Web Scraping: implementação de um modelo em R para extração de notícias jornalísticas.

Carlos Alberto Alves de Meneses,20180003202

2023-04-02

## Web Scraping

Nessa primeira fase do projeto iremos nos concentrar na obtenção dos dados de apenas um site, neste caso do *portal do Jornal da Paraíba*, deixando a implementações de outros sites de noticias para trabalhos futuros.

**Executando a raspagem da web passo a passo, usando o pacote rvest R escrito por Hadley Wickham**.

O *rvest* é uma biblioteca R muito útil que ajuda a coletar informações da páginas da web.

O primeiro passo é baixar e carregar os pacotes necessários.

# Baixando e carregando o pacote  
library( rvest )  
library( xml2 )  
library(shiny)  
library(httr)  
library(httr2)

##   
## Attaching package: 'httr2'

## The following object is masked from 'package:xml2':  
##   
## url\_parse

O pacote rvest define o link da página da web como o primeiro passo. Depois disso, os rótulos apropriados devem ser definidos. A linguagem HTML edita o conteúdo usando várias tags e seletores.

Esses seletores devem ser identificados e marcados para armazenamento de seu conteúdo. Em seguida, todos os dados gravados podem ser transformados em um conjunto de dados apropriado e a análise pode ser realizada.

Coletaremos um conjunto de dados de um portal de notícias (www.jornaldaparaiba.com.br). Este site fornece informações gerais e principalmente do estado da Paraíba.

Vamos comerçar a coletar informações para descobrir quais as principais noticias que estão sendo vinculadas neste portal.

* Para coletar as informações sobre as principais manchetes do portal, usaremos a URL da página de destino do site.

# Link para:  
url <- " https://jornaldaparaiba.com.br/ "  
url

## [1] " https://jornaldaparaiba.com.br/ "

Como mencionamos, estamos interessado em coletar dados sobre as materias jornalisticas publicadas no site.

Agora, a parte que mais nos importa: a coleta dos dados!

O script a seguir, fornece os seguintes procedimentos: visite o URL da página da Web, coletando nós HTML usando a função *read\_html*.

Para analisar nós HTML, estamos usando as regras XPath.

# Ler o HTML da página da web  
url <- c("https://www.jornaldaparaiba.com.br")  
  
pagina <- read\_html(url)  
  
elementos <- pagina %>%  
 html\_nodes("div.container a") #Aplicando as regras XPath

XPath lida principalmente com os nós das árvores XML 1.0 ou XML 1.1. Ë usado para representar a estrutura hierárquica de um documento XML.

XPath usa sintaxe não XML e funciona na estrutura lógica de documentos XML. XPath é projetado para ser usado embutido em uma linguagem de programação.

XPath tem sete tipos diferentes de nós: elemento, atributo, texto, namespace, instrução de processamento, comentário e nós de documento.

Para este projeto, estamos usando a função *html\_nodes* e definindo nossas regras XPath , que já temos, dentro da função:

*html\_nodes(““div.container a”“)*

O código a seguir realiza as duas funções anteriores e exibe o conteúdo da raspagem dos dados através da função *cat*.

# Carregar o pacote rvest  
library(rvest)  
library(xml2)  
library(shiny)  
library(httr)  
library(httr2)  
  
# Ler o HTML da página da web  
url <- "https://www.jornaldaparaiba.com.br"  
  
pagina <- read\_html(url)  
  
dt <- html\_nodes(pagina, "div.container a")  
  
# Exibir o conteúdo dos elementos selecionados  
cat("Conteúdo dos elementos selecionados:\n")

## Conteúdo dos elementos selecionados:

cat(html\_text(dt), sep = "\n")

