```
Análise textual com LDA (Latent Dirichlet Allocation)
Este projeto utiliza o modelo de Latent Dirichlet Allocation (LDA) para análise de tópicos em um conjunto de textos. O objetivo é identificar os
principais tópicos presentes nos documentos e visualizar os termos mais representativos de cada tópico.
  library(data.table)
 library(tm)
  ## Carregando pacotes exigidos: NLP
  library(tidyverse)
  ## — Attaching core tidyverse packages -
                                                                   tidyverse 2.0.0 —
  ## √ dplyr
                 1.1.4
                           ✓ readr
                                        2.1.4
  ## √ forcats 1.0.0

√ stringr

                                       1.5.0
  ## √ ggplot2 3.5.1
                           √ tibble
                                       3.2.1
  ## ✓ lubridate 1.9.2
                           √ tidyr
                                       1.3.1
  ## √ purrr
                 1.0.1
  ## — Conflicts —
                                                             tidyverse_conflicts() —
  ## X ggplot2::annotate() masks NLP::annotate()
                            masks data.table::between()
  ## X dplyr::between()
                            masks stats::filter()
  ## X dplyr::filter()
  ## X dplyr::first()
                            masks data.table::first()
  ## X lubridate::hour()
                            masks data.table::hour()
  ## X lubridate::isoweek() masks data.table::isoweek()
  ## X dplyr::lag()
                            masks stats::lag()
  ## X dplyr::last()
                            masks data.table::last()
  ## X lubridate::mday()
                            masks data.table::mday()
  ## X lubridate::minute() masks data.table::minute()
  ## X lubridate::month()
                            masks data.table::month()
  ## X lubridate::quarter() masks data.table::quarter()
  ## X lubridate::second() masks data.table::second()
  ## X purrr::transpose()
                            masks data.table::transpose()
  ## X lubridate::wday()
                            masks data.table::wday()
  ## X lubridate::week()
                            masks data.table::week()
  ## X lubridate::yday()
                            masks data.table::yday()
  ## X lubridate::year()
                            masks data.table::year()
  ## i Use the conflicted package (<a href="http://conflicted.r-lib.org/">http://conflicted.r-lib.org/</a>) to force all conflicts to become errors
 library(dplyr)
 library(stopwords)
  ## Attaching package: 'stopwords'
  ##
  ## The following object is masked from 'package:tm':
  ##
  ##
        stopwords
 library(topicmodels)
 library(broom)
 library(writexl)
 library(tidytext)
 library(ggpubr)
 library(readxl) # Para ler o arquivo Excel
Visualizando os dados extraídos do site de notícias https://olhardigital.com.br
  dados <- read_excel("C:/Users/User/Downloads/dados_combinados.xlsx")</pre>
 head(dados)
  ## # A tibble: 6 × 4
      url
                                                               titulo descricao texto
      <chr>>
                                                               <chr> <chr>
 ## 1 https://olhardigital.com.br/2024/12/28/ciencia-e-espac... Explo... Explosõe... "Na ...
  ## 2 https://olhardigital.com.br/2024/12/28/internet-e-rede... Mega ... Quem ace... "A \...
  ## 3 https://olhardigital.com.br/2024/12/28/ciencia-e-espac... Explo... Explosõe... "Na ...
 ## 4 https://olhardigital.com.br/2024/12/26/reviews/ofertas... Ofert... Troque s... "Pro...
 ## 5 https://olhardigital.com.br/2024/12/26/ciencia-e-espac... Sonda... A sonda ... "O N...
 ## 6 https://olhardigital.com.br/2024/12/26/carros-e-tecnol... Produ... Montador... "A T...
Preparação dos dados, incluindo a remoção de palavras irrelevantes (stopwords) e a criação de uma matriz de termos de documentos (DTM).
  corp <- VCorpus(VectorSource(dados$texto))</pre>
  stopword_pt <- stopwords("pt")</pre>
  corpus_base <- tm_map(corp, removeWords, stopword_pt)</pre>
  dtm <- DocumentTermMatrix(corpus_base,</pre>
                            control = list(tolower = TRUE,
                                           removeNumbers = TRUE,
                                          removePunctuation = TRUE,
                                           stripWhitespace = TRUE,
                                           bounds = list(global = c(15, Inf))))
 # Converter a DTM
  dtm.matrix <- as.matrix(dtm)</pre>
 head(dtm.matrix,2)
  ## Docs abaixo ação acesso acima ações acontece acordo agência agora água ainda
                                       0
                                                                         0
                                                                               1
        Terms
  ## Docs ajudam ajudar além alerta algo algumas alguns alta amazon ambiente análise
                                      1
        Terms
  ## Docs and ano anos antes apenas apesar aplicativo aplicativos após aposta
                          1
       2 0 2
                                  4
        Terms
  ## Docs apresentou aqui área áreas artificial artigo assim assistir até atenção
                                                                    0 1
        Terms
  ## Docs atual atualmente aumentar aumento avanço base bastante bem bilhões bom
      Terms
  ## Docs brasil brasília busca cada caixa capacidade capaz casa caso casos causa
                                        0
                                        1
                                                          0 1 1
        Terms
  ## Docs causar celular cenário central centro cerca chamado chance china ciência
       2
              0
                                              0
                                                                   1
  ## Docs cientistas cinco clique com combinação como comum comunicação comunicado
                                   0
                               0 1
                                                  1
  ## Docs comuns condições confira conforme conheça conhecida conhecido conta
                                                           0
  ## Docs contato contém conteúdo continua contra controle corpo crédito crescimento
               0
  ## Docs criação criar crucial dados data dentro depois desafios descoberta
                0
                      0
                                        0
  ## Docs descobertas desde desenvolvimento desse destaca deste detalhes deve devem
        Terms
  ## Docs devido dezembro dia dias dicas diferença diferentes digital direção
                           0
        Terms
  ## Docs diretamente disney disponíveis disse disso diversas diversos divulgação
        Terms
  ## Docs diz dois duas durante efeitos eficiente ele eles elétricos embora empresa
  ## Docs empresas energia enquanto entanto então entenda entender entre época
                0
                                  0
                                                        1
