# Mineração de texto em pedidos de Lei de Acesso à informação - LAI

# Packages for this routine

## BASE DE DADOS

## Importação dos dados

Caminho do projeto

```
PATH = "..;/proj_eSIC_v10/textmining_pt/DATA/"
```

• Pedidos e-SIC

Respostas e-SIC

```
FILE1 = "relatorio_respostas.xlsx"
db1_raw = readxl::read_excel(paste0(PATH,FILE1), sheet = "DADOS", col_names = TRUE);
# dim(db1_raw); names(db1_raw)
colnames(db1_raw) = c("ID", "DATA", "SOLICITACAO", "DIRETORIA", "DATA_RESPOSTA")
#View(head(db1_raw))
LAI1 = db1_raw
```

• Stopwords

```
FILE2 = "stopwords_PT_FINAL.csv"
stopwords_pt = read.csv(paste0(PATH,FILE2), sep = ';', header = F, encoding = "UTF-8")
stopwords_pt = stopwords_pt[,-2];
cat(paste0("0 nosso vetor de stopwords contém ",length(stopwords_pt), " palavras únicas"))
## O nosso vetor de stopwords contém 605 palavras únicas
## dim(stopwords_pt); class(stopwords_pt)
stopwords_pt = as.character(stopwords_pt)
stopwords_pt[1:14]
  [1] ","
                  "a"
                            "à"
                                                                     "aí"
##
                                       "acerca"
                                                 "adeus"
                                                           "agora"
```

"algmas"

"algo"

"algumas" "alguns"

## Pré-processamento dos dados

"alem"

#### Pedidos por diretoria

[8] "ainda"

• Tabela 01 número de solcitações/pedidos de informação

"além"

```
LAI1 %>%
count(DIRETORIA, sort = TRUE, name = "total_pedidos") %>%
```

Table 1: Quantitativo de solicitações por Diretoria/EPE via e-SIC

| DIRETORIA | total_pedidos |
|-----------|---------------|
| DEA       | 210           |
| DEE       | 197           |
| DGC       | 115           |
| DPG       | 24            |
| OUTROS    | 19            |
| SIC       | 1             |

```
diretorias0 = levels(as.factor(LAI1$DIRETORIA))
```

Verificamos a existência de 5 diretorias, sendo elas: *DEA*, *DEE*, *DGC*, *DPG*, *SIC* e *OUTROS*. Essa última é devido a existência de informações solicitadas que não são de competência direta de nenhuma das cinco diretorias, daí a necessidade de uma última categoria *OUTROS* para atender essas demandas.

A seguir, um passo importante de reclassificação será executado devido ao número pequeno de solicitações para a diretoria SIC. Apenas uma solcitação existente no nosso banco de dados para essa diretoria. Iremos, portanto, unificar essa demanda à categoria OUTROS.

• Respostas e-SIC - Reclassificação Diretorias

```
LAI1 = LAI1 %>%
  mutate(DIRETORIA = ifelse(DIRETORIA == diretorias0[6], diretorias0[5], DIRETORIA))
diretorias = levels(as.factor(LAI1$DIRETORIA))
#dim(LAI1)
#View(head(LAI1))
```

• Tabela 02 número de solcitações/pedidos de informação - após reclassificação

Table 2: Quantitativo de solicitações por Diretoria/EPE via e-SIC - após reclassificação

| DIRETORIA | $total\_pedidos$ |
|-----------|------------------|
| DEA       | 210              |
| DEE       | 197              |
| DGC       | 115              |
| DPG       | 24               |
| OUTROS    | 20               |

As constatações anteriores foram feitas apenas na base de dados referente às respostas, donde temos a classificação das diretorias responsáveis por responder cada uma das demandas. É necessário, agora, unificar as bases de dados pertinentes a solicitações e respostas.

#### Unificando as duas bases

```
LAI = LAI %>% select(-DATA_RESPOSTA); #dim(LAI)

LAI1 = LAI1 %>% select(-DATA); #dim(LAI1)

DB = left_join(x = LAI, y = LAI1, by = "ID")

#View(head(DB))
```

Ver Anexo 01 c/ amostra dos dados da tabela que serpá utilizada para manipulação daqui pra frente.

## Mineração de texto

• Palavras por diretoria

```
library(tidytext)
diretoria_palavras <- DB %>%
  unnest_tokens(palavra, DESCRI_PEDIDO) %>%
  count(DIRETORIA, palavra, sort = TRUE) %>%
  ungroup()
```

Iniciamos as manipulações utilizando recursos da função unnest\_tokens() do pacote library(tidytext) que nos permite trabalhar com textos em um formato tidy que coloca uma palavra por linha e cada coluna um conjunto de caracteres de text a serem separados por palavra, formando, assim, termos/palavras por linha. Utilizamos, aindam dos recursos do pacote library(diplyr) para agrupar esses termos por diretoria e quantificar as suas repetições.

A tabela a seguir mostra a frequência das 10 palavras de maior ocorrência de todos os pedidos, agregadados por diretoria.

• Tabela 03 Palavras mais frequentes no conjunto de solicitações

Table 3: Palavras mais frequentes no conjunto de solicitações

| DIRETORIA | palavra | n     |
|-----------|---------|-------|
| DEA       | de      | 1'127 |
| DEE       | de      | 979   |
| DGC       | de      | 736   |
| DEE       | a       | 364   |
| DEA       | a       | 350   |
| DEA       | e       | 329   |
| DGC       | a       | 304   |
| DEE       | e       | 273   |
| DGC       | e       | 266   |
| DEA       | О       | 262   |

Verificamos que exatamente as 10 palavras mais frequentes em todos os pedidos realizados são palavras sem muito interesse contextual pois não acrescentam nenhum sentido semântico, são essas: preposição (de), conjunção (e) e artigos(o,a). Veremos mais a frente (indicar sessão) como remover essas palavras que são ditas *stopwords* e trabalhar apenas com palavras de sentido maior semântico, acrescentando assim maior assertividade na classificação do nosso modelo a ser proposto no capítulo (indicar capítulo).

• Total de palavras

```
total_palavras = diretoria_palavras %>%
group_by(DIRETORIA) %>%
summarize(total_palavras = sum(n))
```

• Tabela 04 Total de palavras por diretoria

Table 4: Palavras mais frequentes no conjunto de solicitações

| DIRETORIA | total_palavras |
|-----------|----------------|
| DEA       | 14.079         |
| DEE       | 12.907         |
| DGC       | 10.355         |
| DPG       | 1.434          |
| OUTROS    | 1.022          |

É importante ressaltar aqui, a diferença extrema entre o número de palavras existente por diretoria, isso se dá devido ao número de pedidos realizados por diretoria já constatado anteriormente. Temos 210 solicitações registradas para a DEA, 197 para a DEE, 115 para a DGC e, apenas, 24 e 20 pedidos para a DPG e OUTROS, respectivamente.

Vamos, portanto, visualizar o número médio de palavras por pedido e diretoria. Para isso, vamos pegar o total de palavras por diretoria e dividir pelo total de pedidos por diretoria.

• Tabela 04 Número médio de palavras por pedido e diretoria

Table 5: Palavras mais frequentes no conjunto de solicitações

| DIRETORIA | $total\_pedidos$ | total_palavras | palavras_por_pedido |
|-----------|------------------|----------------|---------------------|
| DEA       | 210              | 14'079         | 67,04               |
| DEE       | 197              | 12'907         | $65,\!52$           |
| DGC       | 115              | 10'355         | 90,04               |
| DPG       | 24               | 1'434          | 59,75               |
| OUTROS    | 20               | 1'022          | 51,10               |

- Junta informações

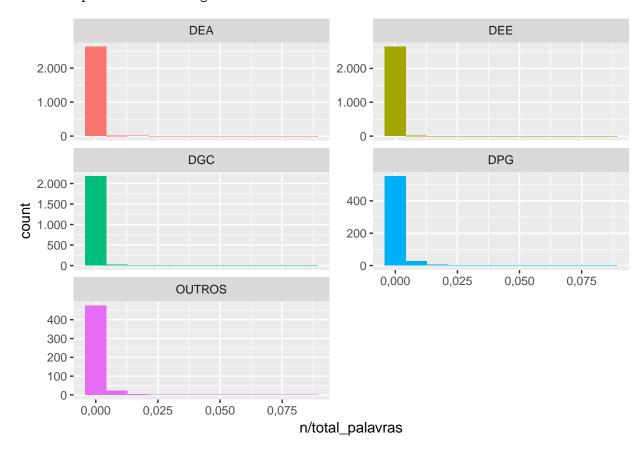
```
diretoria_palavras = left_join(diretoria_palavras, total_palavras, by = "DIRETORIA")
```