##   
##   
##   
## Notícias  
## Justiça determina bloqueio de R$ 399 milhões em bens da Fiji Solutions  
##   
## 14/04/2023 05:46  
##   
##   
##   
##   
## Educação  
## UEPB divulga 3ª chamada da lista de espera do Sisu 2023; veja lista de classificados  
##   
## 14/04/2023 05:49  
##   
##   
##   
## Tecnologia  
## Vazamento de fotos do corpo de Marília Mendonça: advogado explica crimes  
##   
## 14/04/2023 05:08  
##   
##   
##   
## Clima e Tempo  
## Inmet emite alerta de perigo de chuvas intensas para cidades da Paraíba  
##   
##   
##   
##   
## Meio Ambiente  
## 8 praias estão impróprias para banho no Litoral da Paraíba  
##   
##   
##   
##   
## Clima e Tempo  
## Previsão do tempo para o fim de semana, na Paraíba  
##   
##   
##   
##   
##   
## Esportes  
## Série B do Campeonato Brasileiro 2023 tem 14 jogadores paraibanos  
##   
##   
##   
##   
##   
## Esportes  
## Nacional de Patos na Série D 2023: confira o novo elenco do Naça  
##   
##   
##   
##   
## Clima e Tempo  
## Inmet emite alerta de perigo de chuvas intensas para cidades da Paraíba  
##   
## 14/04/2023 01:05  
##   
##   
##   
## Meio Ambiente  
## 8 praias estão impróprias para banho no Litoral da Paraíba  
##   
## 14/04/2023 03:22  
##   
##   
##   
## Clima e Tempo  
## Previsão do tempo para o fim de semana, na Paraíba  
##   
## 14/04/2023 08:50  
##   
##   
##   
##   
## Esportes  
## Série B do Campeonato Brasileiro 2023 tem 14 jogadores paraibanos  
##   
## 14/04/2023 07:58  
##   
##   
##   
##   
## Esportes  
## Nacional de Patos na Série D 2023: confira o novo elenco do Naça  
##   
## 14/04/2023 11:11  
##   
##   
##   
##   
##   
## Esportes   
## Times da Paraíba: conheça a história dos maiores clubes  
##   
## 09/07/2021 05:33  
##   
##   
##   
##   
## Esportes  
## Botafogo-PB: Juliana Bandeira convoca torcida para apoiar o futebol feminino  
## 14/04/2023 06:06  
##   
##   
##   
##   
##   
## Tecnologia  
## WhatsApp anuncia novo recurso para proteção contra roubo de conta  
## 14/04/2023 04:14  
##   
##   
##   
##   
##   
## Economia  
## Volume de serviços na PB tem 4ª maior queda do Brasil em janeiro  
## 14/04/2023 03:36  
##   
##   
## Carregar mais  
##   
##   
##   
## Silvio Osias  
## Turnê Gilberto Gil in Concert passa por João Pessoa nesta quinta-feira  
##   
## 13/04/2023 08:07  
##   
##   
##   
##   
## Conversa Política  
## ‘Blocão’ de Lira na contraofensiva a Pacheco divide os Ribeiro no Congresso  
##   
## 13/04/2023 10:45  
##   
##   
##   
##   
## Pleno Poder  
## Veneziano garante apoio a empréstimos de CG e estreita laços com Bruno Cunha Lima  
##   
## 13/04/2023 07:56  
##   
##   
##   
##   
## Pleno Poder  
## Governo troca aliado de Adriano Galdino por indicado de Lucas Ribeiro na gerência do Amigão  
##   
## 13/04/2023 01:07  
##   
##   
##   
##   
##   
## Esportes   
## Times da Paraíba: conheça a história dos maiores clubes  
##   
## 09/07/2021 05:33  
##   
##   
##   
##   
## Esportes  
## Botafogo-PB: Juliana Bandeira convoca torcida para apoiar o futebol feminino  
## 14/04/2023 06:06  
##   
##   
##   
##   
##   
## Tecnologia  
## WhatsApp anuncia novo recurso para proteção contra roubo de conta  
## 14/04/2023 04:14  
##   
##   
##   
##   
##   
## Economia  
## Volume de serviços na PB tem 4ª maior queda do Brasil em janeiro  
## 14/04/2023 03:36  
##   
##   
##   
##   
##   
## Qual é a Boa?  
## Dia do Café: aprenda a fazer receita de bolo de café  
## 14/04/2023 03:22  
##   
##   
##   
##   
##   
## Botafogo-PB  
## Novo goleiro do Botafogo-PB, Mota já foi campeão com o técnico Surian em 2016; relembre  
## 14/04/2023 01:42  
##   
##   
##   
##   
##   
## Cultura  
## Inaldo Dantas lança ‘O Livro do Síndico’ em João Pessoa  
## 14/04/2023 01:40  
##   
##   
##   
##   
##   
## Notícias  
## Bandeira da Paraíba: a história do ‘Nego’  
## 14/04/2023 01:39  
##   
##   
##   
##   
##   
## Comunidade  
## Lei garante sigilo a filhos de mulheres vítimas de violência em transferência escolar  
## 14/04/2023 01:09  
##   
##   
##   
##   
##   
## Pleno Poder  
## Alpargatas alega sazonalidade para demissões e mantém planejamento de investimentos na Paraíba  
## 14/04/2023 01:06  
##   
##   
##   
##   
##   
## Silvio Osias  
## No palco, com dois filhos, um neto e uma neta, Gil diz pra gente que ainda é preciso virar o mundo em festa, trabalho e pão  
## 14/04/2023 12:29  
##   
##   
## Carregar mais  
##   
##   
##   
##   
## Qual é a Boa?  
## Bixarte lança o álbum ‘Traviacardo’; veja entrevista  
##   
##   
##   
##   
##   
##   
## Qual é a Boa?  
## Agnes Nunes fala sobre turnê internacional  
##   
##   
## Ver todos  
##   
##   
##   
##   
## Conversa Política  
## Fatos políticos de forma clara e opinativa. Angélica Nunes e Laerte Cerqueira  
##   
##   
##   
##   
##   
##   
## Pleno Poder  
## Notícias da cobertura política de Campina Grande e de todo o interior da Paraíba, por João Paulo Medeiros.  
##   
##   
##   
##   
##   
##   
## Caderno Animal  
## Fabi e Miguel Cavalcanti abordam comportamento e saúde dos pets. Dicas, curiosidades, tutoria responsável e harmonização entre animais e família.  
##   
##   
##   
##   
##   
##   
## Sílvio Osias  
## Muita cultura e uma pitada de entretenimento, com abordagem de obras e questões relevantes da música e do cinema.  
##   
##   
##   
##   
##   
##   
## Saúde Alerta  
## Dicas e orientações para cuidar da saúde e bem estar com o médico André Telis.  
##   
##   
##   
## Ver mais

