Acessando Dados da Web em R

Acessando e Analizando Dados do Twitter

Tiago Ventura | venturat@umd.edu

University of Maryland, College Park

Acessando a API do Twitter.

1. O Twitter é uma das companhias em que dados são mais facilmente acessados.
2. Resultado: A maior parte das análises acadêmicas sobre redes sociais usam dados do Twitter.
3. Exemplos de estudos aplicados com dados de Twitter:
   - Ativação em Redes e Tempo de Retuite
   - Posicionamento Ideológico de usuários
   - Polarização e Uso de bots
   - Assédio on-line e Racismo
   - Efeitos de Celebridades em Diminuição de Agressividade nas Redes

Introdução às APIs do Twitter

1. Há duas APIs disponibilizadas pelo Twitter.
    - RESTful API: Buscas de sete dias anteriores. (Search)
    - Streaming: Buscas de mensagens instantaneamente (Filter)

Ambos possuem rate-limits, e retornam somente uma porcentagem do total de tuítes.

Credenciais

```
app_name <- "seu_app_name"
consumer_key <- "seu_consumer_key"
consumer_secret <- "seu_consumer_secret"
access_token <- "seu_acess_token"
access_token_secret <- "seu_token_secret"
```

rtweet

- Para acessar a API, iremos utilizar o pacote `rtweet`.
- Há diversos outros pacotes para acessar a API do Twitter.
- 'Twarc' em Python é um excelente pacote.

```
# Faça o download do pacote
library(devtools)
install_github("mkearney/rtweet")
```

Ative as Credenciais

```
library(rtweet)
library(tidyverse)

create_token(app=app_name,
             consumer_key=consumer_key,

             consumer_secret=consumer_secret,

             access_token = access_token,

             access_secret = access_token_secret)
```

```
<Token>
<oauth_endpoint>
 request:   https://api.twitter.com/oauth/request_token
 authorize: https://api.twitter.com/oauth/authenticate
 access:    https://api.twitter.com/oauth/access_token
<oauth_app> RetornoBot
```

Search Tuítes (Sete Dias)

```
bolsonaro_tweets <-search_tweets("Bolsonaro",

                    n=100, include_rts = FALSE)
```

```
Observations: 98
Variables: 90
$ user_id                 <chr> "1177402598430785536", "24212583
$ status_id               <chr> "1206694563974385677", "12066945
$ created_at              <dttm> 2019-12-16 21:56:18, 2019-12-16
$ screen_name             <chr> "RobsonQueirz1", "ariquilha", "i
$ text                    <chr> "Bolsonaro bolsonaro bolsonaro b
$ source                  <chr> "Twitter for Android", "Twitter
$ display_text_width      <dbl> 100, 150, 89, 98, 25, 23, 85, 11
$ reply_to_status_id      <chr> NA, "1206693868659453952", NA, N
$ reply_to_user_id        <chr> NA, "43340387", NA, NA, NA, NA,
$ reply_to_screen_name    <chr> NA, "DCM_online", NA, NA, NA, NA
$ is_quote                <lgl> TRUE, FALSE, FALSE, FALSE, TRUE,
$ is_retweet              <lgl> FALSE, FALSE, FALSE, FALSE, FALS
$ favorite_count          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ retweet_count           <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
$ quote_count             <int> NA, NA, NA, NA, NA, NA, NA, NA,
$ reply_count             <int> NA, NA, NA, NA, NA, NA, NA, NA,
$ hashtags                <list> [NA, NA, NA, NA, NA, NA, NA, NA
$ symbols                 <list> [NA, NA, NA, NA, NA, NA, NA, NA
```

Filter Tweets: Live-time

```
bolsonaro_tweets <- stream_tweets("Bolsonaro",

                  n=100, include_rts = FALSE)

bolsonaro_tweets$text[1]
```

```
[1] "@marcielivxx @moura_101 Ta, mas o que Bolsonaro tem a ver c
```

