Integração de Dados

2021/22

Extração de dados: Wrappers e Expressões Regulares

Recursos:

- **Web:**
  - http://www.regular-expressions.info/
  - https://reulison.com.br/regex/
  - https://www.piazinho.com.br/
- **Para testar ER:**
  - http://piazinho.com.br/ed4/exemplos.html#1
  - http://www.regextester.com/
  - http://myregexp.com/

Extração de dados: motivação

- **Grande quantidade de informação baseada em texto sem estrutura:**
  - Páginas Web
  - Emails
  - *Log files*
  - Protocolos de comunicação
  - Ficheiros: .doc, .rtf, .xls, .txt, ...
- **Como é representada a informação**
  - Dados estruturado ou semiestruturados: Bases de dados, XML
    - Pesquisa de informação: SQL, XPath ou XQuery
  - Texto sem estrutura
    - Como fazer pesquisas?

Extração de dados: wrappers

- **Para extrair dados de páginas web, fontes de dados semi-estruturadas ou sem estrutura:**
  - Usar Wrappers
    - **podem ser construídos manualmente:**
      - Expressões regulares, XPath
    - **podem ser construídos automaticamente:**
      - aprendizagem automática

Expressões regulares: definição

- É um método formal de se especificar um padrão de texto.
- É uma forma de procurar um texto que segue um determinado padrão ou possui determinadas restrições.
- É um conjunto de símbolos, caracteres literais e caracteres com funções especiais que, agrupados entre si, formam uma sequência, uma expressão. Essa expressão é interpretada como uma regra, que indicará sucesso se uma entrada de dados obedecer exatamente a todas as condições dessa regra.

Ou ainda...

- ER = uma forma de procurar um texto do qual não se possui toda a informação, mas apenas algumas variações possíveis (ex: *Pesquisar numa pauta o nome dos alunos que tenham o número de aluno começado por a2101*)
- ER = uma forma de um programador especificar padrões complexos que podem ser procurados e numa cadeia de caracteres (ex: *retirar de um log file todas as pessoas que acederam entre as 9:00 e as 12:00*)

Expressões Regulares: suporte

- **Presente em (quase) todos os programas e linguagens de programação**
  - *Exemplos*: lex, vim, awk, emacs, grep, Perl, C, C++, **Java**, Ruby, Python, .NET

Expressões Regulares: versões

- **POSIX** (***IEEE Portable Operating System Interface, 1986)***
  - É um padrão mais "humano legível", mas não suportada por algumas linguagens de programação
- **PCRE** ***(Perl Compatible Regular Expressions)***
  - Padrão simplificado, usado por algumas linguagens de programação (Perl)
- **Exemplo**
  - ER para encontrar todos os dígitos de 0 a 9:
    - POSIX: **[0-9]**
    - PCRE: **\d**

Expressões Regulares

- **Vários tipos de caracteres:**
  - Literais
  - Metacaracteres:
    - âncoras
    - representantes
    - quantificadores
    - outros

Expressões Regulares: Literais

- **Caracteres com correspondência literal**
- **No entanto são *case sensitive*!**

- ***Exemplo*:**
  'isec' → "estudar no isec"
  'isec' → "estudar no ISEC"

Expressões Regulares: Literais

- **ER: ana**

A ana foi com a analisa e a susana à cabana da tia Ana.
Foram à loja da Ruana comprar bananas e escreveram um recado: voltamos para a semana
ana e susana
ananana

Expressões Regulares: Metacaracteres

- **Caracteres especiais:**
  - **Âncoras**: marcam uma posição específica na linha ou palavra;
  - **Representantes**: representam um ou mais caracteres;
  - **Quantificadores**: indicam o numero de repetições permitidas de um caracter um bloco de caracteres;
  - **Outros**: caracteres de escape, de agrupamento, etc;

Expressões Regulares: Metacaracteres

- **Âncoras:**

^	circunflexo	início da linha
$	cifrão	fim da linha
\b	borda	início ou fim de palavra

