Mineração de texto em pedidos de Lei de Acesso ä informação - LAI

Packages for this routine

Importação e preparação dos dados

Pedidos e-SIC

```
PATH = "C:/proj_eSIC_v10/textmining_pt/DATA/"
#PATH = "/Users/ewersonpimenta/Desktop/ESIC_TCC/Pedidos_LAI_EPE/BASE_DADOS/"
FILE = "relatorio_pedidos.ods"
db_raw = readODS::read.ods(file = paste0(PATH,FILE), sheet = 1); # dim(db_raw)
dbnames = db_raw[1,]; db_raw = db_raw[-1,];
colnames(db_raw) = c("ID", "DATA_PEDIDO", "DATA_PRAZOATEND", "DESCRI_PEDIDO",
                     "RESUMO_PEDIDO", "DATA_RESPOSTA")
#View(head(db_raw))
LAI = db_raw
```

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
```

Freq. de solicitações no e-SIC por Diretoria da EPE

```
LAI1 %>%
 count(DIRETORIA, sort = TRUE) %>%
  kable("latex", caption = "Frequência de solicitações e-SIC por Diretoria/EPE",
        booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

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

DIRETORIA	n
DEA	210
DEE	197
DGC	115
DPG	24
OUTROS	19
SIC	1

Respostas e-SIC - Reclassificação Diretorias

```
diretorias = levels(as.factor(LAI1$DIRETORIA))
LAI1 = LAI1 %>%
```

```r
  mutate(DIRETORIA = ifelse(DIRETORIA == diretorias[6],diretorias[5],DIRETORIA))
diretorias = levels(as.factor(LAI1$DIRETORIA))
LAI1 = LAI1 %>%
  mutate(DIR_NEW = ifelse(DIRETORIA == diretorias[1],diretorias[1],
                          ifelse(DIRETORIA == diretorias[2],diretorias[2], "OUTRAS")))
#dim(LAI1)
#View(head(LAI1))
```

```r
LAI1 %>%
 count(DIRETORIA, sort = TRUE) %>%
  kable("latex", caption = "Frequência de solicitações e-SIC por Diretoria/EPE - sem SIC",
        booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

Table 2: Frequência de solicitações e-SIC por Diretoria/EPE - sem SIC

DIRETORIA	n
DEA	210
DEE	197
DGC	115
DPG	24
OUTROS	20

```r
LAI1 %>%
 count(DIRETORIA, sort = TRUE) %>%
  kable("html", caption = "Frequência de solicitações e-SIC por Diretoria/EPE - sem SIC",
        booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped", "hold_position"))
```

```r
LAI1 %>%
 count(DIR_NEW, sort = TRUE) %>%
  kable("latex", caption = "Frequência de solicitações e-SIC por Diretoria/EPE (TOP2)",
        booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

Table 3: Frequência de solicitações e-SIC por Diretoria/EPE (TOP2)

DIR_NEW	n
DEA	210
DEE	197
OUTRAS	159

```r
LAI1 %>%
 count(DIR_NEW, sort = TRUE) %>%
  kable("html", caption = "Frequência de solicitações e-SIC por Diretoria/EPE (TOP2)",
        booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped", "hold_position"))
```

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))
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 4: Amostra dos dados a serem pré-processados

