从大的文本文件中找出感兴趣的（有问题的）内容进行分析。
- 垃圾邮件过滤
- 编程时迅速查找，替换变量名称等
- …

- 很多工具都支持正则表达式： vi, grep, awk, sed, find 等

find 对正则表达式的支持

- find /bin/ -regex '.*/e.*'
- find /bin -regex '/bin/.....'

File name matches regular expression pattern. This is a match on **the whole path, not a search**. For example, to match a file named ‘./fubar3’, you can use the regular expression ‘.*bar.’ or ‘.*b.*3’, but not ‘f.*r3’. The regular expressions understood by find are **by default Emacs Regular Expressions**, but this can be changed with the -regextype option.

-regextype
emacs (this is the default), posix-awk, posix-basic, posix-egrep and posix-extended.

正则表达式的分类

- 根据严谨度可分为
  - 基础正则表达式
  - 扩展正则表达式

以 grep 为例，学习基础正则

- 简单用法
  - dmesg | grep 'eth'
  - dmesg | grep -n –color=auto 'eth'
  - dmesg | grep -n -A3 -B2 –color=auto 'eth'

- grep 以整行为单位进行字符串的对比和选取

以 grep 为例，学习基础正则

- 查找特定字符串
  - grep -n 'the' regular_express.txt
  - grep -vn 'the' regular_express.txt
  - grep -in 'the' regular_express.txt

以 grep 为例，学习基础正则

- 利用中括号 [] 查找集合字符
  - grep -n 't[ae]st' regular_express.txt
  - grep -n 'oo' regular_express.txt
  - grep -n '[^g]oo' regular_express.txt
  - grep -n '[^a-z]oo' regular_express.txt
  - grep -n '[0-9]' regular_express.txt
  - grep -n [^[:lower:]]oo' regular_express.txt
  - grep -n '[^[:digit:]]' regular_express.txt

以 grep 为例，学习基础正则

- 行首和行尾字符 ^$
  - grep -n '^the' regular_express.txt
  - grep -n '^[a-z]' regular_express.txt
  - grep -n '\.$' regular_express.txt
  - grep -n '^$' regular_express.txt
  - grep -v '^$' /etc/syslog.conf | grep -v '^#'

以 grep 为例，学习基础正则

- 任意一个字符 “.” 与重复字符 “*”
  - grep -n 'g..d' regular_express.txt
  - grep -n 'ooo*' regular_express.txt
  - grep -n 'goo*g' regular_express.txt
  - grep -n 'g*g' regular_express.txt
  - grep -n 'g.*g' regular_express.txt
  - grep -n '[0-9][0-9]*' regular_express.txt
  -

以 grep 为例，学习基础正则

- 限定连续 RE 字符范围 “ {}”
  - grep -n 'o\{2\}' regular_express.txt
  - grep -n 'go\{2,5\}g' regular_express.txt
  - grep -n 'go\{2,\}g' regular_express.txt

基础正则总结

- 行首 ^
- 行尾 $
- 任意一个字符 .
- 转义字符（特殊字符转为普通字符）\
- 前一个 RE 字符重复零次到无穷多次 *
- 集合中任意一个字符 []
- 此范围的集合的 [x-y]
- 任意非集合中的字符 [^]
- 前一个字符重复的范围 {}

扩展正则表达式

- + ：前一个 RE 字符重复一次到无穷
  - egrep -n 'go+d' regular_express.txt
- ？：前一个 RE 字符出现 0 次或 1 次
  - egrep -n 'go?d' regular_express.txt
- | ：以“或 or” 的方式找出数个字符串
  - egrep -n 'gd|good' regular_express.txt
  - egrep -n 'gd|good|dog' regular_express.txt
-

扩展正则表达式

- () ：找出“组”字符串
  - egrep -n 'g(la|oo)d' regular_express.txt
- ()+ ：字符串重复
  - echo 'AxyzxyzxyzxyzC' | egrep 'A(xyz)+C'