Nesse ponto, concluímos a primeira fase do nosso projeto, realizando a raspagem das manchetes principais do *portal do Jornal da Paraíba*.

**Processamento dos dados**

Para Moreira e Rocabado (2022), a análise de conteúdo só é possível através da transformação do texto bruto em estruturas de dados convenientes para análise.

* Temos basicamente três estruturas num texto:
* **Tokens:** O texto pode ser armazenado em n-grams ou cadeias de caracteres (*strings*).
* **Corpus:** Estruturas que armazenam tanto o conteúdo de documentos como seus metadados.
* **Matris de documentos e termos (DFM ou DTM):** É uma matriz esparsa com uma linha para cada documento e uma coluna para cada termo.

A análise do conteúdo do texto como dado exige versatilidade na transformação entre estruturas.

Utilizaremos especialmente os pacotes *tidytext* e *quanteda*, ambos estruturais para análise de conteúdo.

Antes de aplicarmos a tokenizaão nos nossos dados (texto) coletados, precisamos transformá-los para o formato aceito pelo *tidytext* que é um data.frame.

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.1 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ readr::guess\_encoding() masks rvest::guess\_encoding()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ httr2::url\_parse() masks xml2::url\_parse()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidytext)  
library(quanteda)

## Package version: 3.3.0  
## Unicode version: 14.0  
## ICU version: 70.1  
## Parallel computing: 4 of 4 threads used.  
## See https://quanteda.io for tutorials and examples.

library(dplyr)  
library(rvest)  
library(xml2)  
  
texto <- (html\_text(dt))  
texto <- str\_replace\_all(texto,"\\(",")")  
texto <- str\_replace\_all(texto, "\\)","")  
texto <- str\_replace\_all(texto,"\n","")  
texto <- str\_replace\_all(texto,"\t","")  
texto

