Ficha 4 – Expressões Regulares

Tópicos abordados:

– Expressões regulares:
- Na linha de comandos;
- No PERL:
  - Caracteres especiais;
  - Quantificadores;
  - Grupo de caracteres;
  - Delimitadores;
  - Opções;
  - Precedências;
  - Exemplos.

– Exercícios;
- Documentação;
- Bibliografia.

Duração prevista: 2 aulas

NOTA: em http://files.meetup.com/120908/Regexp%20Quick%20Reference.pdf encontrará uma PERL Quick Reference Card orientada para as expressões regulares.

1. Expressões regulares (*REGEX – REGular EXpressions*)

1.1. Expressões regulares na linha de comandos

A execução do comando **grep** efetua a pesquisa de todas as linhas de um ficheiro (ex: "/etc/passwd") que contém a palavra **root**, da seguinte forma:

$ grep "root" /etc/passwd

root:x:0:0:root:/root:/bin/bash

Neste caso, a expressão a procurar foi uma palavra (root), mas poderemos também especificar uma expressão mais complexa. Com o uso do mesmo comando **grep** e alterando a expressão de procura, podemos então obter resultados mais latos, mais detalhados ou específicos, como por exemplo:

$ grep "so.*turno" /etc/passwd

so*d1:x:1001:1007:SO_****turno****_PL1_diurno,,,,:/home/sod1:/bin/bash*

so*d2:x:1002:1008:SO_****turno****_PL2_diurno,,,,:/home/sod2:/bin/bash*

so*d3:x:1003:1009:SO_****turno****_PL3_diurno,,,,:/home/sod3:/bin/bash*

so*d4:x:1004:1010:SO_****turno****_PL4_diurno,,,,:/home/sod4:/bin/bash*

so*d5:x:1005:1011:SO_****turno****_PL5_diurno,,,,:/home/sod5:/bin/bash*

so*n1:x:1006:1012:SO_****turno****_PL1_nocturno,,,,:/home/son1:/bin/bash*

so*n2:x:1007:1013:SO_****turno****_PL2_nocturno,,,,:/home/son2:/bin/bash*

so*n3:x:1008:1014:SO_****turno****_PL3_nocturno,,,,:/home/son3:/bin/bash*

Neste caso, o comando grep vai encontrar expressões que obedecem ao critério especificado, ou seja, **encontra palavras** nas linhas do ficheiro **/etc/passwd** que comecem por **so**, e que de seguida tenham qualquer caracter (.) zero ou mais vezes (*), e que de seguida apresentem a sequência de caracters "**turno"**.

1.2. Expressões regulares no PERL

As expressões regulares (REGEX) têm um papel fundamental em PERL, e são usadas sempre se que deseja trabalhar com padrões de texto, substituição, etc.

O código seguinte apresenta exemplo de **pesquisa**[1] ("$a =~ **m**/ola") e de **substituição** ("**s**/ola/bom dia/" – substituição no conteúdo da variável $a de "ola" para "bom dia"). Finalmente, o operador PERL "tr" é similar ao utilitário "tr".

```
# Verifica se a string inclui a expressão ola
if ($a =~ m/ola/){ print "ola";  }

# Verifica se a string não inclui a expressão ola
if ($a !~ m/ola/){ print "sem ola";  }

# Substituir "ola" por "bom dia"
$a =~ s/ola/bom dia/;

# Alterar a letras minúsculas para maiúsculas
$a =~ tr/a-z/A-Z/;
```

O formato de uma expressão regular é 'sempre' **/expressões-regulares/**, excepto se:

^	Usado no início do padrão obriga a próxima expressão a verificar-se no início da string (ou da linha se usado o modificador 'm') Usado no **início** após o meta-carácter '[' corresponde à **negação** de todas a expressões até ao meta-carácter']', isto é, indica todos os carateres que não pertencem ao grupo definido por [...]
[]	Permite definir um conjunto de caracteres, que podem ser verificados (ou não, caso se inicie por '^'). Por exemplo, **[a-z,A-E]**
$	Usado no final do padrão obriga a expressão anterior a verificar-se no final da string (ou da linha se usado o modificador 'm')
.	Significa qualquer carácter excepto o newline '\n'. Incluí também o newline se for usado o modificador 's'
|	Permitir definir expressões alternativas que devem ser verificadas. Funciona pois como um "or".
( )	Expressões entre parêntesis corresponde à captura, agrupando expressões. Após execução da operação as variáveis pré-definidas $1,$2,$..,$9 contém o valor da string verificada (capturada) no 1º,2º,..º,9º grupo respetivamente.