• Distribuição do nº de palavras usadas em solicitações por diretoria (histograma)

```
library(ggplot2)
gcomma <- function(x) format(x, big.mark = ".", decimal.mark = ",", scientific = FALSE)

ggplot(diretoria_palavras, aes(n/total_palavras, fill = DIRETORIA)) +
geom_histogram(show.legend = FALSE, binwidth = .0085) + xlim(NA, .025) +
facet_wrap(~DIRETORIA, ncol = 2, scales = "free_y") +
    scale_y_continuous(labels=gcomma) +
    scale_x_continuous(labels=gcomma)</pre>
```

## Scale for 'x' is already present. Adding another scale for 'x', which ## will replace the existing scale.



• Palavras mais frequentes por diretoria

```
PROP_PALAVRA = diretoria_palavras %>%
    mutate(palavra = str_extract(palavra, "[a-z']+")) %>%
    count(DIRETORIA, palavra) %>%
    group_by(DIRETORIA) %>%
    mutate(proportion = n / sum(n)) %>%
    select(-n) %>%
    spread(DIRETORIA, proportion)
```

## Gráficos de comparação de frequência de palavras por diretorias (2 a 2)

## COM STOPWORDS

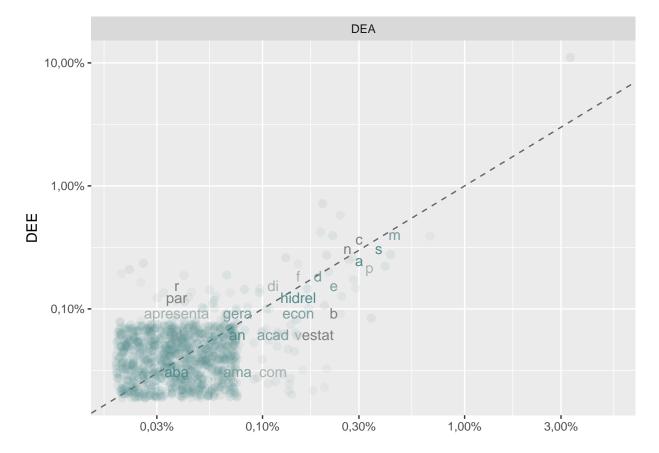
É importante ressaltar que os gráficos a seguir mostram, apenas, a comparação de frequência de palavras

existentes em ambas diretorias. Ou seja, palavras existentes em apenas uma diretoria serão desconsideradas para a geração destes.

#### • DEE X DEA

```
freq00 <- PROP PALAVRA %>%
   gather(DIRETORIA, proportion, c(`DEA`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq00, aes(x = proportion, y = `DEE`,
                        color = abs(`DEE` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
    geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
   scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
   labs(y = "DEE", x = NULL)
```

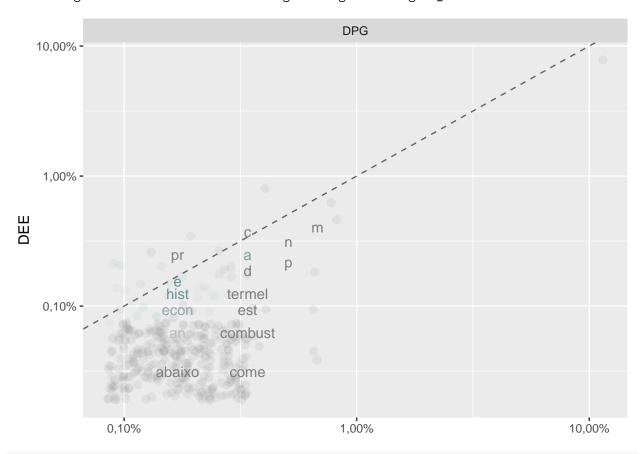
- ## Warning: Removed 3487 rows containing missing values (geom\_point).
- ## Warning: Removed 3488 rows containing missing values (geom\_text).



#### • DEE X DPG

```
freq01 <- PROP_PALAVRA %>%
    gather(DIRETORIA, proportion, c(`DPG`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq01, aes(x = proportion, y = `DEE`,
                        color = abs(`DEE` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
   geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
   scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
   labs(y = "DEE", x = NULL)
```

- ## Warning: Removed 4235 rows containing missing values (geom\_point).
- ## Warning: Removed 4236 rows containing missing values (geom\_text).

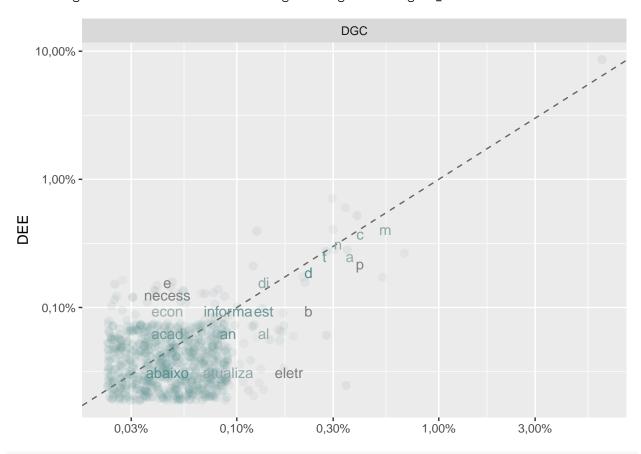


## Warning messages:

- 1: Removed 4235 rows containing missing values (geom\_point).
- 2: Removed 4236 rows containing missing values (geom\_text).
  - DEE X DGC

```
freq02 <- PROP_PALAVRA %>%
    gather(DIRETORIA, proportion, c(`DGC`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq02, aes(x = proportion, y = `DEE`,
                        color = abs(`DEE` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
   geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
   labs(y = "DEE", x = NULL)
```

- ## Warning: Removed 3794 rows containing missing values (geom\_point).
- ## Warning: Removed 3795 rows containing missing values (geom\_text).



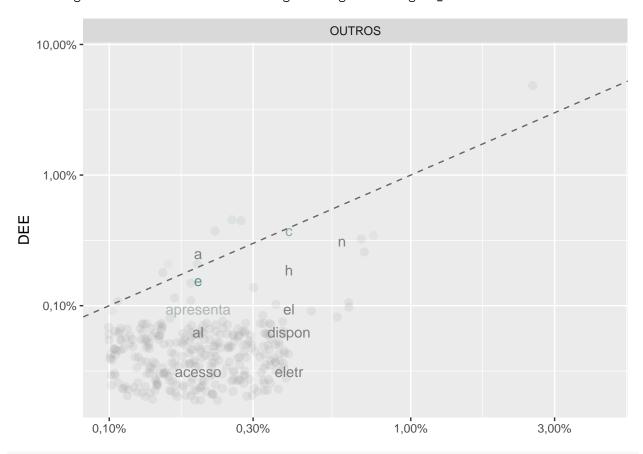
## Warning messages:

- 1: Removed 3794 rows containing missing values (geom\_point).
- 2: Removed 3795 rows containing missing values (geom\_text).

