Scraping Twitter Data

Bennur Kaya

2022-10-11

```
#install.packages("rtweet")
#install.packages("magrittr") # package installations are only needed the fir
st time you use it
#install.packages("dplyr")    # alternative installation of the %>%
#install.packages("stringr")
#install.packages("tidytext")
#install.packages("rvest")

library(rtweet)
library(magrittr) # needs to be run every time you start R and want to use %>
%
library(dplyr)    # alternatively, this also loads %>%

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(stringr)
library(tidytext)
library(rvest)
```

Comparing Twitter Accounts: I will use Twitter data of Boris Johnson who is former Prime Minister of the United Kingdom of Great Britain and Liz Truss who is the current Prime Minister.

```
#Alexander Boris de Pfeffel Johnson Hon FRIBA is a British politician who ser
ved as Prime Minister of the United Kingdom and Leader of the Conservative Pa
rty from 2019 to 2022.

b_johnson_tweets <- get_timeline("BorisJohnson", n = 3200)
b_johnson_tweets

## # A tibble: 3,199 × 43
##    created_at                 id id_str       full_…¹ trunc…² displ…³ entiti
es
##    <dttm>                  <dbl> <chr>        <chr>   <lgl>     <dbl> <list>
##  1 2022-09-30 15:54:42 1.58e18 15758466041… "Vladi… FALSE       273 <named
list>
##  2 2022-09-21 16:52:27 1.57e18 15725996455… "I too… FALSE        67 <named
list>
##  3 2022-09-11 18:54:38 1.57e18 15690065165… "Let u… FALSE       202 <named
list>
##  4 2022-09-09 17:22:30 1.57e18 15682585509… "She s… FALSE       248 <named
list>
##  5 2022-09-08 20:21:38 1.57e18 15679412441… "State… FALSE        62 <named
list>
##  6 2022-09-08 20:21:13 1.57e18 15679411417… "State… FALSE        62 <named
list>
##  7 2022-09-08 20:20:59 1.57e18 15679410807… "State… FALSE        62 <named
list>
##  8 2022-09-05 17:49:28 1.57e18 15668157892… "It's … FALSE       254 <named
list>
##  9 2022-09-05 14:13:51 1.57e18 15667615275… "Congr… FALSE       271 <named
list>
## 10 2022-09-05 14:13:51 1.57e18 15667615252… "I hav… FALSE       237 <named
list>
## # … with 3,189 more rows, 36 more variables: source <chr>,
## #   in_reply_to_status_id <dbl>, in_reply_to_status_id_str <chr>,
## #   in_reply_to_user_id <dbl>, in_reply_to_user_id_str <chr>,
## #   in_reply_to_screen_name <chr>, geo <list>, coordinates <list>,
## #   place <list>, contributors <lgl>, is_quote_status <lgl>,
## #   retweet_count <int>, favorite_count <int>, favorited <lgl>,
## #   retweeted <lgl>, lang <chr>, possibly_sensitive <lgl>, …

# select variables of interest

b_johnson_short <- b_johnson_tweets %>%
  filter(retweeted == "FALSE") %>%
  mutate(screen_name = "b_johnson") %>%
  select(created_at, screen_name, text,
         favorite_count, retweet_count)

b_johnson_short

## # A tibble: 3,199 × 5
##    created_at          screen_name text                          favor…¹ r
etwe…²
##    <dttm>              <chr>       <chr>                           <int>
<int>
##  1 2022-09-30 15:54:42 b_johnson   "Vladimir Putin your speech …  175019
21852
##  2 2022-09-21 16:52:27 b_johnson   "I took the oath today in th…   30280
2334
##  3 2022-09-11 18:54:38 b_johnson   "Let us always remember the …   23132
1864
##  4 2022-09-09 17:22:30 b_johnson   "She showed the world not ju…   44327
```

```
5413
##  5 2022-09-08 20:21:38 b_johnson   "Statement on the death of H…   54020
5686
##  6 2022-09-08 20:21:13 b_johnson   "Statement on the death of H…   44854
4846
##  7 2022-09-08 20:20:59 b_johnson   "Statement on the death of H…  121657
15404
##  8 2022-09-05 17:49:28 b_johnson   "It's been a privilege to wo…   17870
2002