[1] A pesquisa numa REGEX é indicada através de "m" que corresponde à palavra Anglo-Saxónica "match".

1.2.1. Caracteres especiais

\ → É o carácter de escape para testar caracteres que são do ponto de vista das expressões regulares meta-caracteres
\s → um 'white-space', exemplo: um espaço ou um tab
\w → um alfanumérico incluindo o '_'
\d → um digito
\S → qualquer carácter que não seja um 'white-space'
\W → qualquer carácter que não seja um alfanumérico ou '_'
\D → qualquer carácter que não seja um digito

1.2.2. Quantificadores

? → zero ou uma ocorrência
+ → uma ou mais ocorrências
* → zero ou mais ocorrências
{n,} → n ou mais ocorrências
{n,m} → entre n e m ocorrências
{n} → exatamente n ocorrências

Exemplos:

m/Si+E?s/ # Identifica um **S** seguido por um ou mais "**i**" seguido por um ou nenhum "**E**" seguido por um "**s**". (Um resultado possível seria a palavra **Sistemas**)

Os quantificadores *****, **+**, e **?** são ditos "gananciosos" (*greedy*) identificando o maior número de caracteres possível:

$a = "INICIO xxxxxxxxx FIM";

$a =~ s/x+/XPTO/; # devolve: INICIO XPTO FIM

Isto é, toda a sequência "xxxxxxxxx" foi especificada por "x+" e substituída por XPTO

Exercício 1

Recorrendo a expressões regulares, elabore a função PERL "IsNumeric1", que receba como parâmetro de entrada a expressão a validar e devolva: 1 se a expressão for um número e 0 se a expressão não for um número.

Nota: para já devem ser apenas avaliados números inteiros positivos, ou seja, não deve entrar para validação nem o caracter "-" nem o caracter "+".

Exercício 2

Com base na função "IsNumeric1", escreva a função "IsNumeric2", que possa avaliar todos os números inteiros (positivos e negativos).

1.2.3. Grupo de caracteres

- Para especificar, uma gama de caracteres usam-se **"[ ]"** (parêntesis rectos)

  Exemplos

  /[abc]/

  "Verdadeiro" para qualquer string que contenha (pelo menos) um dos caracteres "a" "b" "c"

  /[0123456789]/

  "Verdadeiro" para qualquer string que contenha um algarismo

- Caso se pretenda especificar **"]"** no grupo deve-se empregar a barra de escape (\), ou em alternativa colocá-lo como o primeiro carácter do grupo:

  /[abc\]]/ # inclusão de "]" na gama de caracteres

  /[]abc]/ # idem

- **"-"** para especificar, gama de caracteres :

  /[0123456789]/ # Qualquer dígito

  /[0-9]/ # idem

- Para especificar **"-",** na lista, coloca-se **"\"** antes, coloca-se **"-"** no início ou no fim:

  /[X\-Z]/ # X, -, Z

  /[XZ-]/ # X, Z, -

/[-XZ]/ # -, X, Z

- Mais, alguns exemplos:

  /[0-9\-]/ # 0-9, ou sinal "menos"

  /[0-9a-z]/ # dígito ou letra (minúscula)

  /[a-zA-Z0-9_]/ # qualquer letra, qualquer digito

- Negação de grupo (símbolo **^)** imediatamente após o parêntesis recto esquerdo. Inverte o efeito do grupo (i.e. identifica qualquer carácter não presente no grupo).

  /[^0123456789]/ # tudo que não seja dígito

  /[^0-9]/ # idem

  /[^aeiouAEIOU]/ # tudo exceto vogais

  /[^\^]/ # tudo exceto o próprio ^

1.2.4. Delimitadores

Precedidas por **\b**, apenas serão encontradas as cadeias de caracteres que estiverem, no início de uma palavra.
Seguidas por **\b**, apenas serão encontradas as cadeias de caracteres que estiverem, no final de uma palavra.
Precedidas por **\B**, serão encontradas apenas as palavras que não forem iniciadas, pela cadeia de caracteres.
Seguidas por **\B**, serão encontradas apenas as palavras que não forem terminadas, pela cadeia de caracteres.