        Terms
  ## Docs equipe escolha espacial espaço especial especialistas especialmente essa
                                       0
                                                3
        Terms
  ## Docs essas esse essencial esses estados estar este estudo estudos eua evitar
        Terms
  ## Docs exemplo existe existem expectativa experiência explica família fase
                0
                              0
  ## Docs fatores faz fazer federal feito fenômeno ferramentas fez ficar filmes fim
                0
        Terms
  ## Docs final forma foto funciona futuro garantir geral geralmente google governo
                                 1
        Terms
  ## Docs grande grandes grupo havia história hoje hora horário horas humanos ideal
        Terms
  ## Docs identificar imagem impacto importante incluindo indústria informações
                                                        1
       2
                           2
                                              0
  ## Docs início inteligência internet isso janeiro jogar jogo jogos lançamento leia
                                                                                  1
  ## Docs links locais local logo loja longa longo luz maior maiores maioria mais
        Terms
  ## Docs maneira marca mas matéria média medida meio melhor melhorar melhores menos
        Terms
  ## Docs mercado mês mesma mil milhões modelo modelos momento mostra mostram muda
                            0
                0 0
                         1 0
                                     1
        Terms
  ## Docs mudanças muitas muitos mundial mundo nacional não natal natural necessário
        Terms
  ## Docs necessidade nenhuma nesse nesta neste níveis notícias nova novas novembro
  ## Docs novo novos número objetivo ocorre oferece ofertas olhar onde online opções
  ## Docs operações outra outras outro outros país países papel para parece parte
                                                                  1
  ## Docs partes partir passado paulo período permite permitindo perto pesquisa
                      0
              1
                     1
                                            0
                                                                     0
  ## Docs pesquisadores pessoas planeta pode podem poder pois ponto por porém porque
                                        1
                                                 0
                                                                      0
  ## Docs possibilidade possíveis possível possuem possui potencial pouco precisa
                     1
                                        0
                     1
                                        1
  ## Docs preço presentes primeira primeiro principais principal principalmente
                                 0
  ## Docs problema problemas processo produto produtos programa promete própria
       Terms
  ## Docs proteger próximo próximos publicado quais qual qualidade qualquer quando
                                 0
                                                  1
        Terms
  ## Docs quanto quase quatro quer questão rápido real realidade realizado realizar
                           2
        Terms
  ## Docs receber recursos reduzir região regiões relação responsável resposta
                                       1
               0
                                 0
                                        0
                                                0
                                                        0
                                                                            1
  ## Docs resultado resultados revela revista risco riscos saber são saúde segue
      2
                                                               0 0
                                                                               2
        Terms
  ## Docs seguir segunda segundo segurança seis sempre sendo ser série serviços
                              1
                                            3
        Terms
  ## Docs setor shutterstock sido significa simples sistema sistemas site sobre
                          1 0
                          2 0
       Terms
  ## Docs sociais sol soluções sucesso sugere sul tamanho também tanto tão
               0 3
               0 0
                            0
                                               0
        Terms
  ## Docs tecnologia tela têm temperatura tempo ter terra testes the tipo toda todas
                  0 0 0
                                                 0
                  0 0 1
                                       0
                                              0 0
                                                               0 0
        Terms
  ## Docs todo todos tornar tornou total trabalho tratamento três tudo últimos uma
       2 0
                                 0
                                       0
                                                0
                                                           0 5
  ## Docs única unidos universidade universo usando usar uso usuário usuários
  ## Docs utilizam utilizando vai vale valor vários veículos veja vem ver verdade
                           0
                           0 1
  ## Docs vez vezes via vida vista você whatsapp
                 0 0
                        1
                                 0
modelo LDA com 15 tópicos
  modelo_Ida <- LDA(dtm, k = 15, method = "Gibbs", control = list(seed = 1234))</pre>
  ### probabilidades dos tópicos
  beta_topics <- tidy(modelo_Ida, matrix = "beta")</pre>
  beta_topics
  ## # A tibble: 6,975 × 3
       topic term
       <int> <chr>
                         <dbl>
  ## 1
           1 abaixo 0.0000775
  ##
     2
           2 abaixo 0.0000675
  ## 3
           3 abaixo 0.0000784
           4 abaixo 0.0000756
  ## 5
           5 abaixo 0.0000702
  ## 6
           6 abaixo 0.000657
  ## 7
           7 abaixo 0.0000665
           8 abaixo 0.0000740
  ## 8
  ## 9
           9 abaixo 0.0000671
          10 abaixo 0.0190
  ## 10
  ## # i 6,965 more rows
Os 10 principais termos por tópico:
  beta_top_terms <- beta_topics %>%
   group_by(topic) %>%
   slice_max(beta, n = 10) %>%
   ungroup() %>%
   arrange(topic, -beta)
  beta_top_terms
  ## # A tibble: 155 × 3
       topic term
                           beta
        <int> <chr>
                           <dbl>
  ## 1
                         0.0427
           1 sistema
  ## 2
           1 alta
                         0.0334
  ## 3
                         0.0319
           1 leia
  ## 4
           1 jogo
                         0.0303
  ## 5
                         0.0295
           1 com
                         0.0288
  ## 6
           1 the
  ## 7
                         0.0264
           1 além
  ## 8
                         0.0264
           1 onde
  ## 9
           1 experiência 0.0249
  ## 10
           1 imagem
                         0.0241
```

