Expressoes regulares

Ivanildo Batista

13 de marco de 2021

Expressoes Regulares na linguagem R

O objetivo e identificar se existem padroes em um texto.

```
#criando um vetor de texto
str = c("Expressoes","regulares","em linguagem R","permitem busca de padroes",
        "e exploracao de textos","podemos buscar padroes em digitos",
        "como por exemplo","10992451280")

length(str) #comprimento do vetor de texto
```

```
## [1] 8
```

```
str #imprimindo o vetor
```

```
## [1] "Expressoes"                      "regulares"
## [3] "em linguagem R"                  "permitem busca de padroes"
## [5] "e exploracao de textos"          "podemos buscar padroes em digitos"
## [7] "como por exemplo"                "10992451280"
```

Trabalhando com as funcoes

```
#Funcao grep(): retorna o vetor de indices dos caracteres que contem o padrao especificado

grep("ex",str,value = F) #indice do padrao no texto
```

```
## [1] 5 7
```

```
grep("ex",str,value = T) #texto onde o padrao foi encontrado
```

```
## [1] "e exploracao de textos" "como por exemplo"
```

```
grep("\\d",str,value = F) #indice dos digitos
```

```
## [1] 8
```

```
grep("\\d",str,value = T) #texto onde estao os digitos
```

```
## [1] "10992451280"
```

```
#Funcao grepl(): quando um padrao e encontrado ou procura um padrao

grepl("\\d",str) #procurando um padrao de digitos
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
```

```
grepl("\\D",str) #buscando um padrao de nao digitos
```

```
## [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
```

```
#Funcao gsub(): Substitui os caracteres encontrados de acordo com o padrao especificado

gsub("em","|",str) #onde tiver "em" ele vai substituir por "|" em "str"
```

```
## [1] "Expressoes"                    "regulares"
## [3] "| linguag| R"                  "permit| busca de padroes"
## [5] "e exploracao de textos"        "pod|os buscar padroes | digitos"
## [7] "como por ex|plo"               "10992451280"
```

```
gsub("ex","EX",str, ignore.case = T) #substituindo "ex" por "EX" em "str" e
```

```
## [1] "EXpressoes"                     "regulares"
## [3] "em linguagem R"                 "permitem busca de padroes"
## [5] "e EXploracao de tEXtos"         "podemos buscar padroes em digitos"
## [7] "como por EXemplo"               "10992451280"
```

```
#ignora letras maiusculas ou minusculas
```

```
#Funcao sub(): Substitui o primeiro caracter encontrado de acordo com o padrao
             #especificado

sub("em","EM",str)
```

```
## [1] "Expressoes"                     "regulares"
## [3] "EM linguagem R"                 "permitEM busca de padroes"
## [5] "e exploracao de textos"         "podEMos buscar padroes em digitos"
## [7] "como por exEMplo"               "10992451280"
```

```
#Funcoes regexpr() e gregexpr(): vao mostrar a localizacao do padrao no texto

frase = "Isso 玫将 uma string"
regexpr(pattern = "u", frase)
```

```
## [1] 11
## attr(,"match.length")
## [1] 1
## attr(,"index.type")
## [1] "chars"
## attr(,"useBytes")
## [1] TRUE
```

```
gregexpr(pattern = "u", frase)
```

```
## [[1]]
## [1] 11
## attr(,"match.length")
## [1] 1
## attr(,"index.type")
## [1] "chars"
## attr(,"useBytes")
## [1] TRUE
```

```
#nesse caso o "pattern" esta na posicao 8
```

```
#Outros exemplos: eliminando elementos de um texto

str2 = c("2678 e maior que 45-@????!慎犁$",
         "Vamos escrever 14 scripts na linguagem R")

str2
```

```
## [1] "2678 e maior que 45-@????!<U+00A7>$"
## [2] "Vamos escrever 14 scripts na linguagem R"
```

```
gsub("\\d","",str2) #eliminando digitos por espa攻犁o
```

```
## [1] " e maior que -@????!<U+A>$"
## [2] "Vamos escrever  scripts na linguagem R"
```

```
gsub("\\D","",str2) #eliminando nao digitos
```

```
## [1] "267845007" "14"
```

```
gsub("\\S","",str2) #eliminando espacos
```

```
## [1] "    "   "      "
```

```
gsub("[iot]","",str2) #eliminando do texto as letras "i","o" e "t"
```

```
## [1] "2678 e mar que 45-@????!<U+00A7>$"
## [2] "Vams escrever 14 scrps na lnguagem R"
```

```
gsub("[[:punct:]]","",str2) #eliminando pontuacoes
```

```
## [1] "2678 e maior que 45U00A7"
## [2] "Vamos escrever 14 scripts na linguagem R"
```

```
gsub("[^@]","",str2) #eliminando tudo, exceto o @
```

```
## [1] "@" ""
```