Análise de Sentimentos no Twitter sobre COVID-19 - postagens mais recentes

Sobre

Este trabalho busca fazer uma breve análise de sentimentos sobre as postagens mais recentes feitas no Twitter sobre a temática do novo coronavírus.

Baixando pacotes necessários

```
library(rtweet)
library(tidytext)
library(dplyr)
library(tidyr)
library(ggplot2)
library(wordcloud)
library(stringr)
library(lubridate)
library(kableExtra)
library(reshape2)
library(data.table)
```

Autenticação na API do Twitter

Para ter acesso à API do Twitter é preciso ter uma conta de desenvolvedor. Para fazer isso, é só seguir o tutorial disponível em:

https://cran.r-project.org/web/packages/rtweet/vignettes/auth.html

Após aprovação da sua developer account, é necessário autenticar sua sessão no R para que você possa baixar os tweets. Todos os dados expostos abaixo são retirados da sua conta de desenvolvedor:

twitter_token <- create_token(app = appname, consumer_key = key, consumer_secret = secret)

Agora vamos baixar os tweets para análise

Lendo os últimos 10.000 tweets em inglês que envolvam o termo covid19 sem incluir retweets:

```
covid19_tweets <- search_tweets(q = 'covid19',
                                n = 10000,
                                lang = 'en',
                                include_rts = F)
```

Table 1: Amostra do dataset

created_at	text
2021-01-03 14:11:34	@paulkrugman It'll all be fine, this is THE distraction that both sides are using to define the other side, and sadly, it appears to be getting enormous attention from both sides. The fate of the Senate will be fought over a mythical socialist presidency and a proto fascist dictator.
2021-01-03 14:06:21	@AndrewNoymer Paper and explainer for laypeople from UK doc on new variant (does not discuss implications for vaccine): https://t.co/lP0QJEa2Tv
2021-01-03 14:11:33	@JeromeAdamsMD Listening to this guy when his figures are contested on @CNN . It was 'his hope' ðŸ~,ðŸ~, Give the @Conservatives a call Jerome, they're quite adept at manipulating figures and excellent at failing to meet their targets. #COVID19 #COVIDVaccination #ToryShambles
2021-01-03 10:07:31	Except in #London ðŸ~,ðŸ~,ðŸ~,ðŸ~, Wonder where we've seen all this before ðŸ¤" Maybe when you shut the North down while keeping London open. #BorisTheLiar #ToryShambles #COVID19 @BorisJohnson https://t.co/8zB8fChaTv
2021-01-03 14:11:29	@AlHaill_RSA nigga blocked me dor calling out his bullshit .calling himself a king but he wrak3as fuck..#COVID19 #COVIDSecondWave #COVID19
2021-01-03 12:34:02	I can't believe there are idiots out there who ask ""why doesn't HIV/AIDS and cancer have a vaccine"" and in their mind they actually think they making a valid point yet they are just showing their ignorance ðŸ~,ðŸ~,ðŸ~, #lockdown #COVID19 #southafricancoronavirus #southafricanvaccine

Note:
Tweets - Covid-19

```
amostra %>%
  select(created_at, text) %>%
  kbl(caption = 'Amostra do dataset') %>%
  kable_paper("striped", full_width = F) %>%
  row_spec(0, bold = T) %>%
  column_spec(1, width = "10em")  %>% column_spec(2, width = '30em') %>%
  footnote('Tweets - Covid-19')
```

O dataframe gerado com o comando da leitura dos tweets traz diversas colunas que não serão usadas. Para que a apresentação não fique poluída, o dicionário de dados será apresentado posteriormente com o dataframe já filtrado. Acima, preferi trazer na amostra apenas a data de criação do post e o post em si.

Lexicons e Análise de sentimentos

Para análise de sentimentos, existem algumas listas prontas (lexicons) do pacote tidytext que podem ser bem úteis na análise.

Analisando lexicon nrc:

```
nrc_lexicon <- as.data.frame(unique(get_sentiments('nrc')$sentiment))
colnames(nrc_lexicon) <- 'NRC Lexicon'
nrc_lexicon %>% kbl() %>% kable_paper("striped", full_width = F) %>% row_spec(0, bold = T)
```

NRC Lexicon
trust
fear
negative
sadness
anger
surprise
positive
disgust
joy
anticipation

Lexicon nrc foi desenvolvido por Saif Mohammad e Peter Turney.

