Analises_Tese

Ana Lu

9 de setembro de 2018

Análise das entrevistas dos atores da Preferência Hidrossocial

Carregando e preparando o texto

```
library(pdftools)
library(tidyverse)

## -- Attaching packages ---------------------------------------------
----------- tidyverse 1.2.1 --

## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------------
----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(readtext)

textoHidro <- readtext(paste0("Entrevistas_Hidro/*.pdf"))

library(stringr)
library(tm)

## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate

EntrevHidro <- textoHidro %>%
  paste(textoHidro,collapse = " ") %>%
  removeNumbers() %>%
  removePunctuation() %>%
  str_remove_all("\r") %>%
  str_remove_all("\n") %>%
```

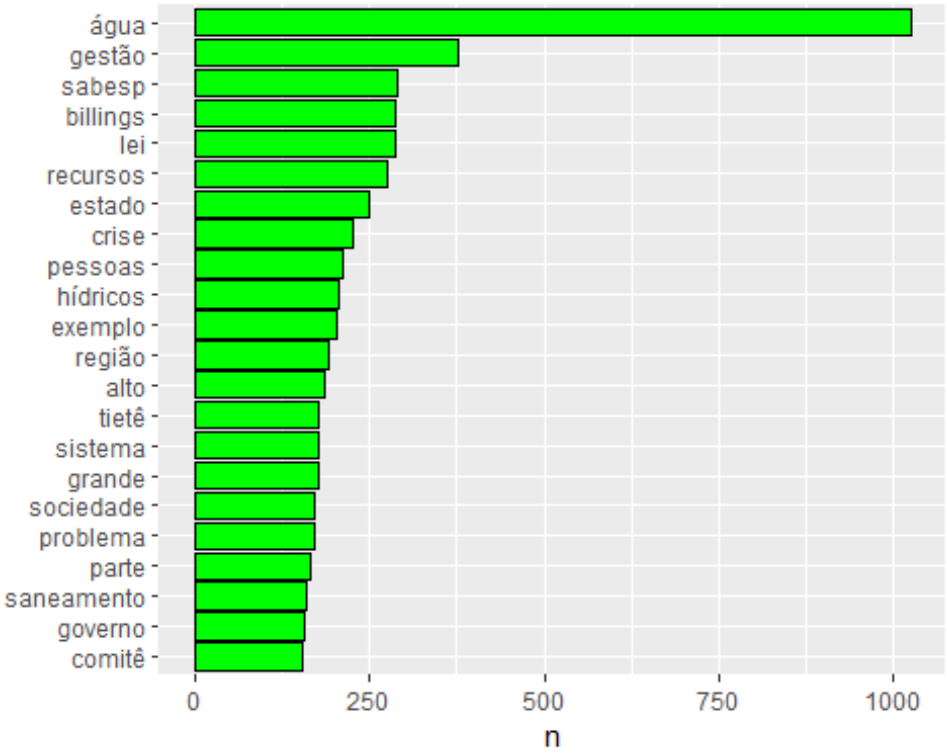
```
  str_to_lower() %>%
  stripWhitespace()
```

Agora estamos trabalhando com um objeto `class(EntrevHidro)`

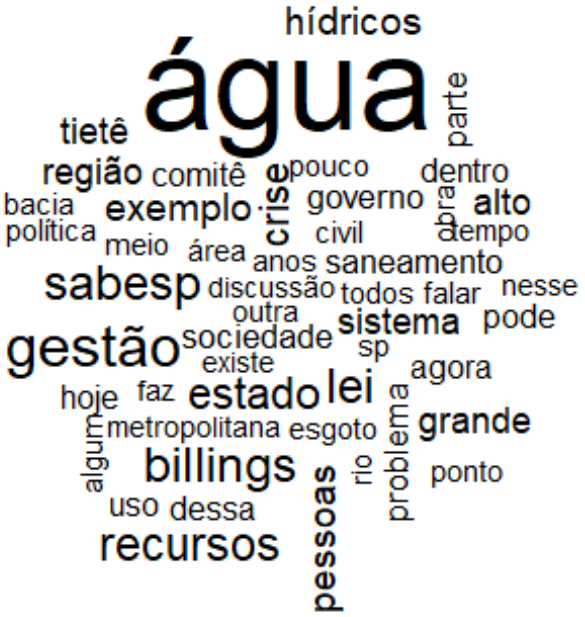
Transformando em dataframe e tokenizando

```
entrevHidro_df <- data_frame(id_discurso = 1:length(EntrevHidro),
                             text = EntrevHidro)

library(tidytext)

entrevHidro_token <- entrevHidro_df %>%
  unnest_tokens(word, text)

stopwords_pt <- c(stopwords("pt"), "que", "é", "entrevistado",
                  "entrevistador", "pra", "porque", "r",
"nentrevistador",
                  "nentrevistado", "n", "questão", "vai", "ai",
                  "aqui", "sobre", "assim", "etc","pois", "desse", "né",
"aí", "paulo",
                  "ainda", "então", "gente", "ser", "joão", "ricardo",
"de", "lá",
                  "acho", "ter", "sim", "coisa", "fazer")

stopwords_pt_df <- data.frame(word = stopwords_pt)

entrevHidro_token <- entrevHidro_token %>%
  anti_join(stopwords_pt_df, by = "word")

## Warning: Column `word` joining character vector and factor, coercing
into
## character vector

entrevHidro_token %>%
  count(word, sort = TRUE)

## # A tibble: 9,977 x 2
##    word         n
##    <chr>    <int>
##  1 água      1024
##  2 gestão     376
##  3 sabesp     290
##  4 billings   288
##  5 lei        286
##  6 recursos   276
##  7 estado     250
##  8 crise      226
##  9 pessoas    210
## 10 hídricos   206
## # ... with 9,967 more rows
```

```
entrevHidro_token %>%
  count(word, sort = TRUE) %>%
  filter(n > 150) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot()+
  geom_col(aes(word, n),colour="black", fill= "green") +
  xlab(NULL) +
  coord_flip()
```




```
library(wordcloud)

## Loading required package: RColorBrewer

entrevHidro_token %>%
  count(word, sort = T) %>%
  with(wordcloud(word, n, use.r.layout = TRUE, max.words = 50))
```


hídricos
água
tietê
parte
região
comitê
pouco
dentro
bacia
exemplo
crise
governo
alto
política
meio
área
civil
tempo
anos
saneamento
sabesp
discussão
todos
falar
nesse
outra
sistema
pode
gestão
sociedade
sp
existe
agora
hoje
faz
estado
lei
problema
metropolitana
esgoto
grande
algum
billings
rio
ponto
uso
dessa
pessoas
recursos