##  9 2022-09-05 14:13:51 b_johnson   "Congratulations to @trussli…   13174
1653
## 10 2022-09-05 14:13:51 b_johnson   "I have been proud to serve …   39582
3399
## # … with 3,189 more rows, and abbreviated variable names ¹favorite_count,
## #   ²retweet_count
```

```
# learning some things about data types
b_johnson_short %>% select(text, retweet_count)
```

```
## # A tibble: 3,199 × 2
##    text                                                                   r
etwe…¹
##    <chr>
<int>
##  1 "Vladimir Putin your speech is a fraud and a disgrace. The world mus…
21852
##  2 "I took the oath today in the House of Commons. God Save the King \U…
2334
##  3 "Let us always remember the victims of terror on 9/11. Let us honour…
1864
##  4 "She showed the world not just how to reign over a people, she showe…
5413
##  5 "Statement on the death of Her Majesty Queen Elizabeth II (3/3) http…
5686
##  6 "Statement on the death of Her Majesty Queen Elizabeth II (2/3) http…
4846
##  7 "Statement on the death of Her Majesty Queen Elizabeth II (1/3) http…
15404
##  8 "It's been a privilege to work with you @ZelenskyyUa, and I look for…
2002
##  9 "Congratulations to @trussliz on her decisive win. I know she has th…
1653
## 10 "I have been proud to serve as leader of the Conservative Party for …
3399
## # … with 3,189 more rows, and abbreviated variable name ¹retweet_count
```

```
class(b_johnson_short$text)
```

```
## [1] "character"
```

```
class(b_johnson_short$retweet_count)
```

```
## [1] "integer"

length(b_johnson_short$text)

## [1] 3199

length(b_johnson_short[1,3])

## [1] 1

str_length(b_johnson_short[1,3])

## [1] 273

b_johnson_short %>%
  select(text) %>%
  head(2) %>%
  pull()

## [1] "Vladimir Putin your speech is a fraud and a disgrace. The world must
never accept your sham referendums or your cruel and illegal attempt to colon
ise Ukraine. We stand with the people of Ukraine and will support them withou
t flinching until their country is whole and free."
## [2] "I took the oath today in the House of Commons. God Save the King GB h
ttps://t.co/3FMPMRtR0N"

# Mary Elizabeth Truss (born 26 July 1975) is a British politician who is the
current prime minister of the United Kingdom and leader of the Conservative P
arty.

liztruss <- get_timeline("trussliz", n = 3200)
liztruss

## # A tibble: 3,199 × 43
##    created_at                 id id_str       full_…¹ trunc…² displ…³ entiti
es
##    <dttm>                  <dbl> <chr>        <chr>   <lgl>     <dbl> <list>
##  1 2022-10-12 23:07:19 1.58e18 15803041291… "Today… FALSE       278 <named
list>
##  2 2022-10-12 11:38:09 1.58e18 15801306945… "We’re… FALSE       246 <named
list>
##  3 2022-10-11 21:29:06 1.58e18 15799170239… "We're… FALSE       263 <named
list>
##  4 2022-10-11 17:45:03 1.58e18 15798606403… "The o… FALSE       276 <named
list>
##  5 2022-10-10 19:48:44 1.58e18 15795293769… "RT @1… FALSE       140 <named
list>
##  6 2022-10-10 18:17:52 1.58e18 15795065122… "Delig… FALSE       160 <named
list>
##  7 2022-10-10 17:40:15 1.58e18 15794970435… "The a… FALSE       252 <named
list>
##  8 2022-10-10 12:37:31 1.58e18 15794208566… "This … FALSE       280 <named
```

```
list>
##  9 2022-10-08 20:22:25 1.58e18 15788130771… "I am … FALSE       238 <named
list>
## 10 2022-10-08 13:09:03 1.58e18 15787040186… "For t… FALSE       277 <named
list>