Selecione Trending Topics

```
tt <- get_trends("Brazil")
tt$trend[1:10]
```

```
 [1] "Paulo Freire"                "#fancamsareoverparty"
 [3] "#BemVindoLeo"                "LANCEI A BRABA"
 [5] "BARBARA LABRES NO SPORTV"    "#CopiarNãoÉRoubar"
 [7] "#OiOiOi51"                   "#MTVHitspectiva150"
 [9] "Sidão"                       "Granata"
```

Acessar Tuítes por Usuário (Banco de Dados da Alepa)

```
library(tidyverse)

deputados <- read_csv("C:/Users/Tiago Ventura/Dropbox/
                      webscraping_workshop_ufpa/
                      html/dia_01/deputados_para.csv")

# Selecionando os que possuem twitter
deputados_twitter <- deputados %>%
                        filter(!is.na(twitter))
# Vamos extrair somente as tags
names <- str_remove_all(deputados_twitter$twitter,

        "https://www.twitter.com/|https://twitter.com/"  )
```

```
 [1] "DepAlexSantiago" "bordalopt"       "cilenecouto"
 [4] "CaveiraDelegado" "dilvandafaroPT"  "dilvandafaroPT"
 [7] "dirceutencaten"  "doutordanielpa"  "drjaques"
[10] "elielfaustino10" "igornormando"    "marinorbrito"
[13] "depcarmona"      "deprsantos"      "deputadothiago"
```

Capturar Tuítes na Timeline

```
dep1_tweets <- get_timelines(names[1], n = 5)
dep1_tweets$text[1]
```

```
[1] "A convite da comunidade evangélica dirigida pelo pastor Ben
```

Informações do Usuário

```
dep1_usuarios <- lookup_users(names[1])
dep1_usuarios$screen_name[1]
```

```
[1] "DepAlexSantiago"
```

Postar Tuites

```
post_tweet("Eu estou postando esse tweet a
           partir do pacote rtweet para
            mostrar aos alunos da UFPA
           o incrível mundo do R")
```

Analizando a Time-Line dos Deputados do Pará

Coletar Tweets

```
# Coletar Tweets

text <- map(names, ~get_timelines(.x, n = 3000))
```

```
# Combinar todos

tweets <- bind_rows(text) %>% select(screen_name, text)
```

bordalopt	CaveiraDelegado	CileneCouto	DepAlexSantiago
2926	21	1717	2
depcarmona	deprsantos	deputadothiago	dilvandafaroPT
1003	1426	80	308
DirceuTenCaten	doutordanielpa	drjaques	ElielFaustino10
2442	125	685	1960
igornormando	marinorbrito		
2978	2991		

Criando um Corpus

```
library(quanteda)
library(tidytext)

# Crie um corpus de textos
corpus <- corpus(tweets$text)
docvars(corpus) <- data_frame(deputados = tweets$screen_name)
```

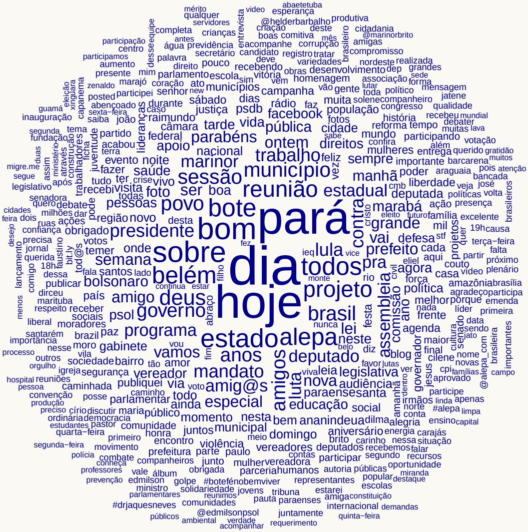
Limpando os Dados (DFM)

```
# Crie uma Document-Feature Matrix
palavras <- c("https","t.co", "http")
dfm <- tokens(corpus, remove_punct = TRUE,

      remove_numbers = TRUE, remove_symbols=TRUE) %>%

  tokens_select(., min_nchar = 3L) %>%

  tokens_remove(., c(stopwords("pt"),
                     palavras)) %>%
      dfm()
summary(dfm)
```

```
   Length     Class      Mode
706954992       dfm        S4
```