#### • DEE X OUTROS

```
freq03 <- PROP_PALAVRA %>%
    gather(DIRETORIA, proportion, c(`OUTROS`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq03, aes(x = proportion, y = `DEE`,
                        color = abs(`DEE` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
   geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
   scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
   labs(y = "DEE", x = NULL)
```

- ## Warning: Removed 4273 rows containing missing values (geom\_point).
- ## Warning: Removed 4274 rows containing missing values (geom\_text).



```
Warning messages:

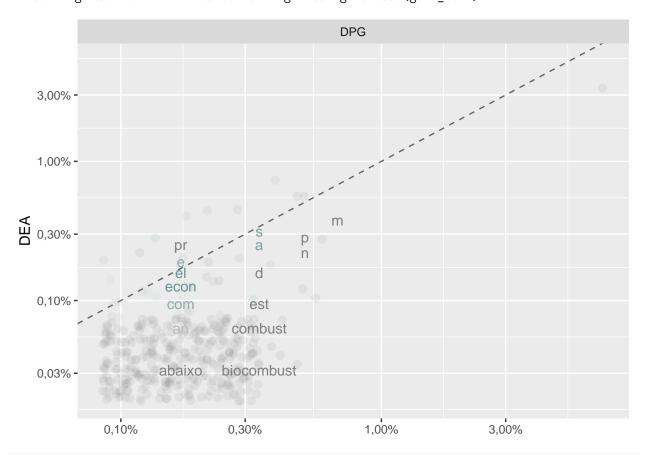
1: Removed 4273 rows containing missing values (geom_point).

2: Removed 4274 rows containing missing values (geom_text).
```

• DEA X DPG

```
freq04 <- PROP_PALAVRA %>%
    gather(DIRETORIA, proportion, c(`DPG`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq04, aes(x = proportion, y = `DEA`,
                        color = abs(`DEA` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
    geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
    labs(y = "DEA", x = NULL)
```

- ## Warning: Removed 4221 rows containing missing values (geom\_point).
- ## Warning: Removed 4222 rows containing missing values (geom\_text).



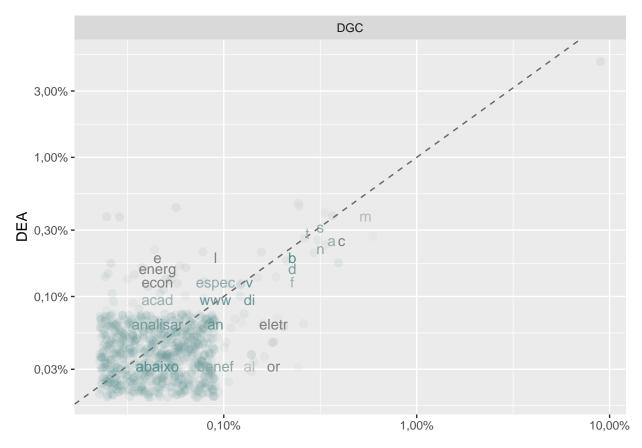
## Warning messages:

- 1: Removed 4221 rows containing missing values (geom\_point).
- 2: Removed 4222 rows containing missing values (geom\_text).

#### • DEA X DGC

```
freq05 <- PROP_PALAVRA %>%
    gather(DIRETORIA, proportion, c(`DGC`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq05, aes(x = proportion, y = `DEA`,
                        color = abs(`DEA` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
    geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
    labs(y = "DEA", x = NULL)
```

- ## Warning: Removed 3812 rows containing missing values (geom\_point).
- ## Warning: Removed 3813 rows containing missing values (geom\_text).

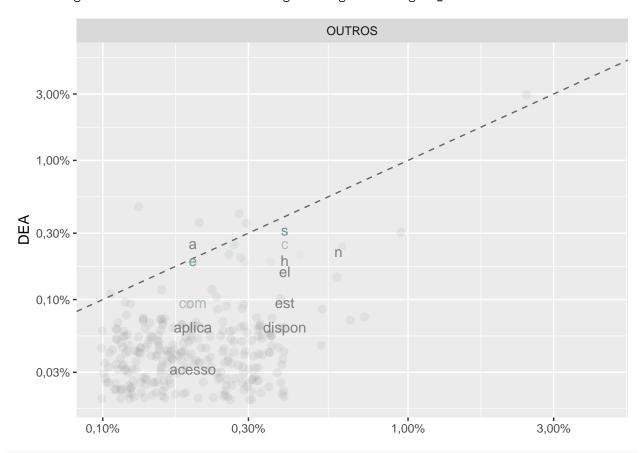


```
Warning messages:
1: Removed 3812 rows containing missing values (geom_point).
2: Removed 3813 rows containing missing values (geom_text).
```

• DEA X OUTROS

```
freq06 <- PROP_PALAVRA %>%
    gather(DIRETORIA, proportion, c(`OUTROS`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq06, aes(x = proportion, y = `DEA`,
                        color = abs(`DEA` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
    geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
   labs(y = "DEA", x = NULL)
```

- ## Warning: Removed 4303 rows containing missing values (geom\_point).
- ## Warning: Removed 4304 rows containing missing values (geom\_text).



```
Warning messages:

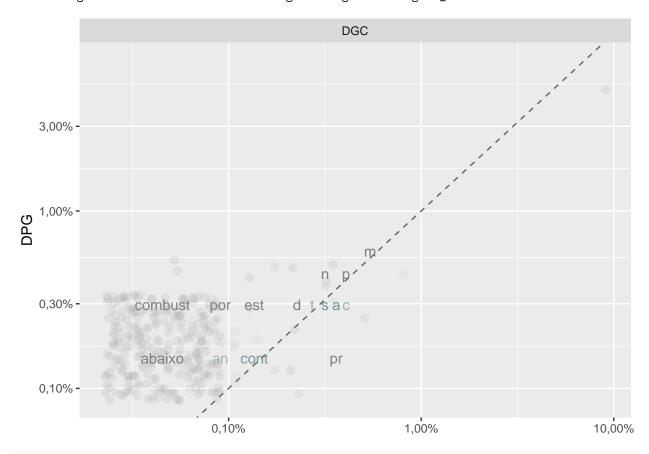
1: Removed 4303 rows containing missing values (geom_point).

2: Removed 4304 rows containing missing values (geom_text).
```

• DPG X DGC

```
freq07 <- PROP_PALAVRA %>%
    gather(DIRETORIA, proportion, c(`DGC`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq07, aes(x = proportion, y = `DPG`,
                        color = abs(`DPG` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
   geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
   labs(y = "DPG", x = NULL)
```

- ## Warning: Removed 4296 rows containing missing values (geom\_point).
- ## Warning: Removed 4297 rows containing missing values (geom\_text).

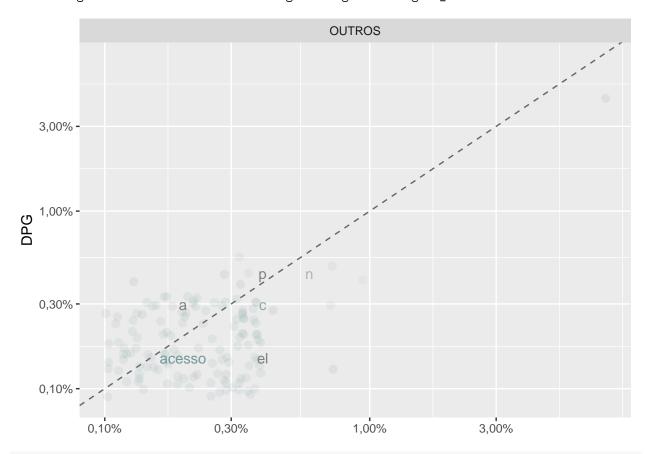


```
Warning messages:
1: Removed 4296 rows containing missing values (geom_point).
2: Removed 4297 rows containing missing values (geom_text).
```

• DPG X OUTROS

```
freq08 <- PROP_PALAVRA %>%
    gather(DIRETORIA, proportion, c(`OUTROS`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq08, aes(x = proportion, y = `DPG`,
                        color = abs(`DPG` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
   geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
   scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
   scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
   labs(y = "DPG", x = NULL)
```

- ## Warning: Removed 4450 rows containing missing values (geom\_point).
- ## Warning: Removed 4451 rows containing missing values (geom\_text).