Foi escolhida pois trata de diversos sentimentos associados às palvaras. São eles: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

Limpeza, organização e transformação dos dados

Gerando novo dataset de trabalho:

- Seleção de colunas: user_id (identificador único do tweet), data de criação, texto da publicação;
- unnest para quebrar o texto em tokens individuais;
- Retirada de palavras que não têm significado na análise de sentimentos (stop words);
- Retirada de @, # e https bem como postagens apenas numéricas.

```
tidy_tweets <- covid19_tweets %>%
    select(user_id, created_at, text) %>%
    unnest_tokens(word, text, token = 'tweets') %>%
    anti_join(stop_words)
```

```
tidy_tweets$word <- tidy_tweets$word %>%
    str_replace_all(pattern = 'https://t.co/\\w+', replacement = '') %>%
    str_replace_all(pattern = '#', replacement = '') %>%
    str_replace_all(pattern = '@', replacement = '') %>%
    str_replace_all(pattern = '[[:digit:]]', replacement = '')

tidy_tweets <- tidy_tweets %>%
    filter(!word == '')
```

```
head(tidy_tweets) %>% kbl() %>% kable_paper("striped", full_width = F)
```

user_id	created_at	word
255925055	2021-01-03 13:20:48	coronavirus
255925055	2021-01-03 13:20:48	ethiopia
255925055	2021-01-03 13:20:48	covid
255925055	2021-01-03 13:20:48	reported
255925055	2021-01-03 13:20:48	ethiopia
255925055	2021-01-03 13:20:48	january

Dicionário de dados

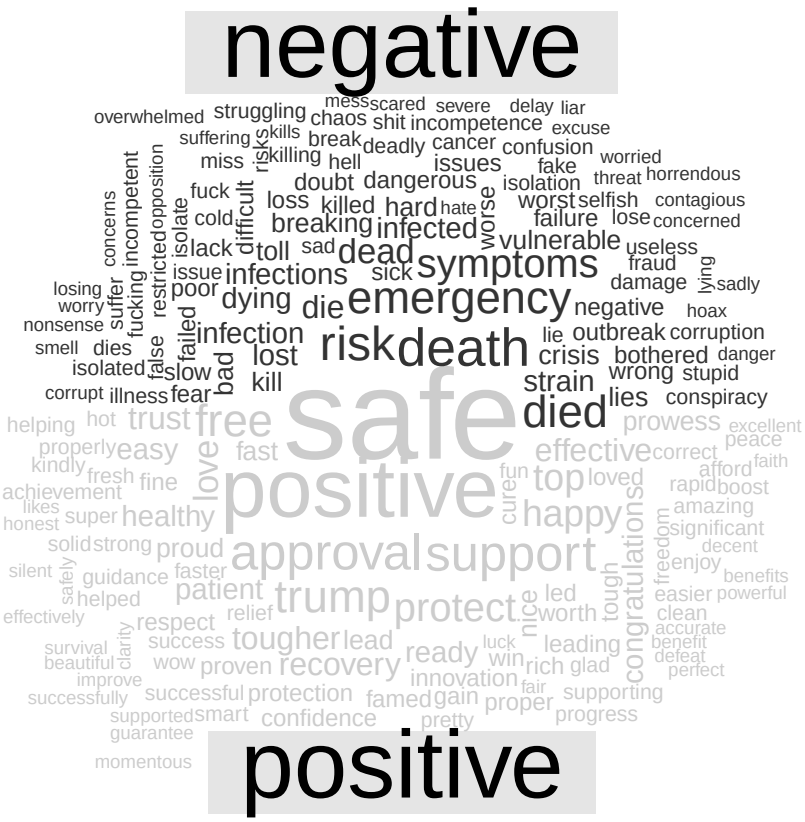
```
dic <- data.frame(variavel = c('user_id', 'created_at', 'word'),
                  descricao = c('identificação única do usuário',
                                'data da postagem',
                                'palavra selecionada da postagem'),
                  formato = c('caracteres', 'datas', 'caracteres'))

dic %>%
  rename('Coluna' = variavel,
         'Descrição' = descricao,
         'Tipo de dado' = formato) %>%
  kbl() %>% kable_paper("striped", full_width = F)
```