Exemplos

ana	A **ana** foi ver as m**ana**s **ana**stácia e lu**ana** que viajam na próxima sem**ana**
^ana$	**sem correspondências!**
\bana\b	A **ana** foi ver as manas anastácia e luana que viajam na próxima semana
ana\b	A **ana** foi ver as manas anastácia e lu**ana** que viajam na próxima sem**ana**
\bana	A **ana** foi ver as manas **ana**stácia e luana que viajam na próxima semana

Expressões Regulares: Metacaracteres

1. A ana foi com a analisa e a susana à cabana da tia Ana.
2. Foram à loja da Ruana comprar bananas e escreveram um recado: voltamos para a semana
3. ana e susana
4. ana

- **O que encontram as seguintes ERs?**
  - **^ana$**
  - **^ana**
  - **ana$**
  - **\bana**
  - **\bana\b**
  - **ana\b**

Expressões Regulares: Metacaracteres

- **Representantes:**

.	ponto	qualquer caracter excepto \n
[...]	lista	lista de caracteres permitidos
[^...]	lista negada	lista de caracteres proibidos

- **Exemplos**

[abcABC]	apenas **um** dos caracteres a, b, c, A, B ou C
[a-z]	**um** caracter minúsculo
[a-zA-Z]	**um** caracter minúsculo ou maiúsculo
[0-9]	**um** digito (0, 1, 2, 3, ...9)
[01ab]	apenas **um** dos caracteres 0, 1, a ou B
[^a-z]	**um** caracter que não seja minúsculo
[^0-9]	**um** caracter que não seja um dígito

Expressões Regulares: Metacaracteres

- **Exemplos**
  - **[ab][a-zA-Z]**
    - **a maria abriu a bela mala**
  - **[abm][abe][^r]**
    - **a maria abriu a bela mala**
  - **[ab][ab][abr]**
    - **a maria abriu a bela mala**

Expressões Regulares: Metacaracteres

- **Representantes:**
- **Exemplo 1**

  \b**[abcABC].[^A-Z]**\b

Ana	Dar	aXn
Bom	AnA	Dna

- **Exemplo 2**

  \b**.[abs][^0-9A-Z]**\b

1sss	5s5	5sa
abs	xbs	abS

Expressões Regulares: Metacaracteres

- **Exemplos**

  \b[a-zA-Z]id[0-9].\b

AZid09	Zaz23	cid2
aid9r	Zid3X	cid22

  \bAna[bela].[123]\b

Anabela	Anabela123	Anabe1
Anae2	anabe3	Anala3

Expressões Regulares: Metacaracteres

- **Quantificadores:**

?	interrogação	zero ou um
*	asterisco	zero, um ou mais
+	mais	um ou mais
{n,m}	chavetas	de n até m
{n}	chavetas	exatamente n

Expressões Regulares: Metacaracteres

Exemplos

\babs?\b
ab abs as bs absss

abs*
ab abs abss absss bssss

\babs{2,5}\b
abs abss absss abssss aaabs abbbssss

a{2}b{3}s{4}
aabbbssss aabs aabbss absss

Expressões Regulares: Metacaracteres

- **Exemplo 1**

\ba?b*c+d{1,4}\b

abbbbcccdd	abcd	aabcd
ccccdddd	acd	abddd

- **Exemplo 2**

\b[ab]?[cd]*[ef]{1,4}\b

ace	bdffff	dddeee
abcdeff	acdddccddcceeef	cccceee

Expressões Regulares: Metacaracteres

- **Exemplos**

\b[a-zA-Z]{2}[0-9]{2}.?\b

aa00xx	aa00xx	AZ99z
aa00	aB123	AZ99

\bgolo{5}!+\b

golooooo!!!!	golooooo

\bgolo{5}!*\b

golooooo!!!!	golooooo

Expressões Regulares: Metacaracteres

- **Outros:**
  - **\c** **escape** **torna literal o caracter c**
  - **|** **ou** **um grupo ou outro**
  - **()** **grupo** **delimita um grupo**

- **Exemplo**

ba*?