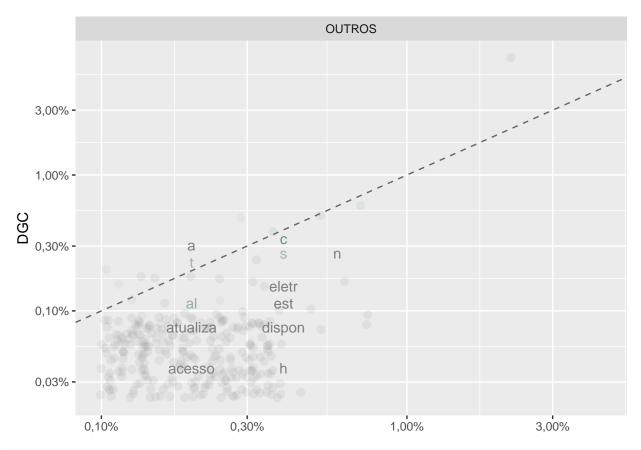


```
Warning messages:
1: Removed 4450 rows containing missing values (geom_point).
2: Removed 4451 rows containing missing values (geom_text).
```

• DPG X OUTROS

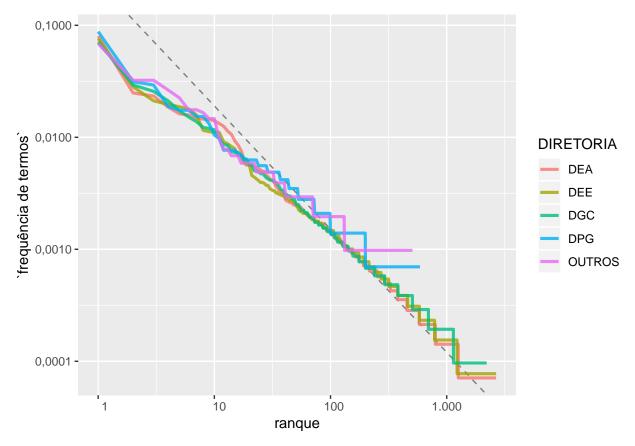
```
freq09 <- PROP_PALAVRA %>%
    gather(DIRETORIA, proportion, c(`OUTROS`))
  library(scales)
  # expect a warning about rows with missing values being removed
  ggplot(freq08, aes(x = proportion, y = `DGC`,
                        color = abs(`DGC` - proportion))) +
   geom_abline(color = "gray40", lty = 2) +
   geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
    geom_text(aes(label = palavra), check_overlap = TRUE, vjust = 1.5) +
   scale_x_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_y_log10(labels = percent_format(big.mark = ".", decimal.mark = ",")) +
    scale_color_gradient(limits = c(0, 0.001),
                         low = "darkslategray4", high = "gray75") +
   facet_wrap(~DIRETORIA, ncol = 1) +
   theme(legend.position="none") +
    labs(y = "DGC", x = NULL)
```

- ## Warning: Removed 4302 rows containing missing values (geom\_point).
- ## Warning: Removed 4303 rows containing missing values (geom\_text).



```
Warning messages:
1: Removed 4302 rows containing missing values (geom_point).
2: Removed 4303 rows containing missing values (geom_text).
```

• Zipf's law



# Frequência de palavras por diretoria

```
diretoria_palavras <- DB %>%
  unnest_tokens(palavra, DESCRI_PEDIDO) %>%
  count(DIRETORIA, palavra, sort = TRUE) %>%
  ungroup()
#diretoria_palavras

plot_diretoria_palavras <- diretoria_palavras %>%
  bind_tf_idf(palavra, DIRETORIA, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(palavra = factor(palavra, levels = rev(unique(palavra)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA, levels = c("DEA",
```

```
"DEE",
                                                                              "DGC",
                                                                              "DPG",
                                                                              "OUTROS")))
#View(head(plot_diretoria_palavras))
#jpeg("02_freq_palavras_dir.jpeg")
plot_diretoria_palavras %>%
group_by(DIRETORIA) %>%
top_n(10, tf_idf) %>%
ungroup() %>%
mutate(palavra = reorder(palavra, tf_idf)) %>%
ggplot(aes(palavra, tf_idf, fill = DIRETORIA)) +
geom_col(show.legend = FALSE) +
labs(x = NULL, y = "tf-idf") +
facet_wrap(~DIRETORIA, ncol = 2, scales = "free") +
coord_flip() +
scale_y_continuous(labels=gcomma)
                                         DEA
                                                                                                            DEE
                                                                            leilão -
      consumo -
                                                                              dee -
       balanço -
                                                                           leilões -
       elétrica -
                                                                            eólica -
     energético -
                                                                               nt -
         carga -
                                                                               r0 -
        região -
                                                                            região -
        estado -
                                                                            fontes -
        fontes -
                                                                          geração -
     município -
        mensal -
                                                                            usina -
              0,0000
                         0,0005
                                    0,0010
                                              0,0015
                                                         0,0020
                                                                                  0,000
                                                                                                     0,002
                                                                                                                       0,004
                                                                                                            DPG
                                        DGC
   funcionários -
                                                                        gasodutos -
      entidade -
                                                                           acúcar -
                                                                            etanol -
      concurso -
         cargo -
                                                                            m.pol -
   empregados -
                                                                          projeção -
        cargos -
                                                                              gás -
                                                                         produção -
       locação -
       salários -
                                                                           bacias -
      contratos -
                                                                        exploração -
       licitação -
                                                                           passar -
      requeiro -
                                                                              sal -
     servidores -
                                                                        toneladas -
                                                                                                0,0025
               0,000
                                 0,002
                                                    0,004
                                                                                 0,0000
                                                                                                              0,0050
                                                                                                                            0,0075
                                      OUTROS
labex -
comissão -
embrapa -
pesquisadores -
autos -
invertido -
prevenção -
trabalhista -
validação -
acordos -
advogados -
cadastramento -
cooperação -
corrente -
profissional -
```

0,000 0,002 0,004 0,006 tf-idf
#dev.off()

#### Filtrando um pedaço de texto

```
DB %>%
filter(str_detect(DESCRI_PEDIDO, "r0")) %>%
select(DESCRI_PEDIDO) %>%
head()
```

```
##
## 1
## 2
## 3 Solicitamos para nossa análise cópias dos relatórios nºs EPE-DEE-RE-147/2008-r0 que trata dos ESTU
## 4
## 5
## 6
```

Uma limpeza removendo palavras sem significado semântico (stopwords) pode auxiliar o algoritmo a retornar palavras ainda mais acertivas

#### Radicais

Podemos diminuir redundâncias por parte do algoritmo ensinando-o a compreender palavras que podem estar escritas de forma diferente mas que em significado semântico são semelhantes. Para isso, analisamos o radical de palavras com um mesmo prefixo mas com sufixos diferentes seja por quisistos como gênero ou plural.

#### Exemplos:

leilão  $\propto$  leilões estado  $\propto$  estados região  $\propto$  regiões

Falta implementar

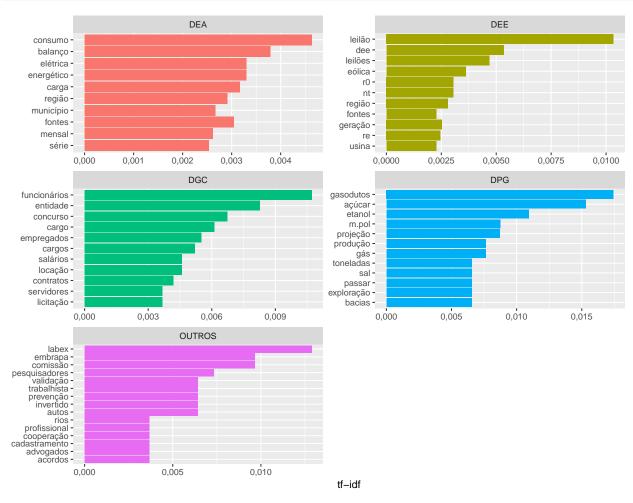
#### Stopwords

Com o arquivo de **stopwords**previamente inserido vamos, primeiramente, transforma-lo em um data\_frame a fim de futuramente utilizá-lo para extrair do texto palavras em comum.