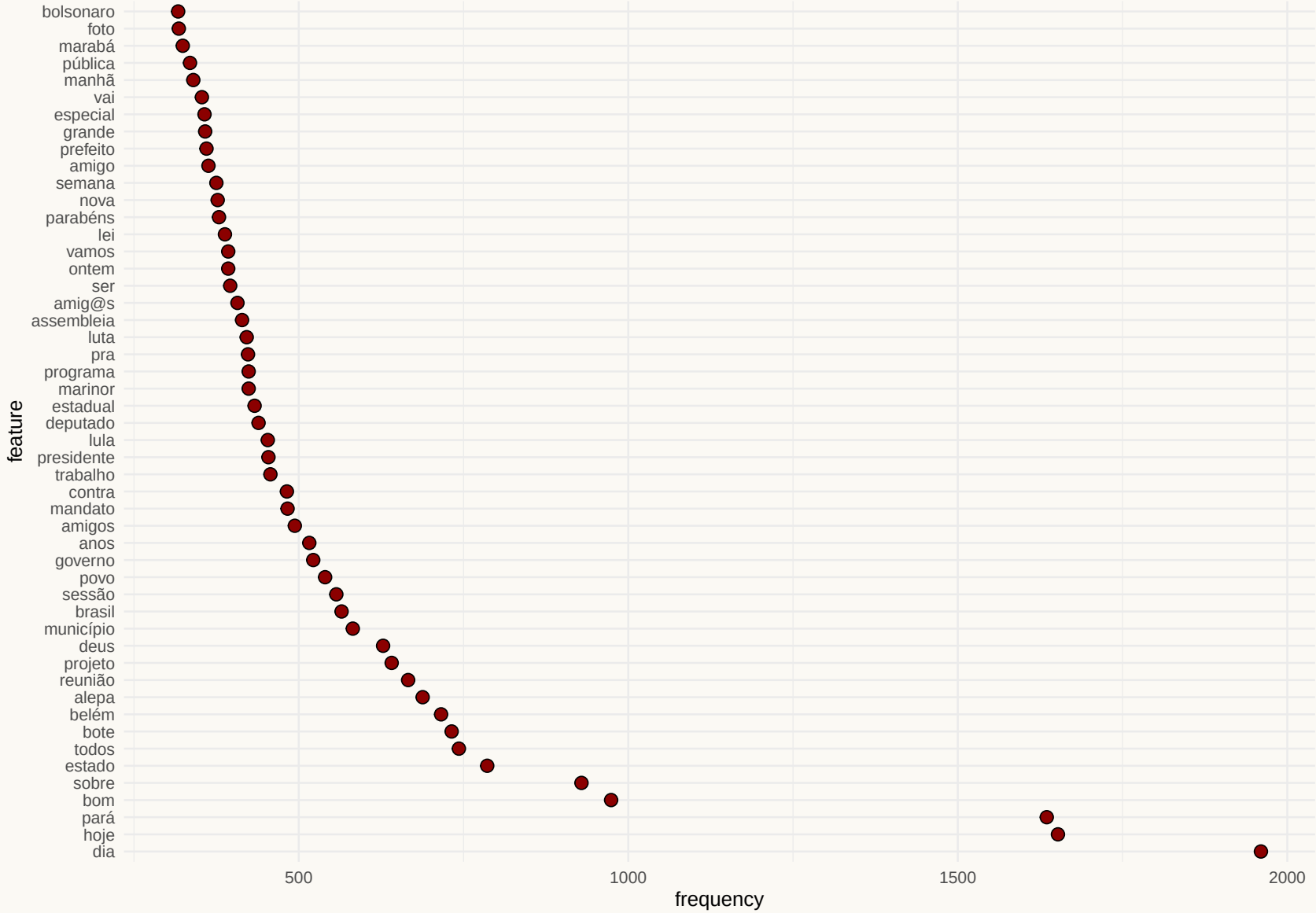
Nuvem de Palavras

```
textplot_wordcloud(dfm)
```


