Module01-03
Linux 基础：正则表达式基础

Linux 基础 - 正则表达式基础

- 常用 Linux 命令
- 深入了解 bash
- ➔ 正则表达式基础
- find、 grep、 sed、 awk

正则表达式基础 - 文件名 vs 样式

- Shell 命令行中的文件通配和正则表达式中样式的区别

  如下例：

  ```
  grep [A-Z]* c[a-c].sql
  ```

  会被 shell 解释为类似下面的结果，[A-Z]* 匹配为文件名：

  ```
  grep Apple.log Que.txt c[a-c].sql
  ```

- 正则表达式的样式要使用双引号或单引号，如下所示：

  ```
  grep '[A-Z]*' c[a-c].sql
  ```

  这样才是在 c[a-c].sql 等文件中找含有大写字母的行

  - 注意：样式可以使用双引号和单引号，但单引号更安全，因为单引号不会解释如变量引用（如 ${}）等特殊字符

正则表达式基础 - 元字符

- 搜索样式的元字符
  - 基本元字符

字符	匹配样式
.	除换行符（\n）外所有单个字符
^	标志行首，如： ^Tiger 匹配以 Tiger 开头的行（ T 为行首字符）
$	标志行尾，如： ing$ 匹配以 ing 结尾的行； ^hello$ 只匹配整行内容为 hello 的行； ^$ 匹配空行
[ ]	匹配或不匹配任意出现在 [ ] 中的单个字符：如 [abd] 匹配 a b c 三个中的任意一个， [^abd] 匹配 a b d 三个字符以外的任意字符， [1-5] - 表示范围，匹配数字 1 到 5
*	字符或子表达式出现的次数： 0 到多次，如 ab* 匹配 a,ab,abb,...
+	字符或子表达式出现的次数： 1 到多次，如 ab+ 匹配 ab,abb,... 但不匹配 a
?	字符或子表达式出现的次数： 0 或 1 次，如 ab? 只匹配 a,ab
{n,m} 或 \{n,m\}	表示前面的字符或子表达式出现的次数： n 到 m 次，如 ab{2,4} 匹配 abb,abbb,abbbb 形式 2： {n} 出现 n 次，如 ab{2} 只匹配 abb 形式 3： {n,} 最少出现 n 次，如 ab{2,} 匹配 abb,abbb,...

正则表达式基础 - 元字符

- 搜索样式的元字符（续）
  - 基本元字符（续）

字符	匹配样式
\	转义字符，如 \n 代表新行符， \. 代表字符 .（这里 . 被转义，已不是元字符）
\(\)	子表达式，如 (ab[5-9])+ 括号内的 ab[5-9] 是一个子表达式
\n	反向引用 (back reference)， n 为数字，指代对应位置的子表达式，如 (ab[5-9])+:(Br[aeiou])?\1 此处的 \1 即等同于 (ab[5-9]) 子表达式，整个表达式等同于: (ab[5-9])+:(Br[aeiou])?(ab[5-9]) 如: (ab+?).*?\1 可以匹配 abb8udabb
\< \>	单词的起、止边界，如 \<Tiger 匹配到单词 Tiger 开始位置
|	匹配 | 之前或之后的样式，如 (abd|p) 匹配 abd 和 abp
( )	字表达式或表达式组

- 搜索样式的元字符（续）
  - POSIX 字符类型

    如 alpha 表示字母 [a-zA-Z] 中任意一个，样式格式：`[[:alpha:]]`

    注意：都是匹配单个字符

类别	匹配	类别	匹配
alnum	字母和阿拉伯数字 [a-zA-Z0-9]	lower	小写字母 [a-z]
alpha	字母 [a-zA-Z]	print	可打印字符
blank	空格或 tab	punct	标点符号，如，
cntrl	控制字符	space	空白字符
digit	十进制数字 [0-9]	upper	大写字母 [A-Z]
graph	除空白字符外的可打印字符	xdigit	十六进制字符 [0-9a-fA-F]

正则表达式基础 - 元字符

- 搜索样式的元字符（续）
  - 几个特殊转义字符

字符	匹配样式
\b	单词边界，同 \< 和 \>，如：\bthe\b 匹配 the first 中的‘the’，但不匹配 there 中的 'the'
\B	单词内部匹配，匹配 2 个构成单词的字符之间，如：ver\B 匹配 version 中的 'ver'，但不匹配 server 中的 'ver'
\s	空白字符，等同于 [[:space:]]
\S	非空白字符，等同于 [^[:space:]]
\w	构成单词的字符 [a-zA-Z0-9_]，等同 [[:alnum:]_]
\W	与上面相反，等同 [^a-zA-Z0-9_]，等同 [^[:alnum:]_]
\`	通常为字符串的开始
\'	通常为字符串的结束

正则表达式基础 - 元字符

- 搜索样式的元字符（续）
  - 搜索样式的元字符简单（不严格）归类：
    - 转义符：\
    - 反向引用，如：\1 \2
    - 控制数量，如：? * + {n} {n,m} {n,}
    - 指定位置，如：^ $ \< \> \b \B \` \'
    - 其它匹配单个字符的样式

正则表达式基础 - 元字符

- 替换样式的元字符
  - 基本替换元字符

字符	匹配样式
\	转义随后的元字符
\n	类似搜索元字符中的 \n （ n 为数字 ）
&	将搜索匹配到的内容作为替换的内容一部分
\u	将匹配样式的第一个字母转为大写
\U	将匹配样式的所有字母转换为大写
\l	将匹配样式的第一个字母转为大写
\L	将匹配样式的所有字母转换为大写
\e	关闭前一次的 \u 或 \l
\E	关闭前一次的 \U 或 \L

正则表达式基础 - 元字符的支持列表

- 元字符的支持列表

  P 代表 POSIX

	ed	ex	vi	sed	awk	grep	egrep
.	y	y	y	y	y	y	y
*	y	y	y	y	y	y	y
^	y	y	y	y	y	y	y
$	y	y	y	y	y	y	y
\	y	y	y	y	y	y	y
[ ]	y	y	y	y	y	y	y
\(\)	y	y	y	y		y	
\n	y	y	y	y		y	
{ }					y(P)		y(P)
\{ \}	y			y		y	
\< \>	y	y	y				

正则表达式基础 - 元字符的支持列表

- 元字符的支持列表

	ed	ex	vi	sed	awk	grep	egrep
+					y		y
?					y		y
( )					y		y
|					y		y

正则表达式基础 - 总结

- Linux 系统中使用正则表达式的工具
  - grep 、 egrep 、 awk 、 sed
  - vi 、 emacs 等编辑器
  - perl 、 python 、 tcl 等语言
- 由于历史原因，各工具使用的样式语法不完全相同，支持的程度也不一致，使用的时候需多参考各自的文档
- 本单元内容仅涉及正则表达式的很小一部分，仅为使用 grep 、 sed 、 awk 、 vi 等工具做准备
- 在 C++ boost.regex 课程中，我们将更深入的研究正则表达式的规则与使用