#### Freq. de palavras sem stopwords por diretoria

```
mystopwords <- data_frame(palavra = stopwords_pt)</pre>
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
diretoria_palavras_noSTOP <- anti_join(diretoria_palavras, mystopwords, by = "palavra")
#View(head(diretoria_palavras_noSTOP))
\#diretoria\_palavras\_noSTOP\_noSTOP
plot_diretoria_palavras_noSTOP <- diretoria_palavras_noSTOP %>%
  bind_tf_idf(palavra, DIRETORIA, n) %>%
  arrange(desc(tf idf)) %>%
  mutate(word = factor(palavra, levels = rev(unique(palavra)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA, levels = c("DEA",
                                                    "DEE",
                                                    "DGC",
                                                   "DPG",
                                                   "OUTROS")))
\#plot\_diretoria\_palavras\_noSTOP
#windows.options(width=10, height=10)
#jpeq("03_freq_palavras_dir_nostop.jpeq")
plot_diretoria_palavras_noSTOP %>%
group_by(DIRETORIA) %>%
```

```
top_n(10, tf_idf) %>%
ungroup() %>%
mutate(palavra = reorder(palavra, tf_idf)) %>%
ggplot(aes(palavra, tf_idf, fill = DIRETORIA)) +
geom_col(show.legend = FALSE) +
labs(x = NULL, y = "tf-idf") +
facet_wrap(~DIRETORIA, ncol = 2, scales = "free") +
coord_flip() +
scale_y_continuous(labels=gcomma)
```



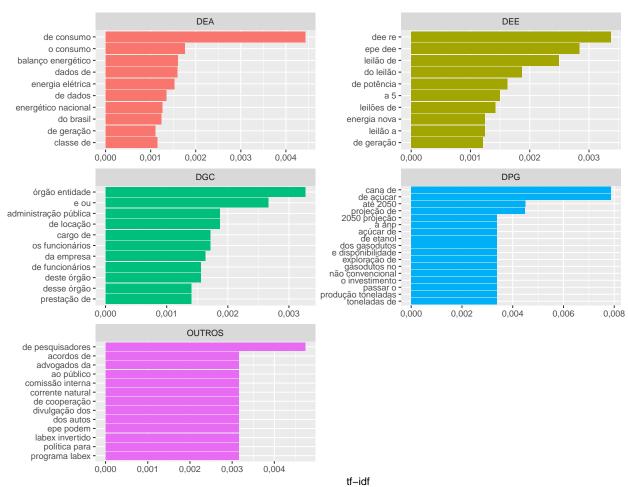
#dev.off()

Usando bigram para n=2 palavras por token

## Frequência de palavras por diretoria

```
diretoria_palavras_bigram <- DB %>%
  select(DESCRI_PEDIDO,DIRETORIA) %>%
  unnest_tokens(BIGRAM, DESCRI_PEDIDO, token = "ngrams", n = 2) %>%
  count(DIRETORIA, BIGRAM, sort = TRUE) %>%
  ungroup()
#diretoria_palavras_bigram
```

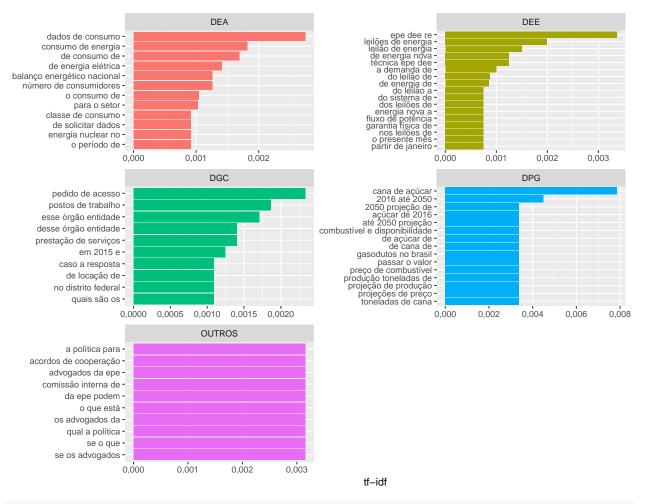
```
plot_diretoria_palavras_bigram <- diretoria_palavras_bigram %>%
  bind_tf_idf(BIGRAM, DIRETORIA, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(BIGRAM = factor(BIGRAM, levels = rev(unique(BIGRAM)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA, levels = c("DEA",
                                                   "DEE".
                                                   "DGC",
                                                   "DPG",
                                                   "OUTROS")))
#View(head(plot_diretoria_palavras_bigram))
#jpeg("02_freq_palavras_dir.jpeg")
plot_diretoria_palavras_bigram %>%
group_by(DIRETORIA) %>%
top_n(10, tf_idf) %>%
ungroup() %>%
mutate(BIGRAM = reorder(BIGRAM, tf_idf)) %>%
ggplot(aes(BIGRAM, tf_idf, fill = DIRETORIA)) +
geom_col(show.legend = FALSE) +
labs(x = NULL, y = "tf-idf") +
facet_wrap(~DIRETORIA, ncol = 2, scales = "free") +
coord flip() +
scale_y_continuous(labels=gcomma)
```



### Usando bigram para n=3 palavras por token

Frequência de palavras por diretoria

```
diretoria_palavras_trigram <- DB %>%
  select(DESCRI_PEDIDO,DIRETORIA) %>%
  unnest_tokens(TRIGRAM, DESCRI_PEDIDO, token = "ngrams", n = 3) %>%
  count(DIRETORIA, TRIGRAM, sort = TRUE) %>%
  ungroup()
\#diretoria\_palauras\_trigram
plot_diretoria_palavras_trigram <- diretoria_palavras_trigram %>%
  bind_tf_idf(TRIGRAM, DIRETORIA, n) %>%
  arrange(desc(tf idf)) %>%
  mutate(TRIGRAM = factor(TRIGRAM, levels = rev(unique(TRIGRAM)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA, levels = c("DEA",
                                                   "DEE",
                                                   "DGC",
                                                   "DPG",
                                                   "OUTROS")))
\#View(head(plot\_diretoria\_palavras\_trigram))
#jpeg("02_freq_palavras_dir.jpeg")
plot_diretoria_palavras_trigram %>%
group_by(DIRETORIA) %>%
top_n(10, tf_idf) %>%
ungroup() %>%
mutate(TRIGRAM = reorder(TRIGRAM, tf_idf)) %>%
ggplot(aes(TRIGRAM, tf_idf, fill = DIRETORIA)) +
geom_col(show.legend = FALSE) +
labs(x = NULL, y = "tf-idf") +
facet wrap(~DIRETORIA, ncol = 2, scales = "free") +
coord_flip() +
scale_y_continuous(labels=gcomma)
```



#### #dev.off()

## tidy object into document-term matrix

#### Nuvem de palavras

Nuvem de palavras por diretoria - s/ steeming e/ c/ stopwords - onegram

```
#View(head(plot_diretoria_palavras))
library(wordcloud)
```

encontraragentes capacidade cmo 16a5 cadastrada térmica nossa medição matemático 2medidos favor houver planejamento déficit estação dia subestações planejamento certificação via la estação dia subestações planejamento atende o conexão 230 ventos apresenta todas fevereiro atende o conexão o conexão marginal novembro de control de co certificaçãoviana pchutilizandorondônia grato state Solar tv suprimento state of solar tv suprimento expansão of s atrasos to atrasos despacho de spacho or2ama mme cálculos usinas física rev.o regime ansmissão porto regin otinas otinas porto regin porto regin per o offontes usina participare participare sa interligados su interligados sus trabalho municípic elétrico 3527 mwmed ou pa oo oo peol dastrados Na Legislation of the projectors era 8°% <u>آ.</u>ي grande definida empreendimentos por reposição estados riachão base por carga explicitado parecer descrita senheros 2026 parecer descrita senhores nordeste série boletins antônio cadastramento

```
## DGC
#jpeg("XX_wordclou_tfidf_dir02_DGC.jpeg")
nuvem2 =
plot_diretorias_tf_dif %>%
  filter(DIRETORIA == "DGC") %>%
  select(-DIRETORIA, word = palavra, freq = tf_idf) %>%
  #top_n(150, freq) %>%
  as.data.frame()
```

```
etrica migo estadoregião estado
```

```
mana petroleo achou biocombustivos petroleo achon piocombustivos petroleo achon p
```



```
#View(head(plot_diretoria_palavras))
library(wordcloud2)
plot_diretorias_tf_dif = plot_diretoria_palavras %>%
    select(palavra, tf_idf, DIRETORIA) %>%
    mutate(palavra = reorder(palavra, tf_idf))
## DEE
\#jpeg("XX\_wordclou\_tfidf\_dir01\_DEE.jpeg")
set.seed(233115)
plot_diretorias_tf_dif %>%
  filter(DIRETORIA == "DEE") %>%
  top_n(150, tf_idf) %>%
  wordcloud2(shuffle = TRUE,
             color = "random-dark",
             shape = "circle")
## DGC
\#jpeg("XX\_wordclou\_tfidf\_dir01\_DGC.jpeg")
set.seed(233115)
plot_diretorias_tf_dif %>%
  filter(DIRETORIA == "DGC") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## DEA
```

```
#jpeg("XX_wordclou_tfidf_dir01_DEA.jpeg")
set.seed(233115)
plot_diretorias_tf_dif %>%
  filter(DIRETORIA == "DEA") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## DPG
#jpeg("XX_wordclou_tfidf_dir04_DPG.jpeg")
set.seed(233115)
plot_diretorias_tf_dif %>%
  filter(DIRETORIA == "DPG") %>%
  top n(150, tf idf) %>%
  wordcloud2()
## OUTROS
#jpeg("XX_wordclou_tfidf_dir01_OUTROS.jpeg")
set.seed(233115)
plot_diretorias_tf_dif %>%
  filter(DIRETORIA == "OUTROS") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
->
Nuvem de palavras por diretoria - s/ steeming e/ou remoção de stopwords - bigram
plot diretorias tf dif bigram = DB %>%
  select(DESCRI PEDIDO, DIRETORIA) %>%
  unnest_tokens(BIGRAM, DESCRI_PEDIDO, token = "ngrams", n = 2) %>%
  count(DIRETORIA, BIGRAM, sort = TRUE) %>%
  bind_tf_idf(BIGRAM, DIRETORIA, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(BIGRAM = factor(BIGRAM, levels = rev(unique(BIGRAM)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA,levels=c("DEA","DEE","DGC","DPG","OUTROS"))) %>%
  select(BIGRAM, tf_idf, DIRETORIA)
## DEE
#jpeq("XX_wordclou_tfidf_dir01_DEE.jpeq")
nuvem1.2 =
plot_diretorias_tf_dif_bigram %>%
  filter(DIRETORIA == "DEE") %>%
  select(-DIRETORIA, word = BIGRAM,freq = tf_idf) %>%
  #top_n(150, freq) %>%
  as.data.frame()
set.seed(231321)
wordcloud(words = nuvem1.2$word, freq = nuvem1.2$freq, min.freq = 0.2,
          max.words=250, random.order=FALSE, rot.per=0.35,
```