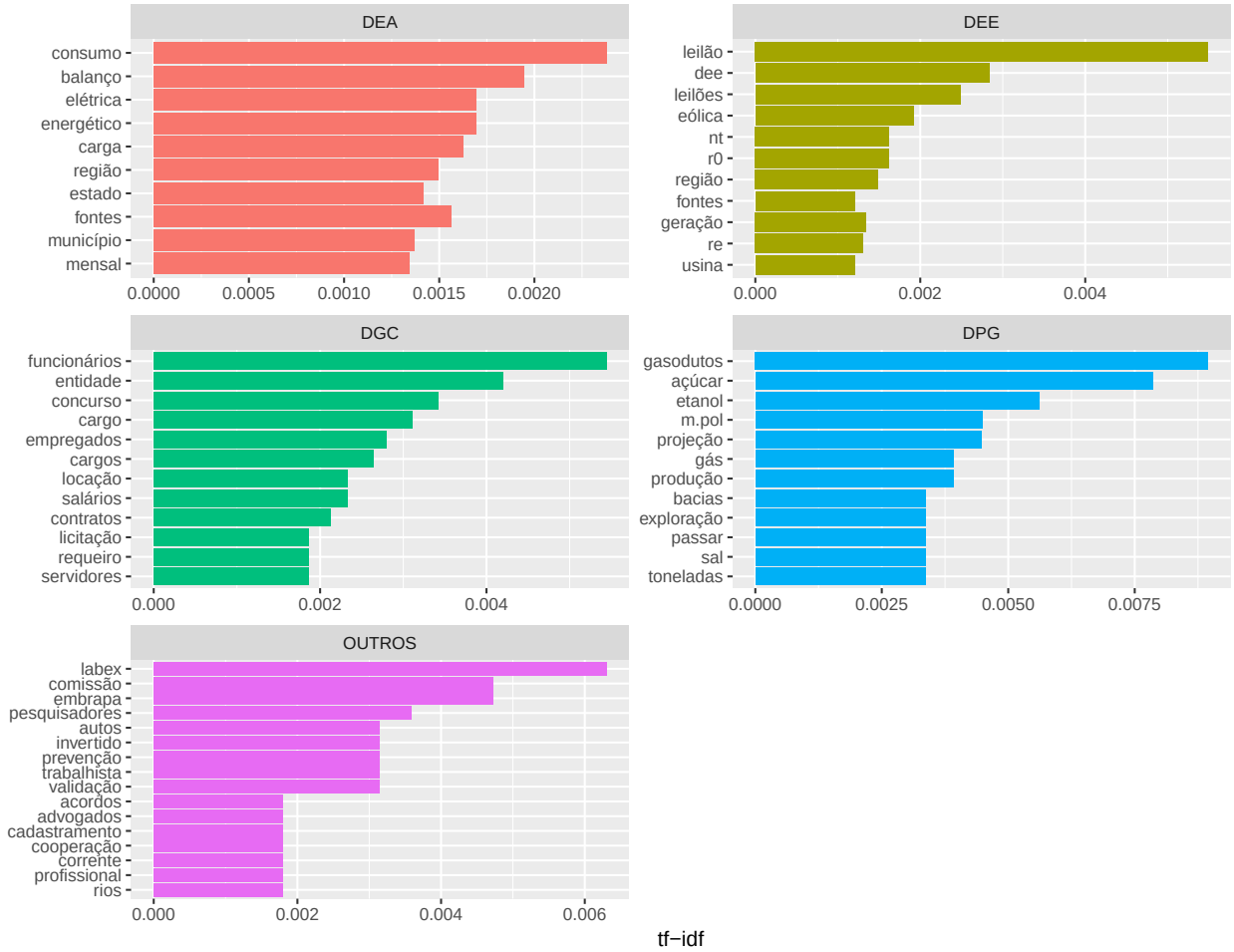
	DATA_PEDIDO	DATA_RESPOSTA	DIRETORIA	DESCRI_PEDIDO
32	30/08/2018 16:44	2018-08-31	DEA	Boa tarde, Gostaria de solicitar os dados históricos de Estatísticas do Consumo de Energia Elétrica (GWh), divulgados pela ONS na Resenha Mensal. A finalidade é estudo econométrico da série histórica de consumo de energia 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! Sirvo-me do presente para solicitar o COP CEC - Suape II - Leilão 01/2007.

Análise de texto

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()
```




```
#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

```
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("O nosso vetor de stopwords contém ",length(stopwords_pt), " palavras únicas"))
```

```
## O nosso vetor de stopwords contém 562 palavras únicas
```

```
## dim(stopwords_pt); class(stopwords_pt)
stopwords_pt = as.character(stopwords_pt)
stopwords_pt[1:14]
```

```
##  [1] "<U+FEFF>a" "à"       "acerca"  "adeus"   "agora"   "aí"      "ainda"
##  [8] "alem"    "além"    "algmas"  "algo"    "algumas" "alguns"  "ali"
```

