Matriz de documentos e termos

Prof. Walmes Zeviani

walmes@ufpr.br

Laboratório de Estatística e Geoinformação
Departamento de Estatística
Universidade Federal do Paraná

Motivação

- A representação espaço vetor é utilizada por muitas técnicas em mineração de texto.
- As coordenadas dos documentos dependem das métricas usadas para construção da matriz de documentos e termos.
- É preciso conhecer métricas mais adotadas e suas propriedades.
- A construção da matriz de documentos e termos deve considerar aspectos do problema em mãos.

Ponderação da MDT

Ponderação da matriz

- Expressa a ocorrência dos termos dentro de um documento.
- Funções da quantidade de vezes que o termo aparece:
  - Linear:
$$n(t) = \text{count}(t, d), \quad n \in \mathbb{N}.$$
  - Indicadora ou binária:
$$b(t) = I(n(t) > 0), \quad b \in \{0, 1\}.$$
  - Logarítmica:
$$l(t) = \log(n(t) + 1), \quad \text{ł} \in \mathbb{R}.$$
  - Outras funções sublineares.
- Funções sublinares diminuem o efeito das palavras mais frequentes.

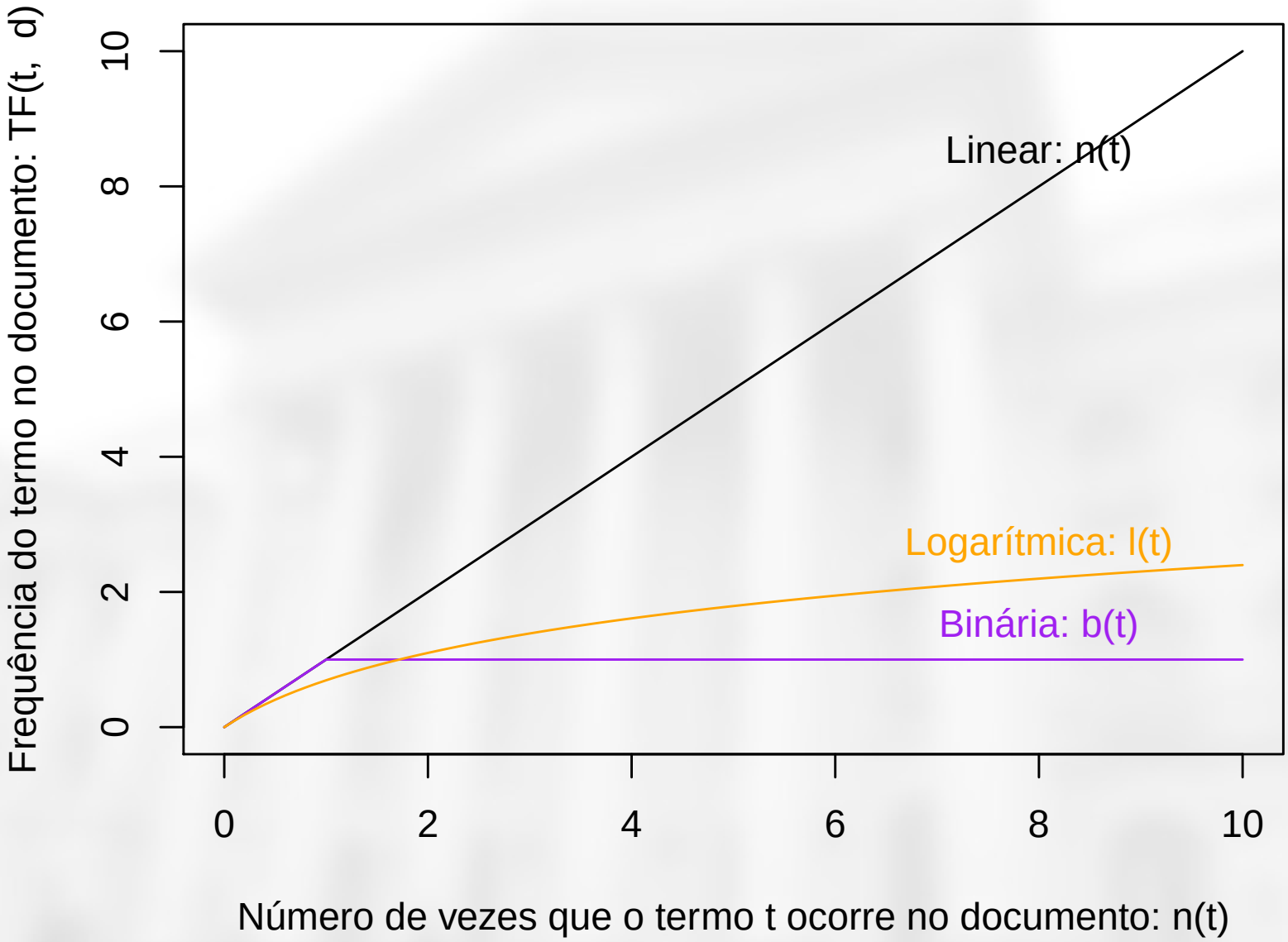

Figura 1. Funções de ponderação conforme a ocorrência de um termo no documento.

Poderar mais os termos raros

- Uma palavra que ocorre em todos os documento diz muito pouco sobre cada um deles.
  - É uma característica que todos os documentos tem, então:
  - Não contribui para classificação/agrupamento dos documentos.
  - Não contribui para como uma variável regressora.
- Usa-se dar mais peso para palavras no corpus que ocorrem em poucos documentos.

A ponderação TF-IDF

- TF-IDF: *term frequency inverse document frequency*.
- A porção IDF é determinada por

$$\text{IDF}(t) = \log\left(\frac{\text{count}(d, C) + 1}{\text{count}(d : t \in d, C)}\right), \quad 0 \leq \text{IDF} \leq \log(\text{count}(d, C)+1).$$

- Em algumas referências pode não aparecer o $+1$ no numerador.
- Numerador: total de documentos.
- Denominador: documentos que possuem o termo t.
- A razão é sempre positiva.