colors=brewer.pal(10, "Dark2"))

```
de atendimento breve retorno de mestrado
cálculo da
               dados do dos leilões copia dos
a memória
                  habilitação técnicaleilões a
 gno leilão
            ĩO
               2016 r0 nota técnica nº epe
             a
                                     da análise of fluxo de 2
    Ō
    energi
gdos estudos
Morio
Partir
                                estado de
     de empreendimentos
                                de entrada
ao estado
            de transmissão
                                técnica epe
   dos empreendimentos
                               energia eólica
    belo monte está definidaart 3º
```

```
sopro de souro distrito do tesouro sopro de solo de locargos de houve a la qual o de solo de solo de solo de solo de locação publica de locação de locação
```

```
## DEA
#jpeg("XX_wordclou_tfidf_dir03_DEA.jpeg")
nuvem3.2 =
plot_diretorias_tf_dif_bigram %>%
    filter(DIRETORIA == "DEA") %>%
    select(-DIRETORIA, word = BIGRAM,freq = tf_idf) %>%
    #top_n(150, freq) %>%
    as.data.frame()

set.seed(543453)
wordcloud(words = nuvem3.2$word, freq = nuvem3.2$freq, min.freq = 0.2,
        max.words=250, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(10, "Dark2"))
```

```
aos custos dos anos
                                     no mundo
  e ar condicionado
                     residencial e
                                       do estado
                                      anuário estatistico
  consumo residencial por classe
                                    \overline{\Phi}
  de mestrado
                consumo mensal
  gno balanço
                    da bahia não é
    2010 a período de a 2017
  nuclear no de uma do brasil
  consumo de dados de
                                          ago
     mensal de O
                                             Ø
                                       efaixa
base d
                            é a e
   estado de classe de 1990 a consumo por 8
                            2007
    4
  união e solicitar dados
e outras
bacia do de obter de carga
                                          de gás
                                       dados no
da energia
                                     étri
                                       bruno arcuri
              dados referentes 500
                                     as fontes
  e estados
                                        mil mw
```

```
de transporte de contribuir de transporte de control de de control de de control de contr
```

de linhas a transmissão ser o labex invertido há no dos autos dia 20 comissão interna advogados da acordos de 22 de acordos de 22 de acordos de cooperação epe podem programa labexima programa programa labexima programa labexima programa labexima programa programa

```
#View(head(plot diretoria palavras))
library(wordcloud2)
plot_diretorias_tf_dif_bigram = DB %>%
  select(DESCRI_PEDIDO,DIRETORIA) %>%
  unnest_tokens(BIGRAM, DESCRI_PEDIDO, token = "ngrams", n = 2) %>%
  count(DIRETORIA, BIGRAM, sort = TRUE) %>%
  bind_tf_idf(BIGRAM, DIRETORIA, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(BIGRAM = factor(BIGRAM, levels = rev(unique(BIGRAM)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA,levels=c("DEA","DEE","DGC","DPG","OUTROS"))) %>%
  select(BIGRAM, tf_idf, DIRETORIA)
## DEE
#jpeg("XX_wordclou_tfidf_dir01_DEE.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_bigram %>%
  filter(DIRETORIA == "DEE") %>%
  top_n(150, tf_idf) %>%
  wordcloud2(shuffle = TRUE,
             color = "random-dark",
             shape = "circle")
## DGC
#jpeg("XX_wordclou_tfidf_dir01_DGC.jpeg")
```

```
set.seed(233115)
plot_diretorias_tf_dif_bigram %>%
  filter(DIRETORIA == "DGC") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## DEA
#jpeg("XX_wordclou_tfidf_dir01_DEA.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_bigram %>%
 filter(DIRETORIA == "DEA") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## DPG
#jpeg("XX_wordclou_tfidf_dir01_DPG.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_bigram %>%
 filter(DIRETORIA == "DPG") %>%
  top_n(150, tf_idf) %>%
 wordcloud2()
```

## Separando palavras de um bigram em "palavra1" e "palavra2" p/ remover stopwords

Considerando já a exclusão de casos onde houver stopwords consecultivos na "palavra1" e "palavra2", ou seja onde  $palavra1 = stopword \land palavra2 = stopword$ 

```
bigrams = DB %>%
  select(DESCRI PEDIDO, DIRETORIA) %>%
  unnest_tokens(BIGRAM, DESCRI_PEDIDO, token = "ngrams", n = 2) %>%
  count(DIRETORIA, BIGRAM, sort = TRUE)
separa_bigrams = bigrams %>%
  separate(BIGRAM, c("palavra1", "palavra2"), sep = " ")
junta_bigrams = separa_bigrams %>%
  unite(BIGRAM, palavra1, palavra2, sep = " ")
# levels(as.factor(junta_bigrams$BIGRAM == bigrams$BIGRAM))
                                                               # CHECK
## remove stopwords
bigrams2 = cbind(separa_bigrams,BIGRAM = junta_bigrams$BIGRAM) %>%
  filter(!palavra1 %in% mystopwords$palavra) %>%
  filter(!palavra2 %in% mystopwords$palavra) %>%
 filter(!palavra1 %in% "a") %>%
 filter(!palavra2 %in% "a") %>%
  filter(!palavra1 %in% "p") %>%
  filter(!palavra1 %in% "s") %>%
  filter(!palavra1 %in% "d") %>%
  filter(!palavra2 %in% "p") %>%
  filter(!palavra2 %in% "s") %>%
  filter(!palavra2 %in% "d") %>%
  filter(!palavra2 %in% "s.a") %>%
  filter(!str_detect(palavra1, "0")) %>%
  filter(!str_detect(palavra1, "1")) %>%
```

```
filter(!str_detect(palavra1, "2")) %>%
filter(!str_detect(palavra1, "3")) %>%
filter(!str_detect(palavra1, "4")) %>%
filter(!str_detect(palavra1, "5")) %>%
filter(!str_detect(palavra1, "6")) %>%
filter(!str_detect(palavra1, "7")) %>%
filter(!str_detect(palavra1, "8")) %>%
filter(!str detect(palavra1, "9")) %>%
filter(!str_detect(palavra2, "0")) %>%
filter(!str_detect(palavra2, "1")) %>%
filter(!str_detect(palavra2, "2")) %>%
filter(!str_detect(palavra2, "3")) %>%
filter(!str_detect(palavra2, "4")) %>%
filter(!str_detect(palavra2, "5")) %>%
filter(!str_detect(palavra2, "6")) %>%
filter(!str_detect(palavra2, "7")) %>%
filter(!str_detect(palavra2, "8")) %>%
filter(!str_detect(palavra2, "9"))
#count(DIRETORIA, BIGRAM)
```