Os exemplos correspondentes são os de (32) a (35).

1.2.5. Opções

g → global;
i → ignorar letras;
\m → interpreta caracteres especiais (tipo \n);

1.2.6. Precedências

1. Precedência: **( )** (parênteses)
2. Precedência: **+ * ? {#,#}** (operadores de agrupamento)
3. Precedência: **abc ^$ \b \B** (caracteres/cadeia de caracteres, início ou término de linha, início ou término de palavras)
4. Precedência: **|** (alternativas)

1.2.7. Exemplos

```
( 1) m/a/ # encontra 'a'
( 2) m/[ab]/ # encontra 'a' ou 'b'
( 3) m/[A-Z]/ # encontra todas as letras maiúsculas
( 4) m/[0-9]/ # encontra números
( 5) m/\d/ # encontra números - como em (4)
( 6) m/\D/ # encontra tudo exceto números
( 7) m/[0-9]\-/ # encontra números ou o sinal de menos
( 8) m/[\[\]]/ # encontra tudo que estiver delimitado por parênteses rectos [ ]
( 9) m/[a-zA-Z0-9_]/ # encontra letras, números ou sinal de sublinhado
(10) m/[\w]/ # encontra letras, números ou sinal de sublinhado - como em (9)
(11) m/[\W]/ # encontra tudo, exceto letras, números e sinal de sublinhado
(12) m/[\r]/ # encontra o sinal de retorno (típico do DOS)
(13) m/[\n]/ # encontra o sinal para quebra de linha
(14) m/[\t]/ # encontra o sinal de tabulação (tab)
(15) m/[\f]/ # encontra o sinal para quebra de página
(16) m/[\s]/ # encontra o sinal de espaço assim como os sinais referidos de (12) a (15)
(17) m/[\S]/ # encontra tudo, exceto sinal de espaço e os de (12) a (15)
(18) m/[äöüÄÖÜ]/ # encontra todos os caracteres com acentuação dupla
(19) m/[^a-zA-Z]/ # encontra tudo que não contiver letras
(20) m/[ab]/s # encontra 'a' ou 'b' também em várias linhas
(21) m/asa/ # encontra 'asa' - também 'casa' ou 'casamento'
(22) m/asa?/ # encontra 'asa', 'casa', 'casamento' e também 'as' e 'asilo'
(23) m/a./ # encontra 'as' e 'ar'
(24) m/a+/ # encontra 'a' e 'aa' e 'aaaaa' - quantos existirem
(25) m/a*/ # encontra 'a' e 'aa' e 'aaaaa' e 'b' - nenhum ou quantos 'a' existirem
(26) m/ca.a/ # encontra 'casa' e 'caça', mas não 'cansa'
```

```
(27) m/ca.+a/ # encontra 'casa', 'caça' e 'cansa'
(28) m/ca.?a/ # encontra 'casa', 'caça' e 'caso'
(29) m/x{10,20}/ # encontra sequências de 10 a 20 'x'
(30) m/x{10,}/ # encontra sequências de 10 ou mais 'x'
(31) m/x.{2}y/ # só encontra 'xxxy'
(32) m/Clara\b/ # encontra 'Clara' mas não 'Clarinha'
(33) m/\bassa/ # encontra 'assa' ou 'assado' mas não 'massa'
(34) m/\bassa\b/ # encontra 'assa' mas não 'assado' e nem 'massa'
(35) m/\bassa\B/ # encontra 'assado' mas não 'assa' e nem 'massa'
(36) m/^Julia/ # encontra 'Julia' apenas no início do contexto da pesquisa
(37) m/Helena$/ # encontra 'Helena' apenas no final do contexto da pesquisa
(38) m/^\s*$/ # encontra linhas constituídas apenas por sinais vazios ou similares
(39) m/$Nome/ # encontra o conteúdo da escalar $Nome
(40) m/asa/s # encontra 'asa', também em várias linhas
(41) m/a|b/ # encontra 'a' ou 'b' - idêntico a m/[ab]/
(42) m/com|sem/ # encontra 'com' e 'descompensar', como também 'sem' e 'semântica'
```

Listagem 1: Expressões regulares

```
#!/usr/bin/perl
use strict;
use warnings;

my $s = 'abcdefcdcd';
my $i = 'X';
my $j = 'W';
#substituição é feita a cada iteração
while ($s =~ s/CD/$i/i){
        print "1: $s\n";
        $i++;
}

#substituição é feita de uma vez só, pois é especificada a opção g
$s = 'abcdefcdcd';
while ($s =~ s/CD/$j/ig){
        print "2: $s\n";
}
```