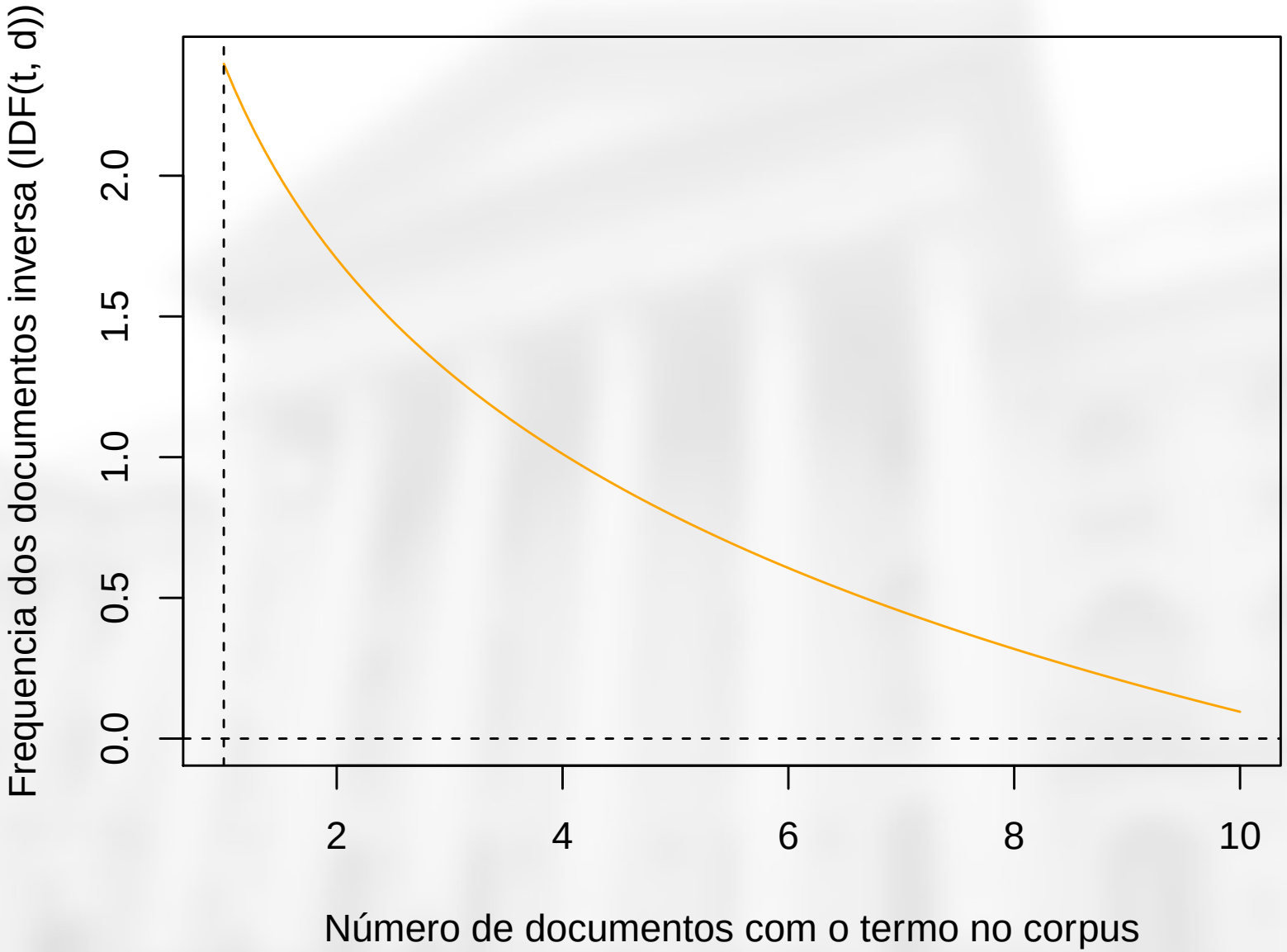

Figura 2. Ponderação IDF considerando um corpus com 10 documentos.

Ponderação de cada termo por TF-IDF

- Combina a frequência no termo (TF) no documento com a ocorrência dele na coleção (IDF).
- O peso de um termo no documento é

$$w(t, d) = \text{TF}(t, d) \times \text{IDF}(t).$$

- Considerando explicitamente todos os termos, tem-se

$$w(t, d) = \text{count}(t, d) \times \log \left(\frac{\text{count}(d, C) + 1}{\text{count}(d : t \in d, C)} \right).$$