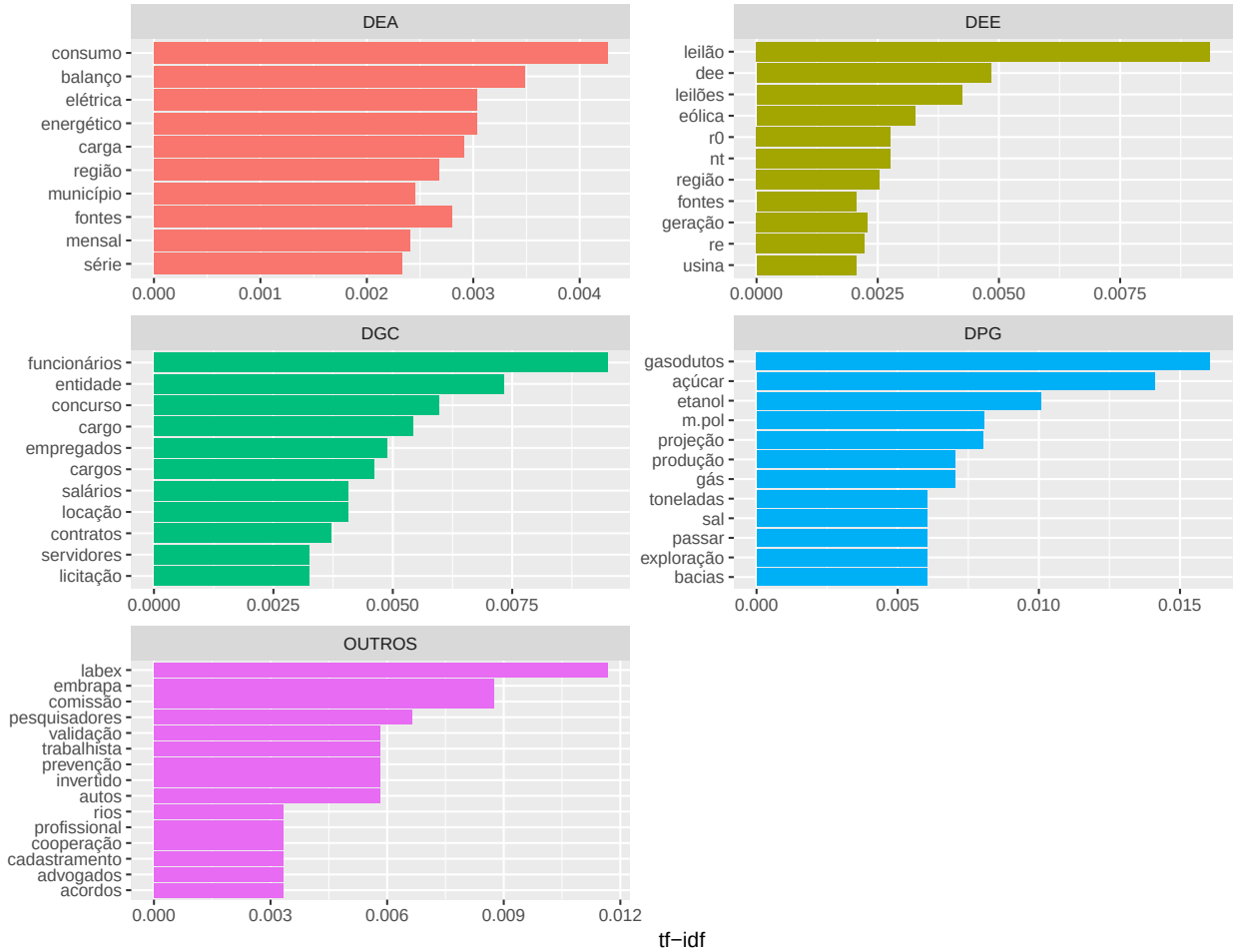
Freq. de palavras sem stopwords por diretoria

```
mystopwords <- data_frame(palavra = stopwords_pt)
```

```
## 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)
#jpeg("03_freq_palavras_dir_nostop.jpeg")
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()
```




```
#dev.off()
```