Coluna	Descrição	Tipo de dado
user_id	identificação única do usuário	caracteres
created_at	data da postagem	datas
word	palavra selecionada da postagem	caracteres

Analisando as palavras mais recorrentes em um mapa de palavras

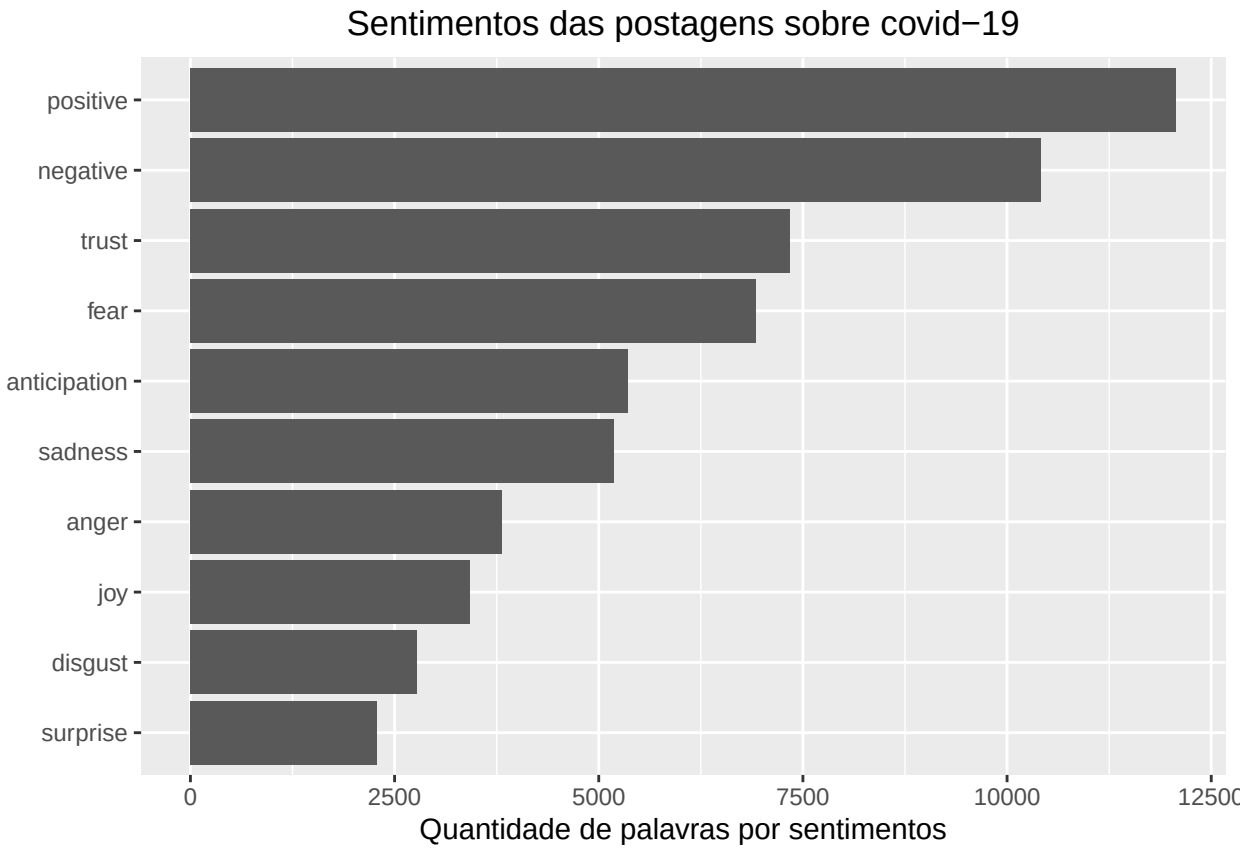
```
tidy_tweets %>%
    inner_join(get_sentiments('bing')) %>%
    count(word, sentiment ,sort = T) %>%
    filter(!word %in% c('covid', 'pandemic', 'coronavirus', 'virus')) %>%
    acast(word ~ sentiment, value.var = "n", fill = 0) %>%
    comparison.cloud(colors = c("gray20", "gray80"), max.words = 200)
```