## [1] " Notícias Justiça determina bloqueio de R$ 399 milhões em bens da Fiji Solutions 14/04/2023 05:46 "   
## [2] " Educação UEPB divulga 3ª chamada da lista de espera do Sisu 2023; veja lista de classificados 14/04/2023 05:49 "   
## [3] " Tecnologia Vazamento de fotos do corpo de Marília Mendonça: advogado explica crimes 14/04/2023 05:08 "   
## [4] " Clima e Tempo Inmet emite alerta de perigo de chuvas intensas para cidades da Paraíba "   
## [5] " Meio Ambiente 8 praias estão impróprias para banho no Litoral da Paraíba "   
## [6] " Clima e Tempo Previsão do tempo para o fim de semana, na Paraíba "   
## [7] " Esportes Série B do Campeonato Brasileiro 2023 tem 14 jogadores paraibanos "   
## [8] " Esportes Nacional de Patos na Série D 2023: confira o novo elenco do Naça "   
## [9] " Clima e Tempo Inmet emite alerta de perigo de chuvas intensas para cidades da Paraíba 14/04/2023 01:05 "   
## [10] " Meio Ambiente 8 praias estão impróprias para banho no Litoral da Paraíba 14/04/2023 03:22 "   
## [11] " Clima e Tempo Previsão do tempo para o fim de semana, na Paraíba 14/04/2023 08:50 "   
## [12] " Esportes Série B do Campeonato Brasileiro 2023 tem 14 jogadores paraibanos 14/04/2023 07:58 "   
## [13] " Esportes Nacional de Patos na Série D 2023: confira o novo elenco do Naça 14/04/2023 11:11 "   
## [14] " Esportes Times da Paraíba: conheça a história dos maiores clubes 09/07/2021 05:33 "  
## [15] " Esportes Botafogo-PB: Juliana Bandeira convoca torcida para apoiar o futebol feminino14/04/2023 06:06 "   
## [16] " Tecnologia WhatsApp anuncia novo recurso para proteção contra roubo de conta14/04/2023 04:14 "   
## [17] " Economia Volume de serviços na PB tem 4ª maior queda do Brasil em janeiro14/04/2023 03:36 "   
## [18] "Carregar mais"   
## [19] " Silvio Osias Turnê Gilberto Gil in Concert passa por João Pessoa nesta quinta-feira 13/04/2023 08:07 "   
## [20] " Conversa Política ‘Blocão’ de Lira na contraofensiva a Pacheco divide os Ribeiro no Congresso 13/04/2023 10:45 "   
## [21] " Pleno Poder Veneziano garante apoio a empréstimos de CG e estreita laços com Bruno Cunha Lima 13/04/2023 07:56 "   
## [22] " Pleno Poder Governo troca aliado de Adriano Galdino por indicado de Lucas Ribeiro na gerência do Amigão 13/04/2023 01:07 "   
## [23] " Esportes Times da Paraíba: conheça a história dos maiores clubes 09/07/2021 05:33 "  
## [24] " Esportes Botafogo-PB: Juliana Bandeira convoca torcida para apoiar o futebol feminino14/04/2023 06:06 "   
## [25] " Tecnologia WhatsApp anuncia novo recurso para proteção contra roubo de conta14/04/2023 04:14 "   
## [26] " Economia Volume de serviços na PB tem 4ª maior queda do Brasil em janeiro14/04/2023 03:36 "   
## [27] " Qual é a Boa? Dia do Café: aprenda a fazer receita de bolo de café14/04/2023 03:22 "   
## [28] " Botafogo-PB Novo goleiro do Botafogo-PB, Mota já foi campeão com o técnico Surian em 2016; relembre14/04/2023 01:42 "   
## [29] " Cultura Inaldo Dantas lança ‘O Livro do Síndico’ em João Pessoa14/04/2023 01:40 "   
## [30] " Notícias Bandeira da Paraíba: a história do ‘Nego’14/04/2023 01:39 "   
## [31] " Comunidade Lei garante sigilo a filhos de mulheres vítimas de violência em transferência escolar14/04/2023 01:09 "   
## [32] " Pleno Poder Alpargatas alega sazonalidade para demissões e mantém planejamento de investimentos na Paraíba14/04/2023 01:06 "   
## [33] " Silvio Osias No palco, com dois filhos, um neto e uma neta, Gil diz pra gente que ainda é preciso virar o mundo em festa, trabalho e pão14/04/2023 12:29 "   
## [34] "Carregar mais"   
## [35] "Qual é a Boa?Bixarte lança o álbum ‘Traviacardo’; veja entrevista"   
## [36] "Qual é a Boa?Agnes Nunes fala sobre turnê internacional"   
## [37] "Ver todos"   
## [38] ""   
## [39] "Conversa PolíticaFatos políticos de forma clara e opinativa. Angélica Nunes e Laerte Cerqueira"   
## [40] "Pleno PoderNotícias da cobertura política de Campina Grande e de todo o interior da Paraíba, por João Paulo Medeiros."   
## [41] "Caderno AnimalFabi e Miguel Cavalcanti abordam comportamento e saúde dos pets. Dicas, curiosidades, tutoria responsável e harmonização entre animais e família."   
## [42] "Sílvio OsiasMuita cultura e uma pitada de entretenimento, com abordagem de obras e questões relevantes da música e do cinema."   
## [43] "Saúde AlertaDicas e orientações para cuidar da saúde e bem estar com o médico André Telis."   
## [44] "Ver mais"   
## [45] ""   
## [46] ""   
## [47] ""   
## [48] ""   
## [49] ""   
## [50] ""   
## [51] ""   
## [52] ""   
## [53] ""   
## [54] ""   
## [55] ""   
## [56] ""