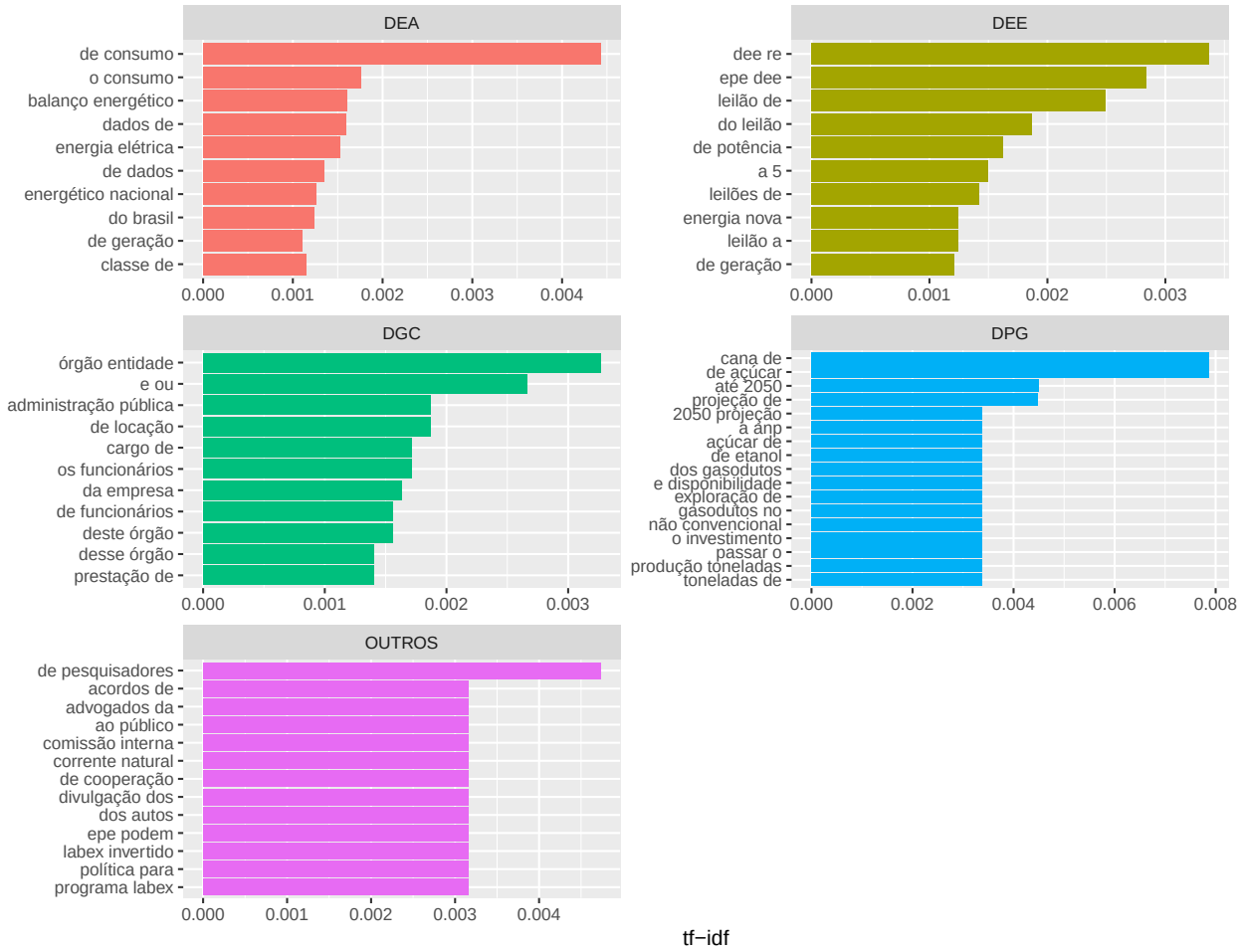
Usando bigram para n=2 palavras por token