Podemos observar palavras negativas como mortes, infecções, risco, vulnerabilidade e sintomas. Já nas palavras positivas vemos segurança, suporte, aprovação, recuperação e saúde. Isso pode indicar o medo das pessoas de se infectarem pelo novo coronavírus bem como suas consequências para saúde, porém também pode apontar um certo sentimento de segurança sobre a recuperação nesta pandemia.

Sentimentos das postagens

```
tidy_tweets %>%
    inner_join(get_sentiments('nrc')) %>%
    count(sentiment, sort = T) %>%
    mutate(sentiment = reorder(sentiment, n)) %>%
    ungroup() %>%
    ggplot(aes(x = sentiment, y = n)) +
    geom_col(show.legend = F) +
    coord_flip() +
    theme(plot.title = element_text(hjust = 0.5)) +
    labs(title = 'Sentimentos das postagens sobre covid-19',
         y = 'Quantidade de palavras por sentimentos', x = NULL)
```

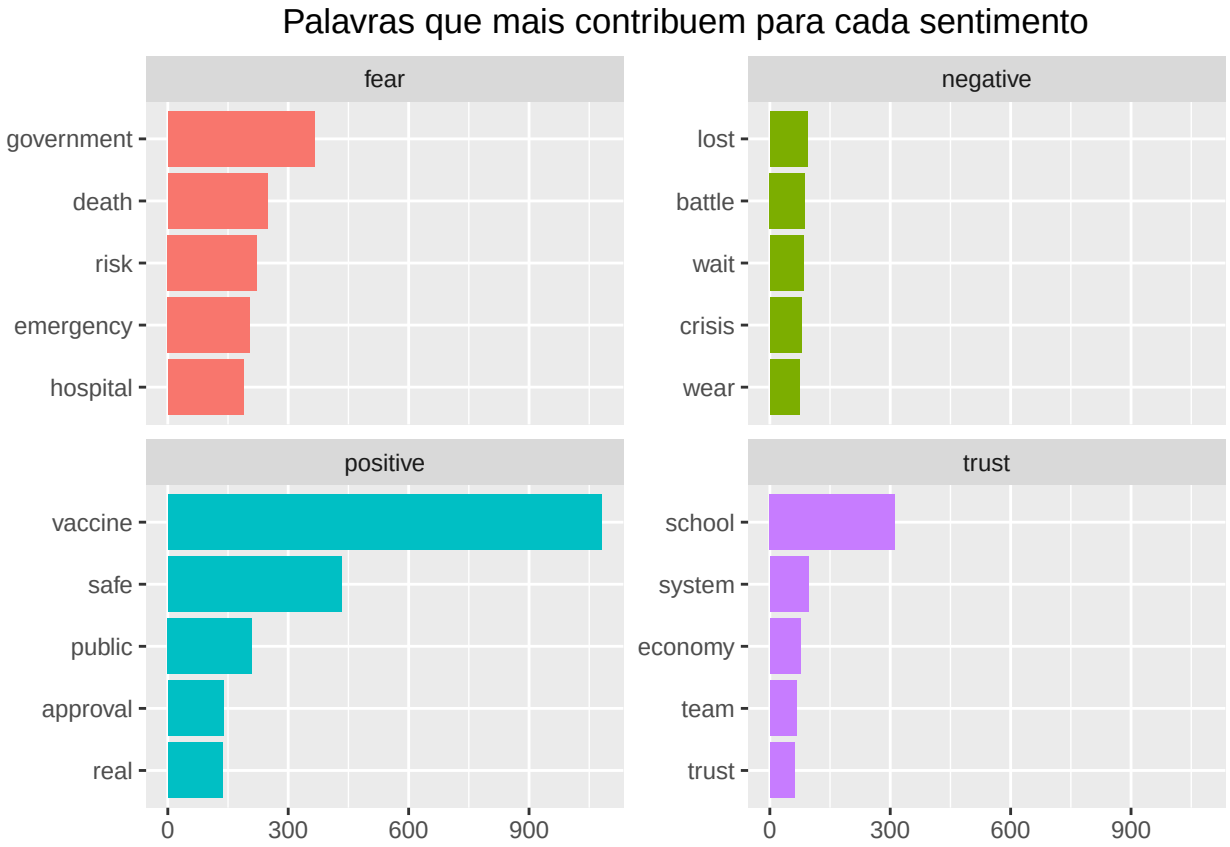


Os sentimentos mais observados nas postagens se dividem entre positivos, negativos, de confiança e medo. Isso pode demosntrar que a população quer acreditar na superação da pandemia, no entanto, ainda há medo do que podemos enfrentar em um futuro próximo.

Palavras mais relevantes de acordo com o sentimento

Selecionando apenas os 4 sentimentos mais relevantes observados nas últimas postagens e filtrando palavras que poderiam atrapalhar a análise, apresentamos aquelas palavras que mais contribuiram para cada sentimento.