texto <- as.vector(texto)  
  
  
text\_df <- tibble(line=1:56, text= texto)  
  
text\_df

## # A tibble: 56 × 2  
## line text   
## <int> <chr>   
## 1 1 " …  
## 2 2 " …  
## 3 3 " …  
## 4 4 " …  
## 5 5 " …  
## 6 6 " …  
## 7 7 " …  
## 8 8 " …  
## 9 9 " …  
## 10 10 " …  
## # ℹ 46 more rows

**Armazenando dados como data.frame**.

* Salvando dados em disco no formato csv

Se os dados já estiverem armazenados como um data.frame:

write.csv(text\_df,file = "text\_df.csv")

Um objeto *tibble* é uma classe moderna de *data.frame* dentro do R, disponível nos pacotes *dplyr* e *tibble*, que possui um método de impressão conveniente, não converte *strins* em fatores e não usa nomes de linha (MOREIRA; ROCABADO, 2022).

Porém, o objeto *tibble* ainda não está no formato aceito pelo pacote *tidytext*.

Precisaremos converter-lo em outro formato que atenda a condição *one-token-per-document-per-row*.

Portanto, cada *token unigram* (cada palavra) deve ser um valor indicado. Utilizaremos a função *unnest\_tokens* do pacote *tidytext* para realizar o processo de *tokenização*.

**Tokens**

* O formato tidytext

Usar os principios do *tidytext* é uma maneira poderosa de tornar o processamento de dados mais ágil e eficaz. Conforme Wickham (2014), os dados organizados têm uma estrutura especifica:

* Cada variável é uma coluna;
* Cada observação é uma linha;
* Cada tipo de unidade de observação é uma tabela.

Assim, o formato do *tidytext* segue a mesma estrutura apresentada, na qual cada linha/observação possui uma unidade de texto significativa, também chamada por *token*, estes organizadas em uma coluna/variável.

O *token* pode ser uma única palavra, um conjunto de palavras, uma frase ou um parágrafo. O processo consiste em realizar a *tokenização*, em que dividimos o texto em *tokens*.

**Token usando a função unnest\_tokens**

library(tidytext)  
  
text\_token <- text\_df %>%  
 unnest\_tokens(word, text)  
  
text\_token

## # A tibble: 675 × 2  
## line word   
## <int> <chr>   
## 1 1 notícias   
## 2 1 justiça   
## 3 1 determina  
## 4 1 bloqueio   
## 5 1 de   
## 6 1 r   
## 7 1 399   
## 8 1 milhões   
## 9 1 em   
## 10 1 bens   
## # ℹ 665 more rows

Iremos remover elementos da nossa base de dados (caso haja) que não agregam valor a depender da análise. esses lementos são chamados de *stopwords* e podemos removê-los utilizando o pacote *quanteda*.

#stopwords  
library(quanteda)  
  
stop\_w <- tibble(word = stopwords(source = "stopwords-iso", language = "pt"))  
  
#retirar do corpus as stopwords  
tidy\_text <- text\_token %>%   
 anti\_join(stop\_w)

## Joining with `by = join\_by(word)`

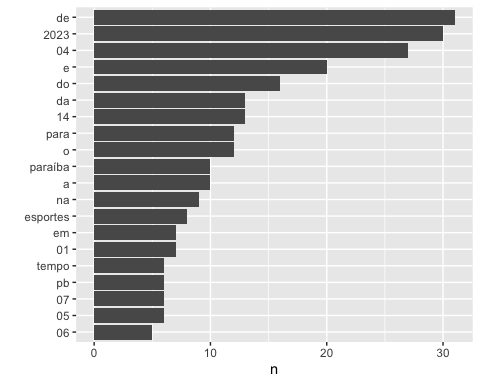
Com a nossa base de dados no formato *tidy* podemos iniciar algumas análises como, por exemplo, a contagem da frequência de palavras ou *tokens*.