pará
dia
hoje
bom
bote
sobre
belém
todos
reunião
município
sessão
estado
alepa
projeto
brasil
deus
governo
presidente
povo
lula
prefeito
trabalho
marinor
estadual
amigos
assembleia
deputado
anos
mandato
amig@s
especial
programa
vamos
bolsonaro
semana
parabéns
ontem
direitos
sempre
manhã
marabá
grande
vai
pra
agora
contra
luta
nova
legislativa
educação
audiência
amigo
psol
pessoas
saúde
foto
ser
boa
apoio
nacional
tarde
vida
pública
cidade
facebook
gabinete
vereador
publiquei
ainda
momento
municipal
nesta
domingo
aniversário
ananindeua
dilma
alegria
política
agenda
frente
lei
neste
social
irmãos

Frequencia de Palavras

```
features_dfm <- textstat_frequency(dfm,
                                   n = 50)

# Sort by reverse frequency order
features_dfm$feature <- with(features_dfm,

                             reorder(feature, -frequency))
ggplot(features_dfm,
       aes(x = feature, y = frequency)) +

  geom_point(shape=21, size=3,
             fill="darkred") +

    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +

    theme_minimal() +
```

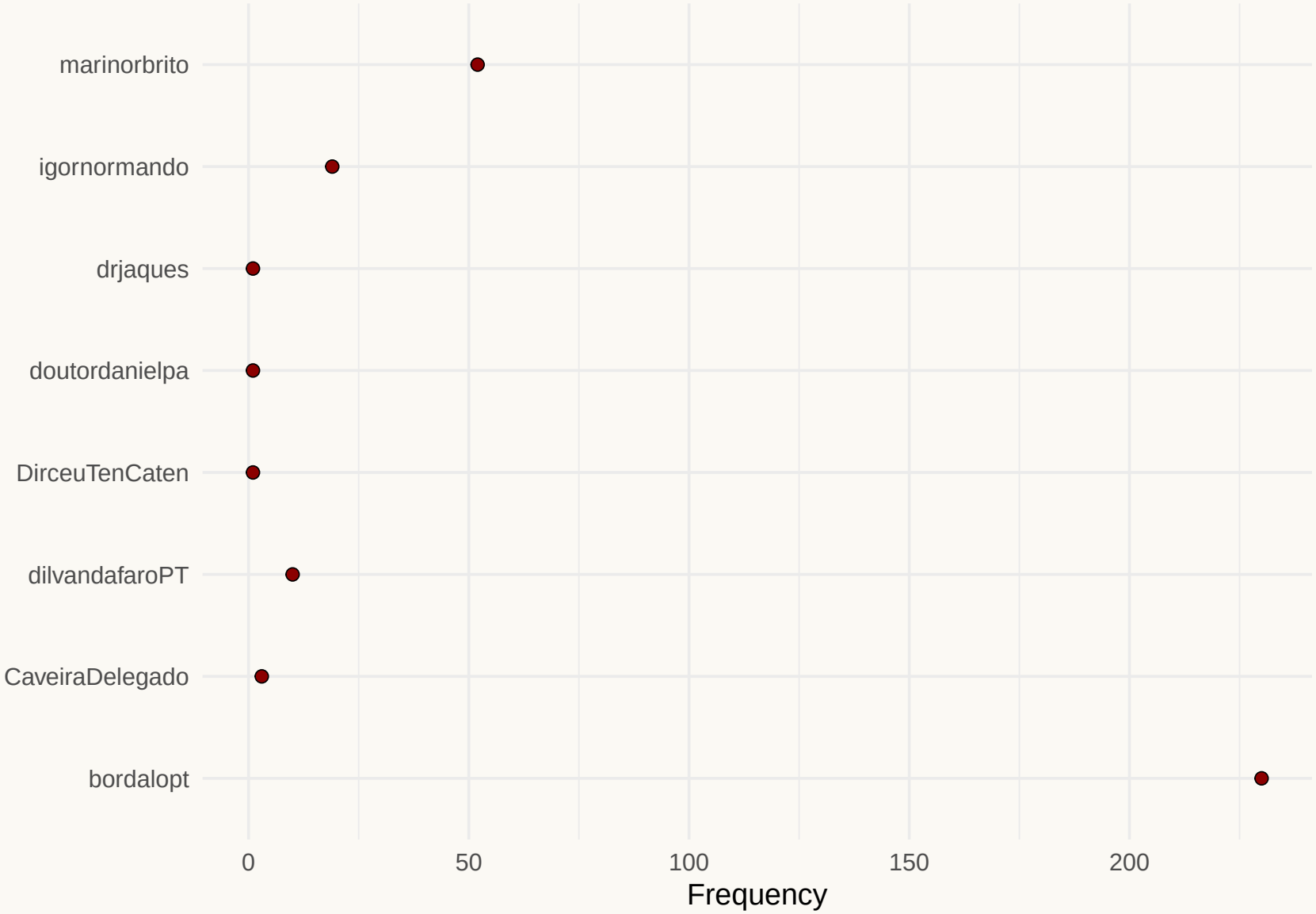

bolsonaro
foto
marabá
pública
manhã
vai
especial
grande
prefeito
amigo
semana
nova
parabéns
lei
vamos
ontem
ser