ggplot(aes(beta, term, fill = factor(topic))) +
geom_col(show.legend = FALSE) +
facet_wrap(~ topic, scales = "free") +
scale_y_reordered() +
labs(x = "Probabilidade (Beta)", y = "Termos", title = "Termos por Tópicos") +
theme_minimal()

print(grupos)

Termos por Tópicos

imagen a graph a

```
Termos
                   0.000.020.040.060.08
                                                                                                                0.000.010.020.030.04
                                                 0.000.020.040.06
                                                                                0.00.00.02.03.04.05
                                                          10
                                                                                         11
                                                                                                                         12
                   0.00\,0.02\,0.04\,0.06
                                                 0.000.030.060.090.12
                                                                                0.0\mathbf{0}.0\mathbf{0}.0\mathbf{2}.0\mathbf{3}.0\mathbf{4}.05
                                                                                                                0.000.030.060.09
                            13
                                                          14
                   0.00 0.02 0.04
                                                0.0000.0205.0500.0705.100
                                                                               0.000.025.050.075
                                                               Probabilidade (Beta)
As distribuições dos tópicos para cada documento:
  documento_topico <- tidy(modelo_Ida, matrix = "gamma")</pre>
  documento_topico
```

```
2 2
                  1 0.0523
## 3 3
                  1 0.0319
## 4 4
                  1 0.0678
## 5 5
                  1 0.0754
## 6 6
                  1 0.0290
## 7 7
                  1 0.0864
## 8 8
                  1 0.0287
## 9 9
                  1 0.0543
## 10 10
                  1 0.0519
## # i 2,990 more rows
#distribuições de tópicos por documento
matriz_documento_topico <- documento_topico %>%
 pivot_wider(names_from = topic, values_from = gamma)
head(matriz_documento_topico)
```

A tibble: 6 × 16

`1`

`2`

`3`

document

i 145 more rows

grupos <- beta_top_terms %>%

mutate(term = reorder_within(term, beta, topic)) %>%

<chr>> <dbl> < ## 1 1 0.0319 0.113 0.0613 0.113 0.0466 0.105 0.0319 0.127 0.0245 0.0539 0.0523 0.300 0.102 0.0523 0.0275 0.0358 0.0275 0.0689 0.0275 0.0358 ## 2 2 ## 3 3 0.0319 0.0613 0.0319 0.105 0.0245 0.0833 0.0613 0.127 0.0245 0.0613 0.0678 0.0271 0.0352 0.0352 0.0271 0.0271 0.0271 0.0271 0.0840 0.0352 ## 4 4 0.0754 0.0754 0.105 0.0389 0.0608 0.0462 0.0462 0.200 0.0608 0.0389 ## 5 5 ## 6 6 0.0290 0.0464 0.0638 0.0812 0.0377 0.0638 0.220 0.0638 0.0551 0.0464 ## # i 5 more variables: `11` <dbl>, `12` <dbl>, `13` <dbl>, `14` <dbl>, ## # `15` <dbl>