text\_token %>%  
 count(word, sort = TRUE)

## # A tibble: 303 × 2  
## word n  
## <chr> <int>  
## 1 de 31  
## 2 2023 30  
## 3 04 27  
## 4 e 20  
## 5 do 16  
## 6 14 13  
## 7 da 13  
## 8 o 12  
## 9 para 12  
## 10 a 10  
## # ℹ 293 more rows

Também podemos visualizar as frequências calculadas:

text\_token %>%  
 count(word, sort = TRUE) %>%  
 mutate(word = fct\_reorder(word, n)) %>%  
 slice(1:20) %>%  
 ggplot(aes(word, n)) +  
 geom\_col() +  
 coord\_flip() +  
 labs(x="")



Com a base de dados em formato *tidy*, podemos utilizar os meta-dados dos documentos em nossas análises.

**Selecionando tokens**

* Tokens sem pontuação e sem números

#Transformando o data.frame em um corpus  
corp <- corpus(texto)  
summary(corp, 10)

## Corpus consisting of 56 documents, showing 10 documents:  
##   
## Text Types Tokens Sentences  
## text1 21 22 1  
## text2 21 25 1  
## text3 18 21 1  
## text4 14 15 1  
## text5 12 12 1  
## text6 14 14 1  
## text7 11 11 1  
## text8 15 15 1  
## text9 21 23 1  
## text10 19 20 1

#Tokens sem pontuação e sem sequência  
toks <- tokens(corp, remove\_punct = T, remove\_numbers = T)

* Tokens sem stopword

# Removendo as stopwords  
toks\_nostop <- tokens\_select(toks, pattern = stopwords('pt'),  
 selection = 'remove')

**Gerando n-grams**

O método de apresentado nâo respeita a ordem em que as palavras foram escritas, no entanto, para algumas análises essa ordem importa. Para garantir essa ordem, podemos criar *tokens* com *N-grams* para garantir que a ordem de palavras esteja presente no processo de *tokenização*.

Através da função *tokens\_ngrams()* é possível aplicar esse método.

# criando n-grams  
toks\_ngram <- tokens\_ngrams(toks, n = 1:4)  
toks\_ngram

## Tokens consisting of 56 documents.  
## text1 :  
## [1] "Notícias" "Justiça" "determina" "bloqueio" "de" "R"   
## [7] "$" "milhões" "em" "bens" "da" "Fiji"   
## [ ... and 34 more ]  
##   
## text2 :  
## [1] "Educação" "UEPB" "divulga" "3ª" "chamada" "da"   
## [7] "lista" "de" "espera" "do" "Sisu" "veja"   
## [ ... and 42 more ]  
##   
## text3 :  
## [1] "Tecnologia" "Vazamento" "de" "fotos" "do"   
## [6] "corpo" "de" "Marília" "Mendonça" "advogado"   
## [11] "explica" "crimes"   
## [ ... and 30 more ]  
##   
## text4 :  
## [1] "Clima" "e" "Tempo" "Inmet" "emite" "alerta"   
## [7] "de" "perigo" "de" "chuvas" "intensas" "para"   
## [ ... and 42 more ]  
##   
## text5 :  
## [1] "Meio" "Ambiente" "praias" "estão"   
## [5] "impróprias" "para" "banho" "no"   
## [9] "Litoral" "da" "Paraíba" "Meio\_Ambiente"  
## [ ... and 26 more ]  
##   
## text6 :  
## [1] "Clima" "e" "Tempo" "Previsão" "do" "tempo"   
## [7] "para" "o" "fim" "de" "semana" "na"   
## [ ... and 34 more ]  
##   
## [ reached max\_ndoc ... 50 more documents ]

**Nuvem de palavras**

Uma forma da visualização de frequência na análise de texto é a nuvem de palavras.

Iremos utilizar a função *textplo\_wordcloud()* do pacote *quanteda*.

library(quanteda.textplots)  
#criando uma DFM com as hashtags  
dfmat\_texto <- dfm(toks)  
set.seed(132)  
textplot\_wordcloud(dfmat\_texto, max\_words = 100)