#### Nuvem de palavras por diretoria - s/ steeming c/ remoção de stopwords - bigram

```
#View(head(plot_diretoria_palavras))
library(wordcloud2)
library(wordcloud)
plot_diretorias_tf_dif_bigram2 = bigrams2 %>%
  select(BIGRAM,n,DIRETORIA) %>%
  bind_tf_idf(BIGRAM, DIRETORIA, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(BIGRAM = factor(BIGRAM, levels = rev(unique(BIGRAM)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA,levels=c("DEA","DEE","DGC","DPG","OUTROS"))) %>%
  select(BIGRAM, tf_idf, DIRETORIA)
## DEE
\#jpeg("XX\_wordclou\_tfidf\_dir01\_DEE.jpeg")
nuvem1.2 =
plot_diretorias_tf_dif_bigram2 %>%
 filter(DIRETORIA == "DEE") %>%
  select(-DIRETORIA, word = BIGRAM, freq = tf_idf) %>%
  #top_n(150, freq) %>%
  as.data.frame()
set.seed(231321)
wordcloud(words = nuvem1.2$word, freq = nuvem1.2$freq, min.freq = 0.2,
          max.words=250, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(10, "Dark2"))
```

```
geração ceg
                                                                                                                   relatório epe projetos eólicos consultoria volga
                                                                recebê lo produção anual mercado livre fontes utv
            epe poderia
                       can'al universitário
                                                                                               geração média parques eólicos boletins mensais
                                                                                                                                                                       módulos fotovoltaicos aspx porém
                                              rev.o estudo USINAS termoelétricas análise técnico
                                                           longo prazo
                                                                                                                                                                                                                                                                                                                                                                                                   biomassa etc
                                                                                               garantia física go energia elétrica es energia elétrica el
                                                                                                          retorno att puc rio última revisão gt roraima
                                                                                     ebe
                                                                                                                                                                                                                                                                                                                                                                                          gwh gerados
                                                                eólica
                                                                                                                                                                                            garantia física o parque eólico anuário estatístico anuário estatístico anuário estatístico estatístic
                s energética epe custo marginal att pe grontes atende santo antônio fonte eó li tais dados documento e
                                           att
   alternativas relatório
                                                                                                                                                                                                                                                                                                                                                                                                σ gas natural
carvão gás
carvão gás
solar pv
                                                                                                                                                                                                                                                                                                                                                                                                                                                      delo matemático agentes participantes
                                                                                                                                                                                                                                                                                                                                                                                                                 epe região norte
                                                                                                                                                                                                                                                                                                                                                                ita
  puc minas
                      epe mdi
                                                                              ccpe ctet energia eólica
                                                                                                                                                                                                                                                                                                                                                                     Ø
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  básic
mauá
quantos gwh mwmed mensal decreto no Control of the 
                                                                                                                                                                                                                                                                                          estudo epe *
        igaporă iii operação comercial
                                                                                                                                                             att mônica tv puc peço informar
                         senhores peço
        joão daniel interligado nacional diariamente despachadas
             ufv eol fontes renováveisacre rondônia processo nº usina eólica
                                                                                                                                                      atendimento elétrico amazonas distribuidora codigo unico subgrupo iv
                                                   energia Itda dados utilizados região nordeste epe utilizando
            extremo sul atlântico verde
```

```
empresa possui gadministrativa empresa distrito federal termos aditivos previsão legal passagens aéreas beneficiados anexar
                                                                      branco localizado
                  empresa possui
                                                               suporte técnico
rio branco
mentidade possui
                                                          possui contratos supra
                                                                   pareceres notas
                                                                   iniciativa privada
                                                                               centro rio
beneficiados anexar
convenções coletivas
estatal direcionou
cargos beneficiados
                                                                               <u>ख</u>
                                                                                  ल df site
                                                                                            era
                                                     Zã
     cidade corporate
                                                      instru
concurso público
             exista listar
                                                    órgão poss
                       previstos
         sul
     site
                 ल
                                                               inteiro teor
ed parque 2
                                                    classe executiva
   scs quadra
                                                   compra caso
                             istar
           orma interna
consulta termo
       caso exista
estatal federal
                                                     modalidades entidades o
                       pisos
                                                    parque cidade
                 total gasto o constituição federal
                                        analista
                                                                              avenida rio
                                                   tais salários srs permita
                                                      tesouro nacional
                                                      abaixo citados
seguintes informações
   funcionários recebem
                                                           processo dl.epe
```

```
universidade federal industrial comercial industrial ocomercial in
```

```
att maurícioartur campos
energia elétrica arquivos shp
nota técnica arquivos shp
adicionalmente gentileza
meno acompanha los
poderia passar
poderia passar
custos fixos
soprade destinada soprade destinada empresa
dopgre gás natural soprado alguma empresa
agradeço willian
agradeço willian
algum erro existe alguma
existe alguma
artigos dados aguardo retorno
http www.epe.gov.br
```

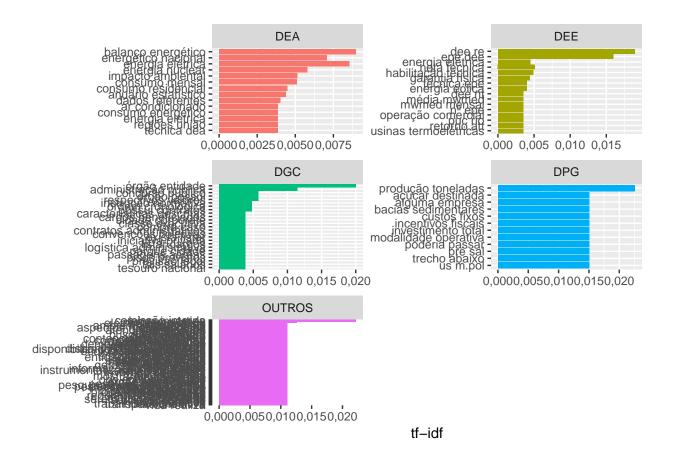
seto deferimento rio brasil busque algum risco ana paula algum atenção ana sere algum algu

```
#View(head(plot diretoria palauras))
library(wordcloud2)
plot_diretorias_tf_dif_bigram2 = bigrams2 %>%
  select(BIGRAM,n,DIRETORIA) %>%
  bind_tf_idf(BIGRAM, DIRETORIA, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(BIGRAM = factor(BIGRAM, levels = rev(unique(BIGRAM)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA,levels=c("DEA","DEE","DGC","DPG","OUTROS"))) %>%
  select(BIGRAM, tf_idf, DIRETORIA)
#jpeg("XX_wordclou_tfidf_dir01_DEE.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_bigram2 %>%
  filter(DIRETORIA == "DEE") %>%
  top n(150, tf idf) %>%
  wordcloud2(shuffle = TRUE,
             color = "random-dark",
             shape = "circle")
## DGC
#jpeg("XX_wordclou_tfidf_dir02_DGC.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_bigram2 %>%
```

```
filter(DIRETORIA == "DGC") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## DEA
#jpeg("XX_wordclou_tfidf_dir03_DEA.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_bigram2 %>%
  filter(DIRETORIA == "DEA") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## DPG
#jpeg("XX_wordclou_tfidf_dir04_DPG.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_bigram2 %>%
  filter(DIRETORIA == "DPG") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## OUTROS
#jpeg("XX_wordclou_tfidf_dir05_OUTROS.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_bigram2 %>%
  filter(DIRETORIA == "OUTROS") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
```

## Gráfico da estatística tf\_idf c/ remoção de stopwords

```
plot_diretorias_tf_dif_bigram2 %>%
group_by(DIRETORIA) %>%
top_n(10, tf_idf) %>%
ungroup() %>%
mutate(BIGRAM = reorder(BIGRAM, tf_idf)) %>%
ggplot(aes(BIGRAM, tf_idf, fill = DIRETORIA)) +
geom_col(show.legend = FALSE) +
labs(x = NULL, y = "tf-idf") +
facet_wrap(~DIRETORIA, ncol = 2, scales = "free") +
coord_flip() +
scale_y_continuous(labels=gcomma)
```



