Регулярные выражения UNIX

Регулярные выражения (regular expressions) – система синтаксического разбора текстовых фрагментов по формализованному шаблону, основанная на системе записи образцов для поиска. Образец (pattern), задающий правило поиска, по-русски также иногда называют шаблоном или маской. Регулярные выражения произвели прорыв в электронной обработке текста в конце XX века.

Сейчас регулярные выражения используются многими текстовыми редакторами и утилитами для поиска и изменения текста на основе выбранных правил. Многие языки программирования уже поддерживают регулярные выражения для работы со строками. Набор утилит (включая редактор sed и фильтр grep), поставляемых в дистрибутивах UNIX, одним из первых способствовал популяризации понятия регулярных выражений.

Справочная страница (man re_format) содержит исчерпывающее описание регулярных выражений, соответствующих стандарту POSIX 1003.2.

Символы базовых регулярных выражений

Звездочка -- * --

Означает любое количество символа в строке, предшествующего “звездочке”, в том числе и нулевое число символов.

Точка -- . --

Означает не менее одного любого символа

Символ -- ^ --

Означает начало строки, но иногда, в зависимости от контекста, означает отрицание в регулярных выражениях.

```
$ grep '^s' /etc/passwd
 sshd:*:22:22:Secure Shell Daemon:/var/empty:/usr/sbin/nologin
```

```
 smmsp:*:25:25:Sendmail Submission
 User:/var/spool/clientmqueue:/usr/sbin/nologin
```

Знак доллара -- $ --

В конце регулярного выражения соответствует концу строки.

```
$ grep 'sh$' /etc/passwd
 root:*:0:0:Charlie &:/root:/bin/csh
```

Квадратные скобки -- [...] --

Предназначены для задания подмножества символов. Квадратные скобки, внутри регулярного выражения, считаются одним символом, который может принимать значения, перечисленные внутри этих скобок. Метасимвол ^ означает отрицание множества

```
$ grep '^[rt]' /etc/passwd
 root:*:0:0:Charlie &:/root:/bin/csh
 toor:*:0:0:Bourne-again Superuser:/root:
 tty:*:4:65533:Tty Sandbox:/:/usr/sbin/nologin
```

Обратный слеш -- --

Служит для экранирования специальных символов, это означает, что экранированные символы должны интерпретироваться буквально, т.е. как простые символы (в некоторых случаях наоборот).

Экранированные "угловые скобки" -- <...> --

Отмечают границы слова (не работает в sed).

```
$ grep 'var' /etc/login.conf
 :setenv=MAIL=/var/mail/
 $,BLOCKSIZE=K,FTP_PASSIVE_MODE=YES:
```

```
 :nologin=/var/run/nologin:
 # Russian Users Accounts. Setup proper environment variables.
 # :setenv=MAIL=/var/mail/$,BLOCKSIZE=K:
 # :nologin=/var/run/nologin:

 $ grep '' /etc/login.conf
 :setenv=MAIL=/var/mail/
 $,BLOCKSIZE=K,FTP_PASSIVE_MODE=YES:
 :nologin=/var/run/nologin:
 # :setenv=MAIL=/var/mail/$,BLOCKSIZE=K:
 # :nologin=/var/run/nologin:
```

Экранированные "круглые скобки" -- ( ) --

Предназначены для выделения групп регулярных выражений. Они полезны при использовании с оператором “|” и при извлечении подстроки.

```
$ grep 'daily|weekly' /etc/crontab
 # Perform daily/weekly/monthly maintenance.
 1 3 * * * root periodic daily
 15 4 * * 6 root periodic weekly
 $ grep 'periodic (daily|weekly)' /etc/crontab
 1 3 * * * root periodic daily
 15 4 * * 6 root periodic weekly
 $ ls /usr/bin | sed 's/(.*)/rm 1/'
 rm CC
 rm Mail
 rm addftinfo
 ...
 $ ls /usr/bin | sed n
 's/(^a.*)/rm 1/p'
 rm addftinfo
 rm addr2line
 rm afmtodit
 ...
 $ cat > catalog.txt
 petrof ivan 2345678
 ivanof sidor 2145678

 $ sed 's/(.*) .* (.*)/1 2/' catalog.txt
 petrof 2345678
 ivanof 2145678
```

Экранированные "фигурные скобки" -- { } --

Задают число вхождений предыдущего выражения.

```
$ grep '(ro.*){2}' /etc/passwd
 root:*:0:0:Charlie &:/root:/bin/csh
 daemon:*:1:1:Owner of many system
 processes:/root:/usr/sbin/nologin
```

Классы символов POSIX.

[:class:] это альтернативный способ указания диапазона символов.

```
$ grep '<[[:alpha:]]{X}>' /etc/login.conf
 :ignorenologin:
 # :ignorenologin:
 # :maxmemorysizecur=128M:
 # :refreshperiod@:
 # :refreshperiod@:

 $ grep '<[AZaz]{X}>' /etc/login.conf
```

Заменяем в файле catalog.txt пробел на TAB

```
$ sed 's/(.*)[[:space:]].*[[:space:]](.*)/1 2/' catalog.txt
 petrof 2345678
 ivanof 2145678
```

Символы расширенных регулярных выражений

Многие символы экранируемые в базовых выражениях – () {} | – но не – <> – используются без экранирования.

Знак вопроса -- ? --

Означает, что предыдущий символ или регулярное выражение встречается 0 или 1 раз.

```
$ grep -E '^r?o' /etc/passwd
 root:*:0:0:Charlie &:/root:/bin/csh
 operator:*:2:5:System &:/:/usr/sbin/nologin
```

Знак "плюс" -- + --

Указывает на то, что предыдущий символ или выражение встречается 1 или более раз (добавляем произвольное количество символов разделителей в файл catalog.txt).

```
$ sed -E 's/([[:alpha:]]+)[[:space:]]+.*[[:space:]]+
 ([[:alpha:]]*)/1 2/' catalog.txt
 petrof 2345678
 ivanof 2145678
```