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))
```

```r
#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()
```




```r
#dev.off()
```

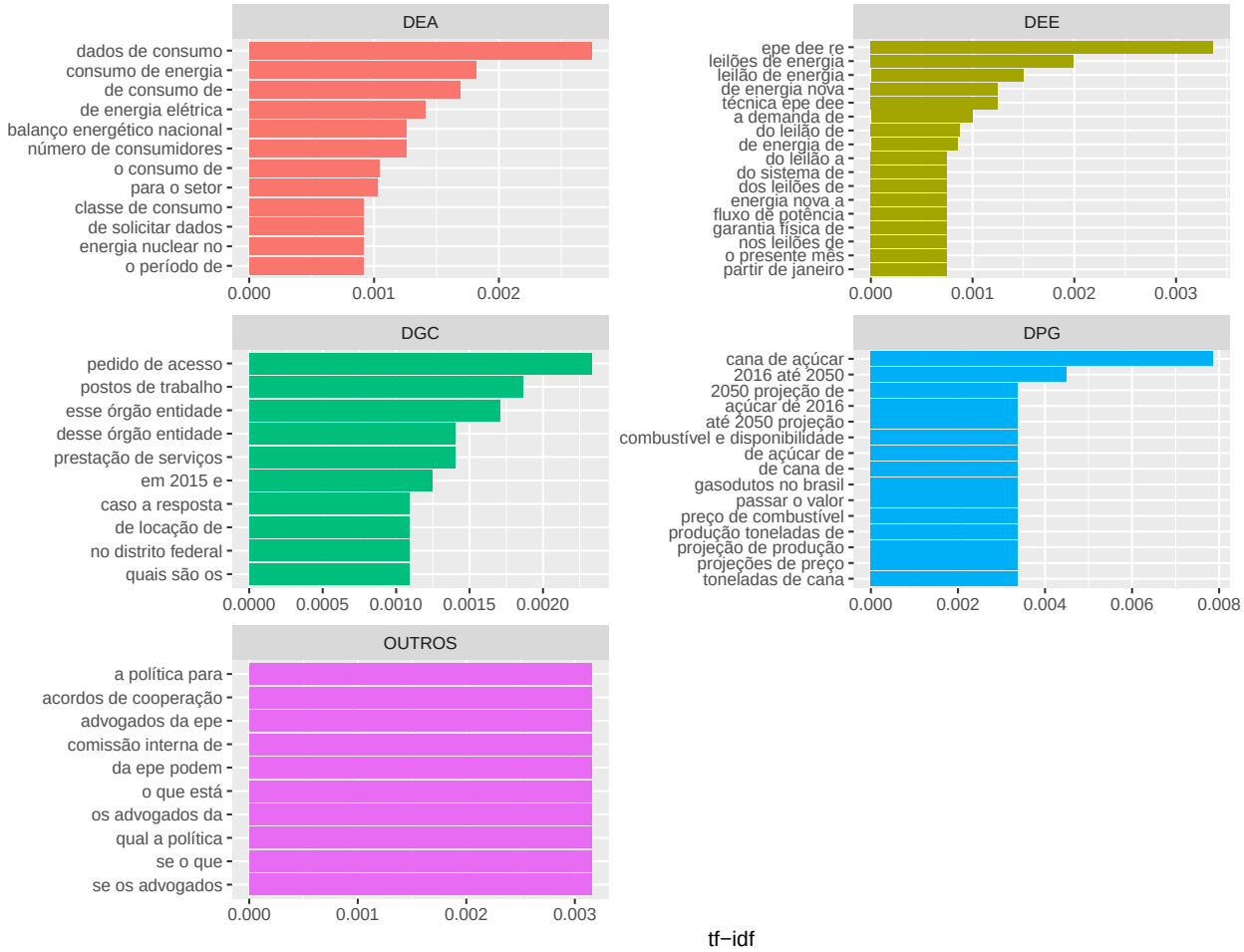
Usando bigram para n=3 palavras por token

Frequência de palavras por diretoria

```r
diretoria_palavras_trigram <- DB %>%
  select(DESCRI_PEDIDO,DIRETORIA) %>%
  unnest_tokens(TRIGRAM, DESCRI_PEDIDO, token = "ngrams", n = 3) %>%
  count(DIRETORIA, TRIGRAM, sort = TRUE) %>%
  ungroup()
```

```r
#diretoria_palavras_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()
```


DEA
dados de consumo
consumo de energia
de consumo de
de energia elétrica
balanço energético nacional
número de consumidores
o consumo de
para o setor
classe de consumo
de solicitar dados
energia nuclear no
o período de
0.000
0.001
0.002
DEE
epe dee re
leilões de energia
leilão de energia
de energia nova
técnica epe dee
a demanda de
do leilão de
de energia de
do leilão a
do sistema de
dos leilões de
energia nova a
fluxo de potência
garantia física de
nos leilões de
o presente mês
partir de janeiro
0.000
0.001
0.002
0.003
DGC
pedido de acesso
postos de trabalho
esse órgão entidade
desse órgão entidade
prestação de serviços
em 2015 e
caso a resposta
de locação de
no distrito federal
quais são os
0.0000
0.0005
0.0010
0.0015
0.0020
DPG
cana de açúcar
2016 até 2050
2050 projeção de
açúcar de 2016
até 2050 projeção
combustível e disponibilidade
de açúcar de
de cana de
gasodutos no brasil
passar o valor
preço de combustível
produção toneladas de
projeção de produção
projeções de preço
toneladas de cana
0.000
0.002
0.004
0.006
0.008
OUTROS
a política para
acordos de cooperação
advogados da epe
comissão interna de
da epe podem
o que está
os advogados da
qual a política
se o que
se os advogados
0.000
0.001
0.002
0.003
tf−idf

```
#dev.off()
```