08

Mãos na massa: Coletando dados

Chegou a hora de você executar o que foi visto na aula! Para isso, execute os passos listados abaixo.

Será utilizado o **RStudio** como interface gráfica, então todos os comandos aqui mostrados deverão ser executados no seu console.

1) Abra o **RStudio** e carregue o script **Coleta_e_Explora.R**, que você pode baixar aqui (https://s3.amazonaws.com/caelum-online-public/731-pipeline-big-data/01/arquivos/Coleta_e_Explora.R). Caso a acentuação do conteúdo do script esteja errada, abra-o novamente, acessando o menu ***File -> Reopen with Enconding...***, escolhendo **UTF-8** em seguida.

2) Especifique o diretório de trabalho, no console do **RStudio**, executando o comando `setwd` do script. **O diretório que está no script provavelmente não será igual ao seu, então modifique-o para o seu diretório**:

```
setwd("C:\\Users\\eduar\\OneDrive\\aBig Data\\aAlura\\R")
```

Para confirmar que o diretório foi criado corretamente, você pode executar o comando `getwd` .

3) Como o **R** funciona a base de *packages*, instale os necessários para busca e extração de dados em páginas web, o **httr** e o **XML**:

```
install.packages('httr')
install.packages('XML')
```

4) Em seguida, execute os comandos `library` , para que os *packages* estejam na memória:

```
library(httr)
library(XML)
```

5) Você vai fazer uma coleta de páginas web dos últimos 20 anos, então primeiramente inicialize as variáveis:

```
df_OVNI  <- data.frame()
mes_corrente = 9
ano_corrente = 1997
ano_mes_corrente = (ano_corrente * 100) + mes_corrente
```

6) Agora faça o *loop*, onde você vai ler cada página, até que o mês/ano chegue a 09/2017:

```
while (ano_mes_corrente  <= 201709) {
  site <- paste("http://www.nuforc.org/webreports/ndxe", as.character(ano_mes_corrente), ".html")
  site <- gsub (" ", "", site)
  html2 <- GET(site)
  parsed <- suppressMessages(htmlParse(html2, asText=TRUE))
  tableNodes <- getNodeSet(parsed, "//table")
  tb <- readHTMLTable(tableNodes[[1]])
  df_OVNI <- rbind(df_OVNI,tb)
```

```
    if (mes_corrente == 12)
    {
        mes_corrente <- 1
        ano_corrente <- ano_corrente + 1
        ano_mes_corrente <- (ano_corrente * 100) + mes_corrente
    }
    else
    {
      mes_corrente <- mes_corrente + 1
      ano_mes_corrente <- ano_mes_corrente + 1
    }
    print(ano_mes_corrente)
  }
```

7) Agora, salve esses dados em um arquivo CSV:

```
write.csv(rbind(df_OVNI), file = "OVNIS.csv")
```