Listagem 2: Substituições recorrendo a expressões regulares

```
#!/usr/bin/perl
$\ = "\n";
$, = ",";

$s = "Isto é uma string\nEsta é uma String\nEsta é a terceira
linha\n";
$s .= " Esta é a quarta linha\nEsta não é a última frase\n";
$s .= "Esta é a última linha\n";
```

```
print $s;

print "--------------\n";

$s =~ m/^Isto/ && print 'verificou-se m/^Isto/';
$s =~ m/^Esta/ || print 'não se verificou m/^Esta/';
$s =~ m/^Esta/ && print 'verificou-se m/^Esta/';

print "--------------\n";
print $s;
print "--------------\n";

@l = $s =~ m/^Esta/mg and $j = @l, print "verificou-se m/^Esta/mg $j
vezes";

print "--------------\n";

@l = $s =~ m/^[EI]st[ao]/mg and $j = @l, print "verificou-se
m/^[EI]st[ao]/mg $j vezes";

print "--------------\n";

@l = $s =~ m/^Isto|^Esta/mg and $j = @l, print "verificou-se
m/^Isto|^Esta/mg $j vezes\t", @l;

print "--------------\n";

@l = $s =~ m/^Isto|Esta/mg and $j = @l, print "verificou-se
m/^Isto|Esta/mg $j vezes\t", @l;
```

Listagem 3: Expressões regulares

Exercício 3

a) Elabore a função "IsNumeric3" (com base na função "IsNumeric2"), que deve avaliar, agora, todos os números **reais** (positivos e negativos). Faça, também, um teste intensivo à função para ter a certeza que esta não tem nenhum "bug".

Como exemplo, pode chamar a função N vezes no seu código de forma similar à indicada na listagem seguinte:

```
  my @Strings_Teste_L = ("","+","+1","-1.23","5.3","5.",".8","0.6","-
.5");
  foreach my $Teste_S(@Strings_Teste_L){
    printf("Teste:'$Teste_S':%d\n", IsNumeric3($Teste_S));
  }
```

Nota: no caso de aparecer alguma situação que esteja omissa nos testes acima indicados, decida o que a função deve devolver e implemente essa validação.

b) Repita a alínea a), mas em lugar de usar o vetor **@Strings_Teste_L** faça uso de uma lista associativa (hash) denominada **%Strings_Teste_H**, em que cada chave

corresponde a uma string de teste e o respetivo valor corresponde ao resultado esperado do teste.

1.2.8. Grupo e Captura

- Para especificar, um grupo de valores a capturar usam-se **"( )"** (parêntesis curvos). Por exemplo, se quisermos obter os valores de uma data "01-05-2011" em componentes individuais temos que conseguir separar, de forma rápida o "01", o "05" e o "2011". Isto pode ser conseguido como os **"( )"** da seguinte forma:

```
#!/usr/bin/perl -w
use strict;

my $str = "01-05-2011";

$str =~ m/^(\d\d)-(\d+)-(\d{4})$/;
print("$1 \n");   # mostra no terminal: 01
print("$2 \n");    # mostra no terminal: 05
print("$3 \n");   # mostra no terminal: 2011
```

Listagem 4: Expressões regulares

Neste exemplo são utilizadas várias expressões (complementares) para captura dos diversos grupos:

- (\d\d) – captura os primeiros dois dígitos;
- (\d+) – captura 1 ou mais dígitos até que apareça um "-", neste caso serão 2 dígitos;
- (\d{4}) – captura os últimos 4 dígitos.

Os valores capturados por estes grupos vão, respectivamente, para as seguintes variáveis especiais: **$1, $2, $3**.

Nota: não se esqueça que a variáveis $0 também é "especial" mas contém o nome do programa que está a ser executado.

Exercício 4

Elabore a função "IsNumeric4" (com base na função "IsNumeric3"), que deve avaliar, todos os números **reais** (positivos e negativos), que deve descriminar no caso de números com parte decimal, a parte inteira da parte decimal.