amig@s
assembleia
luta
pra
programa
marinor
estadual
deputado
lula
presidente
trabalho
contra
mandato
amigos
anos
governo
povo
sessão
brasil
município
deus
projeto
reunião
alepa
belém
bote
todos
estado
sobre
bom
pará
hoje
dia
feature
500
1000
1500
2000
frequency

Frequencia por Grupos

```
# Get frequency grouped by president
freq_grouped <- textstat_frequency(dfm,

                        groups = "deputados")
# Filtrar Bolsonaro
freq_bolsonaro <- subset(freq_grouped,

             freq_grouped$feature %in% "bolsonaro")
```

Grafico

```
ggplot(freq_bolsonaro, aes(x = group, y = frequency)) +
geom_point(shape = 21, size = 3, fill = "darkred") +
xlab(NULL) +
ylab("Frequency") +
theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
theme_minimal(base_size = 16) + coord_flip()
```


marinorbrito
igornormando
drjaques
doutordanielpa
DirceuTenCaten
dilvandafaroPT
CaveiraDelegado
bordalopt
0
50
100
150
200
Frequency

Modelos de Tópicos

Módelo de Tópicos: Modelo Estatístico para análide de texto.

+ Documentos são extraídos de distribuições estatisticas de tópicos.
+ Palavras são associadas há cada tópico.
+ Totalmente não supervisionado.
+ Algoritmo máximiza a probabilidadede palavras aparecerem juntas.