baaaa **ba*** ca

((ab)|(cd))s*

absimoes **cd**silva **ab**sssara **ab**rir **cd**

Expressões Regulares: Metacaracteres

- **Outros:**

\s	**Qualquer caracter de espaçamento**
\d	**qualquer dígito -> *equivalente a [0-9]***
\D	**qualquer caracter excepto dígito *equivalente a [^0-9]***
\n	**new line**
\r	**carriage return**

Expressões Regulares: Metacaracteres

- **Diferenças entre () e []**

Expressão Regular	Padrões encontrados
(ab)|(cd)	ab cd
[ab]|[cd]	a b c d
a|b|c|d	a b c d
[abcd]	a b c d
(abcd)	abcd
[ab]*	aaa bbb ababab aabbaa
(ab)*	ab abab abababab

Expressões Regulares: Metacaracteres

- **Grupos: delimitados por ()**

  `([AaBb].{1,6}):([0-9]{2}):(200[0-9])`

- **Identificar os grupos:**
  - Grupo 1 palavras começadas por A, a, B ou B e com tamanho máximo de 7 caracteres
  - Grupo 2 identificador com dois dígitos
  - Grupo 3 anos entre 2000 e 2009
  - : literal delimitador

Expressões Regulares: Metacaracteres

- **Pessoas que entraram depois das 9:00?**

06:13 – Tomé Lopes
08:40 – Anabela Simões
09:00 – Rui Lopes
09:05 – Carlos Freire
10:00 – Pedro Melo
10:15 – Sofia Limões
11:16 – Filipe Torres
11:45 – Maria Ferro
14:55 – Luís Paulo
16:08 – Gustavo Matos
22:14 – Maria Lurdes

Grupo 1: Identifica as horas
Grupo 2: separador
Grupo 3: nome das pessoas

(09:(0[1-9]|[1-9][0-9])|([1-2][0-9]|1[0-9]|2[0-3]):[0-5][0-9])(-)([a-zA-Zãõíéó\s]+

Expressões Regulares: Metacaracteres

- **Pessoas nascidas entre 1990 e 2000?**

Ana Simões nasceu a 12-12-1995
Rui Matos nasceu a 01-01-1987
José Matias nasceu a 26-11-2002
Daniel Silva nasceu a 10-05-2000
Fernando Grave nasceu a 30-04-1997
Margarida Moço nasceu a 01-01-1990

Grupo 1: Identifica o nome
Grupo 2: nasceu a
Grupo 3 = **subgrupo 3 identifica o ano entre 1990 e 2000**

([a-zA-Zçõéã\s]*) (nasceu a) ([0-9]{2}-[0-9]{2}-**(199[0-9]|2000)**)

Sugestões para construir uma ER

- **Construa a ER por partes**
- **Comece por uma versão simplificada mas funcional**
- **Vá acrescentando os melhoramentos necessários passo a passo**
- **Exemplo:**
  - Construa uma ER que valide datas no formato
    - dd/mm/aaaa ou dd-mm-aaaa

ER para validar uma Data

ER1: [0-9]{2}/[0-9]{2}/[0-9]{4}

ER2: [0-9]{2}[/-][0-9]{2} [/-][0-9]{4}

ER3: [0123][0-9] [/-][0-9]{2} [/-][0-9]{4}

ER4: [0123][0-9] [/-][01][0-9] [/-][0-9]{4}

ER5: [0123][0-9] [/-][01][0-9] [/-][12][0-9]{3}

ER6: ([012][0-9]|3[01]) [/-]([01][0-9])[/-]([12][0-9]{3})

ER7: ([012][0-9]|3[01]) [/-](0[1-9]|1[012]) [/-]([12][0-9]{3})

…..