Exercícios

1. Construa as expressões regulares que permitam validar as seguintes situações:
    a. Código postal
    b. Nº Telefone 91, 92, 93 e 96
    c. Contas de Mail
    d. Endereços IP
    e. Url's https portugueses

2. Recorrendo somente à linguagem PERL e às expressões regulares, elabore o script "**eitube_num_views.pl**", cujo propósito é o de reportar o número de visionamentos ("views") de cada vídeo constante do canal YouTube do curso de Eng. Informática da ESTG (http://www.youtube.com/cursoeiestg), sendo a informação precedida da data corrente. Por exemplo, quando executado o script produzirá a seguinte saída:

```
20090424_19h24:43
286 views
662 views
```

Nota: Para testar poderá efectuar o *download* da página inicial do site da ESTG, através do comando:

$ wget http://www.youtube.com/cursoeiestg **-O** eitube.html

3. Recorrendo à linguagem PERL e às expressões regulares, elabore a script **"encontra_imagens.pl",** que encontre e efetue a contagem de quantas imagens existem numa página HTML. A script recebe por parâmetro o ficheiro, onde se deve procurar as imagens:

$ ./encontra_imagens.pl index.html

As imagens no HTML são identificadas pela TAG <*img*>.

Nota: Para testar poderá efetuar o *download* da página inicial do site da ESTG, através do comando:

$ wget http://www.estg.ipleiria.pt -O index.html

4. Recorrendo à linguagem PERL e às expressões regulares, elabore a script **"extensoes.pl"** que dada uma diretoria de procura, efectue a listagem **apenas dos ficheiros** existentes nessa diretoria, especificando a extensão de cada um deles. A diretoria será indicada por parâmetro, na chamada da script:

$ ./extensoes.pl $HOME

Note que, um ficheiro com o nome **a.bb.cc.zip**, deverá ser separado com o nome **a.bb.cc,** e com extensão **zip**.

O resultado final terá um aspecto similar a este:

Ficheiro	->	Nome	-> Extensão
exercicio1.pl~	->	exercicio1	-> pl~
exercicio2.pl~	->	exercicio2	-> pl~
extensoes.pl	->	extensoes	-> pl
extensoes.pl~	->	extensoes	-> pl~
index.html	->	índex	-> html

5. Recorrendo à linguagem PERL e às expressões regulares, elabore a script **"devolveip.pl"** que deve devolver a lista de ip's encontrada no comando "netstat –na", e contar quantas vezes cada um se repete.

6.

Elabore, com recurso à linguagem PERL, a função "GetExtension($)", que deve devolver uma string contendo a extensão de um nome de ficheiro que lhe seja passado como argumento. Por extensão, entende-se o texto que surge após o último ponto. Por exemplo, no caso do nome "*a.txt*", deverá ser devolvido "*txt*", ao passo que para o nome "*a.b.c.zip*" a função deverá devolver "*zip*".

7. Elabore a função PERL SplitPath($) que recebendo um caminho de ficheiro (seja ele absoluto ou relativo) deve devolver duas strings: uma referente aos diretórios e a segunda referente ao nome do ficheiro (caso exista). Por exemplo, dado o caminho "/root/test/fich.txt", a função deve devolver como *parte diretório "/root/test/"* e como parte ficheiro "*fich.txt*".

2. Documentação PERL

Existe documentação PERL acessível na web. Como sítio de referência, sugere-se www.perl.org (e em particular perldoc.perl.org) .

Em termos de bibliografia, recomenda-se o livro "*Learning PERL*" da O´Reilly (pesquise "learning perl" num motor de busca web), o mais avançado "*Programming PERL*" também da editora O'Reilly e por fim, o livro para especialistas em REGEX, o "*Mastering Regular Expressions*" de Jeffrey Friedl.

3. Bibliografia

- GARS - Expressões Regulares – ESTG – IPLeiria, (Patrício Domingues (2007-2008))

- "Beginning PERL"
  http://www.perl.org/books/beginning-perl/ (Abril 2008)

- "Mastering Regular Expressions", 3rd edition, Jeffrey Friedl, Agosto 2006
  http://oreilly.com/catalog/9780596528126/

- "Perl 5 Tutorial"
  http://www.cbkihong.com/download/perltut.pdf (Abril 2008)

- "Mastering PERL REGEX"
  http://www.cbkihong.com/download/perltut.pdf (Abril 2008)

- Advanced PERL quick reference guide
  http://refcards.com/docs/trusketti/perl-regexp/perl-regexp-refcard-a4.pdf