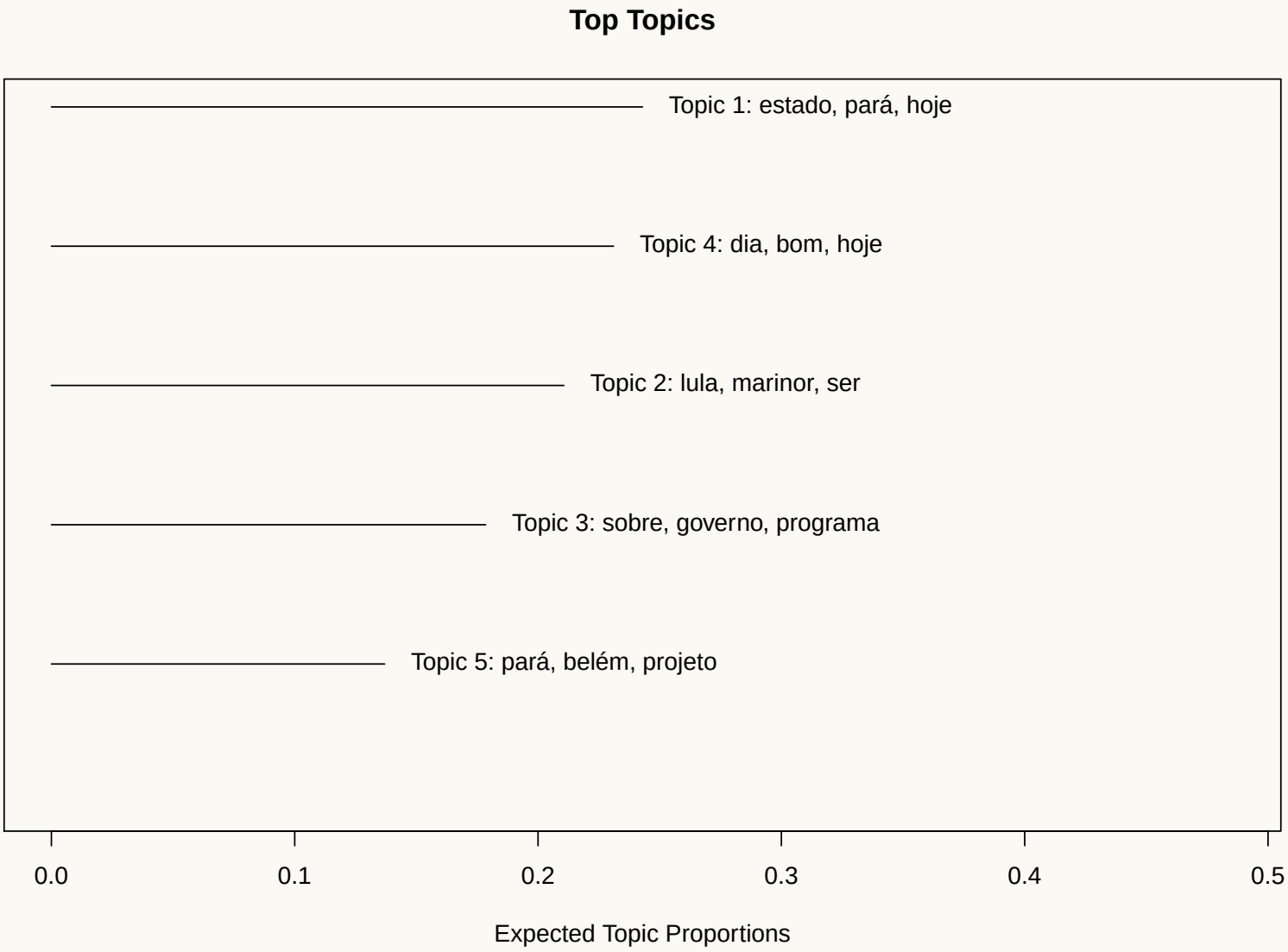
```r
# Para estimar os modelos de tópicos,
# vamos usar o pacote de R `STM`.
# Este pacote exige uma pequena transformação no objeto
library(stm)
dfm_stm <-  quanteda::convert(dfm, to = "stm")

model_parag <- stm(dfm_stm$documents,

                   dfm_stm$vocab, K = 5, data = dfm_stm$meta,
                   init.type = "Spectral",verbose = FALSE)
```

```
Topic 1 Top Words:
     Highest Prob: estado, pará, hoje, bote, reunião, alepa, ses
     FREX: bote, reunião, mandato, marabá, legislativa, comissão
     Lift: eldorado, reunimos, trabalhos, visitando, #13podcast,
     Score: bote, reunião, sessão, marabá, legislativa, alepa, m
```

```
Topic 2 Top Words:
     Highest Prob: lula, marinor, ser, brasil, vai, temer, agora
     FREX: lula, ter, moro, vou, deve, democracia, nunca
     Lift: #14j, #1deabril, #30m, #asemana, #bolsonaro, #bolsona
     Score: lula, temer, moro, ser, diz, bolsonaro, país
```

```
Topic 5 Top Words:
     Highest Prob: pará, belém, projeto, nova, foto, trabalho, f
     FREX: belém, projeto, nova, foto, facebook, publiquei, fren
     Lift: #elenunca, #souhumana, #souhumano, agressões, atuar,
     Score: facebook, publiquei, foto, projeto, publicar, belém,
```

Plotando os Topicos.



Exercícios.

1. Colete dados do Twitter sobre um tópico de seu interesse.
2. Apresente uma análise interessante sobre esses dados. Pode ser uma nuvem de palavra, um gráfico de frequência, fique à vontade.