EXPRESSÕES REGULARES
Exercícios

Exercícios:

- **Escreva uma expressão regular que capture as palavras :**

- **golo**
- **golos**
- **golo!**
- **goooolo!**
- **gooolooooo!!!!!**
- **golos!**
- **goooloooos!**
- **gooooooloos!!!**

Exercícios:

- **Construa uma expressão regular que encontre os nomes cujo primeiro nome comece em "An" e cujo apelido comece com "Si" ou com "Sa"**
  - Anabela Sintra
  - Antonieta Silva
  - Anselmo Sinatra
  - Aniceto Sala
  - Ana Santos

Exercícios:

- **Construa uma expressão regular que encontre todos os números inteiros num texto. Números começados por zero são inválidos? (1 12 2345 343234 1202)**

- **Construa uma ER que encontre todas as strings que contêm as 5 vogais, em qualquer número, mas por sempre por ordem alfabética (aeiou aaaeeiouuuu aaaaeeeeiiiiioouuuu)**

Exercícios:

- **Construa uma expressão reg**ular que permita encontrar todos os números de telemóvel portugueses (iniciados por 91, 92, 93, 96 e com nove dígitos)

- **Altere a expressão regular de forma a aceitar apenas números terminados em 00.**

Exercícios:

- **Construa uma expressão regular que valide *passwords* com os seguintes requisitos**
  - Começar por um caracter maiúsculo
  - Seguido de caracteres maiúsculos, minúsculos e/ou dígitos
  - Termine com um digito
  - Tamanho entre 6 e 10 caracteres

Exercícios:

- **Construa uma expressão regular que permita validar um e-mail que um utilizador introduz num formulário.**
  - **começa por letra e pode ter dígitos, . ou _**
  - **o domínio tem 2 ou 3 caracteres**
    - abs_1234@isec.pt
    - abs@eden.dei.uc.pt
    - abs.bekas_1972@yahoo.com

- **Altere a expressão regular de forma a aceitar apenas emails do domínio pt**

Exercícios:

- **Escreva algumas cadeias binárias que sejam representadas pelas seguintes ERs:**

\b0(0|1)*0\b **\b(01)*\b** **\b[01]*\b**

\b0+10*10*10+\b **\b(0|1)*0(0|1)(0|1)\b**

Exercícios: Analise as ERs e indique quais as correspondências que estas encontram:

^a(ab)*a$

1. abababa
2. aaba
3. aabbaa
4. aba
5. aabababa

^a.[bc]+$

1. abc
2. abbbbbbbb
3. azc
4. abcbcbcbc
5. ac
6. asccbbbbcbcccc

^ab+c?$

1. abc
2. ac
3. abbb
4. bbc

^(abc|xyz).{2}$

1. abcxyzaa
2. abcxy
3. abcd
4. abc|xyz.
5. xyzww

Exercícios: Analise as ERs e indique quais as correspondências que estas encontram:

^[a-z]+[\.\?!]$

1. battle!
2. Hot
3. green
4. swamping.
5. jump up.
6. undulate?
7. is.?

^[a-z][\.\?!]\s+[A-Z]$
(\s == caracter de espaçamento)

1. A. B
2. c! d
3. e f
4. g. H
5. i? J
6. k L

^[a-zA-Z]*[^,]=$

1. Butt=
2. BotHEr,=
3. Ample
4. FIdDlE7h=
5. Brittle =
6. Other.=$

^<[^>]+>$

1. <an xml tag>
2. <opentag> <closetag>
3. </closetag>
4. <>
5. <abs>
6. <with attribute="77">

Exercícios:

- **Analise a seguinte ER e assinale quais as sequências encontradas:**

  ^1{3}0?(1*|2{2})$

  130111
  111
  1110111
  130?1*
  11122
  1110122
  111022

Exercícios:

- **Analise a seguinte ER e escreva algumas frases reconhecidas pela ER**

  ^[aAoO]\s.{3,5}\sestá\s((ap)|(rep))(rovad)([ao])$

Exercícios:

- **Analise a seguinte ER e indique quais as correspondências que esta encontra:**

  (very)+(fat)?(tall|ugly) man

  1) very fat man
  2) fat tall man
  3) very very fat ugly man
  4) very very very tall man

Exercícios:

- **Verdade ou Falso?**

  ^(0|1)*$ == ^(0*|1*)$

  ^(0|1)*$ == ^(0*1*)*$

Exercícios

- **Expressão regular que encontre frases começadas por vogais maiúsculas e terminadas por ponto final. Assuma que na frase podem surgir apenas os caracteres alfabéticos maiúsculos e/ou minúsculos e espaços em branco.**

Exercícios:

- **Escreva uma expressão regular que encontre frases interrogativas de tamanho entre 10 e 20 caracteres. Nas frases assuma que começam por qualquer caracter maiúsculo, seguidos de caracteres minúsculos, espaçamentos ou dígitos e terminando com o sinal de interrogação.**

Exercícios:

- Escreva uma expressão regular que encontre as palavras que comecem por ga, go, gi ou gr seguidas de qualquer caracter minúsculo.
- Na frase seguinte, as palavras a sublinhado mostram exemplos do que a ER deve validar.

o gato preto e grande gosta de gaivotas brancas gigantes

Exercícios:

- Imagine que possui o seguinte ficheiro de coordenadas. Escreva a ER que capture quais as cidades cuja Latitude se situa a N e Longitude a W. Indique os grupos relevantes.

41°9'N	8°38'W	Porto
38°42'N	9°11'W	Lisboa
51°30'25''N	0°07'39''W	Londres
22°54'30''S	43°11'47''W	Rio
55°45'8''N	37°37'56''E	Moscovo

Exercícios:

```
Sofia Melo *** CR *** 15 valores
Pedro Mota *** ID *** 6 valores
Rui Matos *** IIA *** 1 valor
Carlos Lopes *** POO *** 18 valores
Sandra Serra *** IP *** 7 valores
Romeu Torres *** POO *** 20 valores
```

- **Recorrendo a grupos, construa uma expressão regular que encontre os nomes das pessoas, respetivas disciplinas e classificações, mas apenas nas disciplinas com notas superiores ou iguais a 15 valores. Indique os grupos necessários de forma a retirar o nome, a disciplina e a nota que respeitem o critério acima definido.**

Exercícios:

```
Caramelo - gato silvestre - vacinado em Coimbra
Tareco - gato domestico - vacinado em Antanhol, Coimbra
Calvin - cao Labrador - vacinado em Almada, Lisboa
Leia - cadela Cocker - vacinada em Condeixa, Coimbra
Bounty - cadela - vacinada em Aveiro
Tica - gata europeia - vacinada em Coimbra
Tareco - gato - vacinado em Eiras, Coimbra
```

- **Recorrendo a grupos, construa uma expressão regular que encontre os animais vacinados em Coimbra, em qualquer localidade. Indique os grupos necessários de forma a retirar o nome e o tipo de animal (apenas a indicação se é cão/cadela ou gato/gata).**

Exercícios:

log01 - 12/01 - 09:00
log02 - 05/01 - 06:10
log02 - 05/01 - 06:66
log03 - 07/01 - 01:99
log99 - 19/01 - 08:57
log123 - 21/01 - 09:01
log1099 - 30/01 - 08:32
log012 - 04/02 - 07:45

- **Recorrendo a grupos, construa uma expressão regular que encontre os dias do mês de Janeiro em que os logs foram efectuados antes das 9:00. A ER regular deve encontrar apenas horas válidas, No ficheiro anterior há duas horas inválidas que não devem ser consideradas (06:66 e 01:99).**

Exercícios: procurar dados na web

- **Procurar no site bertrand.pt os detalhes do livro com isbn** 9789722043892
  - **Aceder ao site**
  - **Procurar** "9789720040923"
  - **Na página de procura faça Ver-Origem (Source)**
  - **Analise o conteúdo e construa as ERs para encontrar o autor, o titulo, o preço do livro**
  - **Experimente as ERs para outros ISBN e verifique se a informação do autor, titulo e preço é encontrada corretamente**
  - **Teste as ERs no *source* da página usando uma das ferramentas on line:**
    - **Por exemplo: http://piazinho.com.br/ed4/exemplos.html#1**