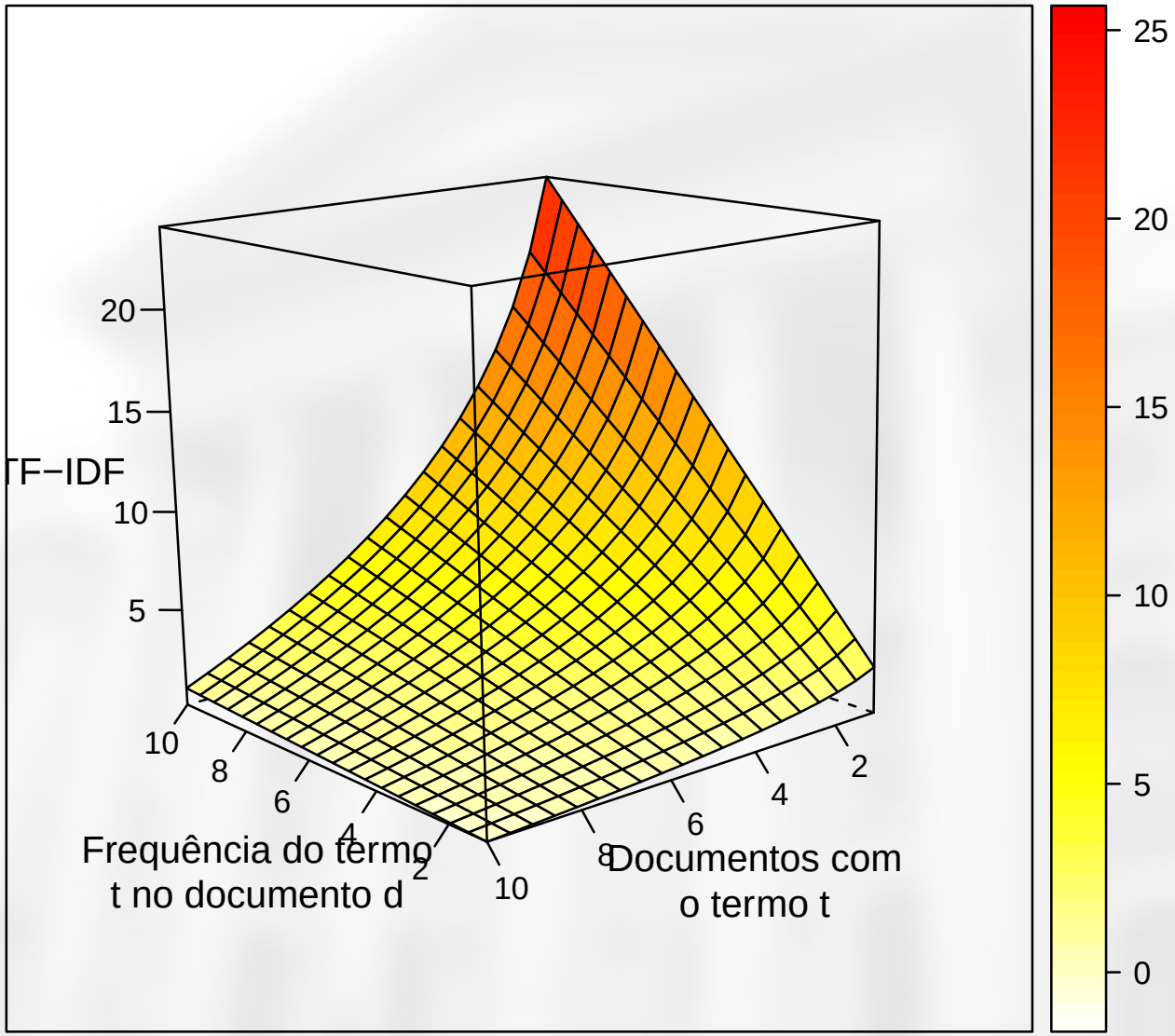

Figura 3. Valor de TF-IDF para uma coleção de $n = 10$ documentos.

Aplicação com o R

Um corpus didático

```
library(tm)

docs <- c("A vida é linda.",
          "A vida é uma aventura.",
          "A vida é uma só.",
          "A vida é linda por ser única.",
          "A vida é minha, minha vida!",
          "A vida é pra ser vivida.")
cps <- VCorpus(VectorSource(x = docs),
               readerControl = list(language = "pt", load = TRUE))
cps

cps <- tm_map(cps, FUN = content_transformer(tolower))
cps <- tm_map(cps, FUN = removePunctuation)
cps <- tm_map(cps, FUN = removeWords,
              words = c("a", "é", "e", "uma", "pra", "por", "ser"))
content(cps[[1]])

# Funções de ponderação.
apropos("^weight[[:upper:]]", ignore.case = FALSE)
```

Ponderação TF

```r
dtm_tf <- DocumentTermMatrix(cps,
                             control = list(weighting = weightTf,
                                            wordLengths = c(1, Inf)))
inspect(dtm_tf)
```

```
## <<DocumentTermMatrix (documents: 6, terms: 7)>>
## Non-/sparse entries: 13/29
## Sparsity           : 69%
## Maximal term length: 8
## Weighting          : term frequency (tf)
## Sample             :
##     Terms
## Docs aventura linda minha só única vida vivida
##    1        0     1     0  0     0    1      0
##    2        1     0     0  0     0    1      0
##    3        0     0     0  1     0    1      0
##    4        0     1     0  0     1    1      0
##    5        0     0     2  0     0    2      0
##    6        0     0     0  0     0    1      1
```

Ponderação binária

```
dtm_bin <- 1 * as.matrix(dtm_tf > 0)
dtm_bin
```

```
##     Terms
## Docs aventura linda minha só única vida vivida
##    1        0     1     0  0     0    1      0
##    2        1     0     0  0     0    1      0
##    3        0     0     0  1     0    1      0
##    4        0     1     0  0     1    1      0
##    5        0     0     1  0     0    1      0
##    6        0     0     0  0     0    1      1
```

```
dtm_bin <- DocumentTermMatrix(cps,
                              control = list(weighting = weightBin,
                                             wordLengths = c(1, Inf)))
# inspect(dtm_bim)
```