## Nuvem de palavras por diretoria - s/ steeming e/ou stopwords - trigram

```
#View(head(plot_diretoria_palavras))
library(wordcloud)
plot_diretorias_tf_dif_trigram = DB %>%
  select(DESCRI_PEDIDO,DIRETORIA) %>%
  unnest_tokens(TRIGRAM, DESCRI_PEDIDO, token = "ngrams", n = 3) %>%
  count(DIRETORIA, TRIGRAM, sort = TRUE) %>%
  bind_tf_idf(TRIGRAM, DIRETORIA, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(TRIGRAM = factor(TRIGRAM, levels = rev(unique(TRIGRAM)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA,levels=c("DEA","DEE","DGC","DPG","OUTROS"))) %>%
  select(TRIGRAM, tf_idf, DIRETORIA)
#jpeg("wordcloud_tfidf_dir01_DEE_trigram_comstop_semstemming.jpeg")
nuvem1.3 =
plot_diretorias_tf_dif_trigram %>%
  filter(DIRETORIA == "DEE") %>%
  select(-DIRETORIA, word = TRIGRAM, freq = tf_idf) %>%
  #top_n(150, freq) %>%
  as.data.frame()
set.seed(8835)
wordcloud(words = nuvem1.3$word, freq = nuvem1.3$freq, min.freq = 0.2,
          max.words=250, random.order=FALSE, rot.per=0.35,
```

```
o com 1.768.000 kw
for correto solicito
                       a memória de
elétrica
                                                os dados do de dados do

    energia de reserva

                                                                                                                          உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
உள்ளிய de reserva
கூறிய de re
                                                       leilão a 5
                                                              a 5 2014
                                                              da epe o to
                                                                                                                                                         fluxo de potência
de energia
                                                   a partir de
                                                                                                                                                                                                                                                                                                           ener
                                                                                                                                                  dos leilões de
                                                                                                                                       de energia de de cme e
                  Φ
                  (1)
  licito os
                                                                                                                                                                                                                                                                                                         ener
                  (1)
                                                                                                                                                                                                                                                                                                          Φ
                                                                                                                                                                                                                                                                                      0
                                                                                                                                                                                                                                                                            site da
                                                                                        do leilão de
                                                                                                                                                                                                                           partir
              se isto for
                                                                                                                                                                                                                                              Ø
                                                                                                                                                                                                                                             emand
   correto
                        garantia física de
                                                                                                                                                                                                                                                                                                         eraç
                                                                                                                                                                                  leilões
                                      partir de janeiro iĕ
                                                                                                                                                                                                                           ത
                                                                                                                                                                   Φ
               acesso ao sistema
                                                                                                                                                                                                                            dados
                                           da geração de
                                                                                                                                                                                     nos
                                                                                                                                                                                                                                                                                                               de
 a 4 de de entrada em
                                                                                                                                                                                                                                                                          2
                                                                                                                                                                                                              ō
                                                                                                                                                                                                                                                               nod
                                                                                                                                                                                                                                                                                                              2
      média mwmed mensa
de 2017 até
                                                                                                                                                        2004 cabe a
                                                                                                                                                                                                                               do sin e
```

## #dev.off()

de incentivo ao O caso exista listar o número de se for o nos termos do lote c ed em 2016 e de cargos de a razão da 09 lote c a compra de 2015 e 201 a sede desse a in 05 de acesso n 2004 a 2015 e a previsao e ou no folha ao inteiro teor e gestão de e ou repórter

```
regiões união ede todas as
entre os anos
           de 2000 a por classe de
de 1980 a 🥨
             de impacto ambiental
   ado
consumo mensa
                 da energia nuclear
                 banco de dados
               o período de
   Φ
      energia nuclear no
       mensal de energia
gos anos de dos dados de de energia eletrica
                                 nos o
    de energia eletrica
      nota técnica dea
 os dados de solicitar dados
```

```
## DPG
#jpeg("XX_wordclou_tfidf_dir04_DPG_trigram_comstop_semstemming.jpeg")
nuvem4.3 =
plot_diretorias_tf_dif_trigram %>%
   filter(DIRETORIA == "DPG") %>%
   select(-DIRETORIA, word = TRIGRAM,freq = tf_idf) %>%
   #top_n(150, freq) %>%
   as.data.frame()

set.seed(75437)
wordcloud(words = nuvem4.3$word, freq = nuvem4.3$freq, min.freq = 0.1,
        max.words=250, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(10, "Dark2"))
```

```
à produção de de etanol de em ago de mago de m
```

da constituição da altera a corrente a vinda de a bancos e SE O QUE SEO QUE Está Esta corrento e a síntese desse fazem a transmissão ancorados em corrente

```
#View(head(plot diretoria palauras))
library(wordcloud2)
plot_diretorias_tf_dif_trigram = DB %>%
  select(DESCRI PEDIDO,DIRETORIA) %>%
  unnest tokens (TRIGRAM, DESCRI PEDIDO, token = "ngrams", n = 3) %>%
  count(DIRETORIA, TRIGRAM, sort = TRUE) %>%
  bind_tf_idf(TRIGRAM, DIRETORIA, n) %>%
  arrange(desc(tf_idf)) %>%
  mutate(TRIGRAM = factor(TRIGRAM, levels = rev(unique(TRIGRAM)))) %>%
  mutate(DIRETORIA = factor(DIRETORIA,levels=c("DEA","DEE","DGC","DPG","OUTROS"))) %>%
  select(TRIGRAM, tf_idf, DIRETORIA)
## DEE
#jpeg("XX_wordclou_tfidf_dir01_DEE.jpeg")
set.seed(233115)
plot diretorias tf dif trigram %>%
  filter(DIRETORIA == "DEE") %>%
  top n(150, tf idf) %>%
  wordcloud2(shuffle = TRUE,
             color = "random-dark",
             shape = "circle")
## DGC
#jpeg("XX_wordclou_tfidf_dir02_DGC.jpeg")
```

```
set.seed(233115)
plot_diretorias_tf_dif_trigram %>%
  filter(DIRETORIA == "DGC") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## DEA
#jpeg("XX_wordclou_tfidf_dir03_DEA.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_trigram %>%
  filter(DIRETORIA == "DEA") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## DPG
\#jpeg("XX\_wordclou\_tfidf\_dir04\_DPG.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_trigram %>%
  filter(DIRETORIA == "DPG") %>%
  top_n(150, tf_idf) %>%
  wordcloud2()
## OUTROS
#jpeg("XX_wordclou_tfidf_dir05_OUTROS.jpeg")
set.seed(233115)
plot_diretorias_tf_dif_trigram %>%
  filter(DIRETORIA == "OUTROS") %>%
  top_n(150, tf_idf) %>%
 wordcloud2()
```

## **ANEXOS**

• Anexo 01: Tabela - Exemplo amostral da tabela unificada

```
DB[c(32,50,66),c(-1,-3,-5,-6,-9)] %>%
    select(DATA_PEDIDO, DATA_RESPOSTA, DIRETORIA, DESCRI_PEDIDO) %>%
kable("latex", caption = "Amostra dos dados a serem pré-processados", booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"), full_width = F) %>%
    column_spec(4:4, width = "2cm") %>%
    column_spec(5:5, width = "10cm") %>%
landscape()
```

Table 6: Amostra dos dados a serem pré-processados

|    | DATA_PEDIDO      | DATA_RESPOSTA | DIRETORIA | DESCRI_PEDIDO  |
|----|------------------|---------------|-----------|--|
| 32 | 30/08/2018 16:44 | 2018-08-31    | DEA       | Boa tarde, Gostaria de solicitar os dados históricos de Estatísticas do Consumo de Energia Elétrica (GWh), divulgados pela ONS na Resenha Mensal. A finalidade é estudo econométrico da série histórica de consumo de energia no Brasil e nos setores da economia. Obrigada. |
| 50 | 25/08/2015 18:35 | 2015-09-04    | DEE       | Prezados, boa tarde! Solicito a cópia da NT EPE-DEE-RE-077/2008. Estou realizando alguns estudos pertinentes a CUR e preciso deste arquivo de referência. Grato pela atenção! Thiago Paulino   |
| 66 | 21/10/2015 14:53 | 2015-10-26    | DEE       | Prezados, bom dia!<br>Sirvo-me do presente para solicitar o COP CEC - Suape II -<br>Leilão 01/2007.  |