```
tidy_tweets %>%
    inner_join(get_sentiments('nrc')) %>%
    count(word, sentiment, sort = T) %>%
    group_by(sentiment) %>%
    filter(sentiment %in% c('positive', 'negative', 'trust', 'fear')) %>%
    filter(!word %in% c('covid', 'pandemic', 'coronavirus', 'virus')) %>%
    mutate(word = reorder(word, n)) %>%
    slice_max(word, n = 5) %>%
    ungroup() %>%
    ggplot(aes(x = n, y = word, fill = sentiment)) +
    geom_col(show.legend = F) +
    facet_wrap(~ sentiment, ncol = 2, scales = 'free_y') +
    theme(plot.title = element_text(hjust = 0.5)) +
    labs(title = 'Palavras que mais contribuem para cada sentimento',
         x = NULL, y = NULL)
```



Do sentimento medo, a palavra que mais apareceu foi governo. Talvez demonstre certa insegurança da população sobre as medidas governamentais.

Do sentimento negativo, observamos palavras como batalha e crise, demonstrando o cenário atual que enfretamos.

Do sentimento positivo, vemos a palavra vacina, podendo indicar a grande esperança da população para esse ano novo.

E, por fim, do sentimento de confiança, vemos palavras como sistema, economia e time, podendo indicar certa confiança da população na ação coletiva e na retomada da economia.