Ponderação TF-IDF

```r
dtm_tf <- as.matrix(dtm_tf)
dtm_bin <- 1 * (dtm_tf > 0)

# ATTENTION: é log base 2 que é usada. Não tem o `+1`.
(idf <- log2((nrow(dtm_bin))/colSums(dtm_bin)))
```

```
## aventura    linda    minha      só    única     vida   vivida
## 2.584963 1.584963 2.584963 2.584963 2.584963 0.000000 2.584963
```

```r
# Divide cada coluna pelo respectivo escalar.
sweep(dtm_tf, MARGIN = 2, STATS = idf, FUN = "*")
```

```
##     Terms
## Docs aventura    linda    minha      só    única vida   vivida
##    1 0.000000 1.584963 0.000000 0.000000 0.000000    0 0.000000
##    2 2.584963 0.000000 0.000000 0.000000 0.000000    0 0.000000
##    3 0.000000 0.000000 0.000000 2.584963 0.000000    0 0.000000
##    4 0.000000 1.584963 0.000000 0.000000 2.584963    0 0.000000
##    5 0.000000 0.000000 5.169925 0.000000 0.000000    0 0.000000
##    6 0.000000 0.000000 0.000000 0.000000 0.000000    0 2.584963
```

Ponderação TF-IDF

```r
# https://www.quora.com/What-are-non-normalized-TF-IDF-weights
weightTfIdf_un <- function(x) weightTfIdf(x, normalize = FALSE)

dtm_tfidf <- DocumentTermMatrix(cps,
                                control = list(
                                    weighting = weightTfIdf_un,
                                    wordLengths = c(1, Inf)))
inspect(dtm_tfidf)
```

```
## <<DocumentTermMatrix (documents: 6, terms: 7)>>
## Non-/sparse entries: 7/35
## Sparsity           : 83%
## Maximal term length: 8
## Weighting          : term frequency - inverse document frequency (tf-idf)
## Sample             :
##     Terms
## Docs aventura    linda    minha       só    única vida   vivida
##    1 0.000000 1.584963 0.000000 0.000000 0.000000    0 0.000000
##    2 2.584963 0.000000 0.000000 0.000000 0.000000    0 0.000000
##    3 0.000000 0.000000 0.000000 2.584963 0.000000    0 0.000000
##    4 0.000000 1.584963 0.000000 0.000000 2.584963    0 0.000000
##    5 0.000000 0.000000 5.169925 0.000000 0.000000    0 0.000000
##    6 0.000000 0.000000 0.000000 0.000000 0.000000    0 2.584963
```

Considerações sobre o TF-IDF

- Funciona como uma padronização multiplicativa por coluna.
- As padronizações *Z* ou min-max (são aditivas-multiplicativas).
- Termos raros serão valorizados e termos frequentes penalizados.
- O argumento `normalize = TRUE` faz uma normalização por linhas, ou seja, considera a proporção do documento com cada termo.

TF-IDF com normalização por linha

```r
dtm_tf <- as.matrix(dtm_tf)
dtm_bin <- 1 * (dtm_tf > 0)

# Normalização por linha (proporção de cada termo no documento).
dlen <- rowSums(dtm_tf)
dtm_tfn <- sweep(dtm_tf, MARGIN = 1, STATS = dlen, FUN = "/")

# Normalização por coluna com IDF.
idf <- log2((nrow(dtm_bin))/colSums(dtm_bin))
sweep(dtm_tfn, MARGIN = 2, STATS = idf, FUN = "*")
```

```
##     Terms
## Docs aventura     linda    minha       só     única vida   vivida
##    1 0.000000 0.7924813 0.000000 0.000000 0.0000000    0 0.000000
##    2 1.292481 0.0000000 0.000000 0.000000 0.0000000    0 0.000000
##    3 0.000000 0.0000000 0.000000 1.292481 0.0000000    0 0.000000
##    4 0.000000 0.5283208 0.000000 0.000000 0.8616542    0 0.000000
##    5 0.000000 0.0000000 1.292481 0.000000 0.0000000    0 0.000000
##    6 0.000000 0.0000000 0.000000 0.000000 0.0000000    0 1.292481
```

```r
dtm_tfidf <- DocumentTermMatrix(cps,
                                control = list(
                                    weighting = weightTfIdf,
                                    wordLengths = c(1, Inf)))
```

Ponderações definidas pelo usuário

- É possível usar outras funções de ponderação.
- Verifique este link: https://stackoverflow.com/questions/39448360/how-do-i-set-up-tf-weight-of-terms-in-corpus-using-the-tm-package-in-r.
- A escolha da ponderação é problema dependente.
- Quando não houver clara preferência por uma ponderação, aplique as disponíveis e avalie os resultados.

Próxima aula

- Análise de sentimentos.
- Introdução ao `tidytext`.
- Sabatina a partir de quinta.