

Disciplina:

# **Processamento de linguagem natural**

Professor: Gabriel Assunção



# Expressão regular (Regex)

## O que são Regex?

- Formas de representar um grupo de strings
- Tem como objetivo identificar padrões comuns de strings.
- Exemplos:
  - Identificar strings que representam valores em reais.
  - Identificar strings que são números em textos.

## Quando usamos?

- Identificar padrões
- Processar os padrões (substituir o valor)
- Validar se o dado possui uma informação
- Realizar quebra do texto a partir de um padrão

## Quando usamos?

- Identificar url em texto
- Identificar e-mails
- Quebrar o texto utilizando pontuação
- Anonimização de dados (documentos)
- Radicalização

# Expressão regular Python

- Biblioteca: Re
- Operações:
  - split
  - findall
  - search
  - sub

## Exemplos

Expressão	Exemplo
A	<b>A</b> na faz aniversário no dia 15 de maio!
15	Ana faz aniversário no dia <b>15</b> de maio!
!	Ana faz aniversário no dia 15 de maio <b>!</b>
a	Ana <b>a</b> faz <b>a</b> aniversário no dia <b>a</b> 15 de maio <b>a</b> !

# Disjunções



## Exemplos

Expressão	Match
[Aa]	A ou a
[abcd]	a ou b ou c ou d
[059]	0 ou 5 ou 9

## Exemplos

Expressão	Match
[A-Z]	caractere maiúsculo
[a-z]	caractere minúsculo
[0-9]	um único dígito

## Exemplos

Expressão	Match
[^A-Z]	não é caractere maiúsculo
[^aA]	nem a e nem A

## Exemplos

Expressão	Match	Exemplo
via.em	qualquer caractere, exceto quebra	viagem viagem

# Âncoras

## Âncoras

Estabelece a posição de referência do regex.

Padrão	Significado
^	Início do texto.
\$	Fim do texto.

# Especificadores

## Especificadores

Especificam um conjunto de caracteres

Padrão	Significado
<code>\d</code>	o mesmo que <code>[0-9]</code>
<code>\D</code>	o mesmo que <code>[^0-9]</code>
<code>\s</code>	espaço, quebra de linha, tabs etc.; o mesmo que <code>[ \t\n\r\f\v]</code>
<code>\S</code>	o mesmo que <code>[^ \t\n\r\f\v]</code>
<code>\w</code>	o mesmo que <code>[a-zA-Z0-9_]</code>
<code>\W</code>	o complemento de <code>\w</code>



# Quantificadores

## Quantificadores

Define o número de repetições da expressão

Padrão	Significado
$\{n\}$	exatamente $n$ ocorrências
$\{n,m\}$	no mínimo $n$ ocorrências e no máximo $m$
$\{n,\}$	no mínimo $n$ ocorrências
$\{,n\}$	no máximo $n$ ocorrências
$?$	0 ou 1 ocorrência; o mesmo que $\{,1\}$
$+$	1 ou mais ocorrência; o mesmo que $\{1,\}$
$*$	0 ou mais ocorrência

## Exemplo

Queremos construir um **aplicativo** para **ajudar** um usuário a **comprar** um **computador** na web.

*O usuário quer “qualquer máquina com pelo menos 6 GHz e 500 GB de HD por menos de R\$ 1000”.*

## Exemplo

Queremos construir um **aplicativo** para **ajudar** um usuário a **comprar** um **computador** na web.

*O usuário quer “qualquer máquina com pelo menos **6 GHz** e **500 GB de HD** por menos de **R\$ 1000**”.*

## Exemplo

Identificar Processador: 6GHz

```
[0-9]+ *([Gg][Hh][Zz])
```

## Exemplo

Identificar HD: 500Gb de HD

```
[0-9]+ *(GB|[Gg]((igabyte)|(iga))s?)
```

## Exemplo

Identificar Preço: ***R\$ 1000***

```
R\$\s?[0-9]+(,[0-9]{2})?
```