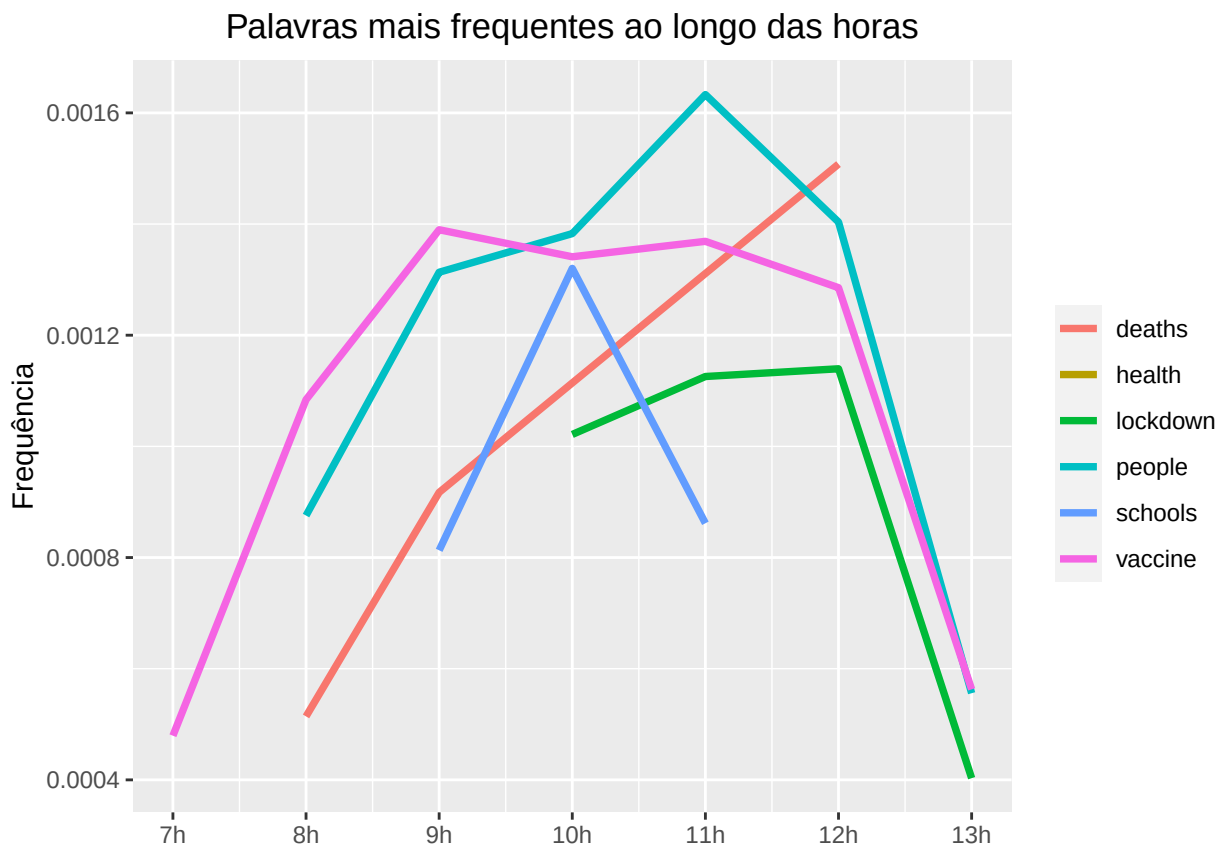
Frequência das palavras ao longo das horas

Como o termo covid-19 é muito falado no mundo inteiro, os 10.000 posts coletados foram todos realizados no dia 03/01/2020 entre às 7h e 13h.

Nesse sentido, podemos analisar como a frequência das palavras mais comentadas mudou a cada hora:

```
tidy_tweets %>%
    mutate(time = floor_date(created_at, unit = '1 hour')) %>%
    mutate(time = hour(time)) %>%
    count(word, time) %>%
    mutate(freq = n / sum(n)) %>%
    filter(!word %in% c('covid', 'pandemic', 'coronavirus', 'virus',
                        'amp', 'india', 'marr','borisjohnson', 'total')) %>%
    filter(freq > .0004) %>%
```

```
    group_by(time) %>%
    slice_max(freq, n = 4) %>%
    ungroup() %>%
    ggplot(aes(x = time, y = freq, color = word)) +
    geom_line(size = 1.3) +
    scale_x_continuous(limits = c(7, 13),
                       breaks = seq(7, 13, 1),
                       labels = paste0(seq(7, 13, 1), 'h')) +
    theme(legend.title = element_blank()) +
    theme(plot.title = element_text(hjust = 0.5)) +
    labs(title = 'Palavras mais frequentes ao longo das horas',
         y = 'Frequência',
         x = NULL)
```



Das palavras mais frequntes, observamos que, diante das postagens coletadas, em nenhum momento deixou-se de falar sobre a vacina. De fato, ela pode ser considerada um dos desejos mais latentes da população mundial e a grande esperança para superação do covid-19 neste ano de 2021.

Outras palavras como pessoas, mortes e lockdown foram observadas também na maior parte das horas, podendo ser consideradas como preocupeções da população neste cenário de pandemia.

Conclusão

Nesta análise rápida buscamos entender o que as pessoas que utilizam o twitter em inglês estavam postando sobre o covid-19 nas últimas horas. Por ser um tópico bastante falado no mundo inteiro, os 10.000 posts

coletados referiram-se às últimas seis horas apenas. Tentei entender quais as palavras foram mais recorrentes, quais eram os sentimentos mais relevantes das postagens e como a frequência das palavras mais postadas mudou ao longo das horas.

Referências

- Text Mining with R: A Tidy Approach - Julia Silge and David Robinson (last built on 2020-11-10)
- Formação Cientista de Dados - Data Science Academy