## # … with 3,189 more rows, 36 more variables: source <chr>,
## #   in_reply_to_status_id <dbl>, in_reply_to_status_id_str <chr>,
## #   in_reply_to_user_id <dbl>, in_reply_to_user_id_str <chr>,
## #   in_reply_to_screen_name <chr>, geo <list>, coordinates <list>,
## #   place <list>, contributors <lgl>, is_quote_status <lgl>,
## #   retweet_count <int>, favorite_count <int>, favorited <lgl>,
## #   retweeted <lgl>, lang <chr>, quoted_status_id <dbl>, …

# select variables of interest
liztruss_short <- liztruss %>%
  filter(retweeted == "FALSE") %>%
  mutate(screen_name = "liztruss") %>%
  select(created_at, screen_name, text,
         favorite_count, retweet_count)

liztruss_short

## # A tibble: 3,199 × 5
##    created_at          screen_name text                            favor…¹ r
etwe…²
##    <dttm>              <chr>       <chr>                             <int>
<int>
##  1 2022-10-12 23:07:19 liztruss    "Today 143 countries stand u…    8857
1227
##  2 2022-10-12 11:38:09 liztruss    "We’re taking action to prot…    2018
343
##  3 2022-10-11 21:29:06 liztruss    "We're making sure millions …    1784
336
##  4 2022-10-11 17:45:03 liztruss    "The overwhelming internatio…    4243
709
##  5 2022-10-10 19:48:44 liztruss    "RT @10DowningStreet: ⚽ The…       0
135
##  6 2022-10-10 18:17:52 liztruss    "Delighted to meet the @Lion…    1721
195
##  7 2022-10-10 17:40:15 liztruss    "The appalling attacks on ci…    4662
806
##  8 2022-10-10 12:37:31 liztruss    "This country has come a lon…    2999
439
##  9 2022-10-08 20:22:25 liztruss    "I am shocked and saddened b…    5513
514
## 10 2022-10-08 13:09:03 liztruss    "For the first time in over …    5136
936
## # … with 3,189 more rows, and abbreviated variable names ¹favorite_count,
## #   ²retweet_count
```

```
# learning some things about data types
liztruss_short %>% select(text, retweet_count)

## # A tibble: 3,199 × 2
##    text                                                                    r
etwe…¹
##    <chr>
<int>
##  1 "Today 143 countries stand united in condemnation of Putin’s illegal…
1227
##  2 "We’re taking action to protect people from the rising energy costs …
343
##  3 "We're making sure millions of people up and down the country keep m…
336
##  4 "The overwhelming international support for Ukraine’s struggle stand…
709
##  5 "RT @10DowningStreet: ⚽ The @Lionesses are an inspiration, encourag…
135
##  6 "Delighted to meet the @Lionesses today.\n\nThe nation thanks you fo…
195
##  7 "The appalling attacks on civilian areas in Kyiv and elsewhere are a…
806
##  8 "This country has come a long way in how we talk about mental health…
439
##  9 "I am shocked and saddened by the tragic loss of life in Donegal.\n\…
514
## 10 "For the first time in over 20 years, millions of people in the USA …
936
## # … with 3,189 more rows, and abbreviated variable name ¹retweet_count

class(liztruss_short$text)

## [1] "character"

class(liztruss_short$retweet_count)

## [1] "integer"

length(liztruss_short$text)

## [1] 3199

length(liztruss_short[1,3])

## [1] 1

str_length(liztruss_short[1,3])

## [1] 278

liztruss_short %>%
  select(text) %>%
  head(2) %>%
  pull()
```

```
## [1] "Today 143 countries stand united in condemnation of Putin’s illegal a
ttempts to annex four regions of Ukraine. \n\nUnited against Russia’s hostile
actions, the @UN General Assembly vote is a clear sign that Putin is isolated
on the international stage.\n\nWe stand with Ukraine GBUA"
## [2] "We’re taking action to protect people from the rising energy costs ca
used by Putin’s barbaric campaign.  \n\nThis government is acting decisively
to make sure people and businesses across the country get secure, affordable
and fairly priced energy. https://t.co/syIval1qRL"
```

```
tweets <- bind_rows(b_johnson_short, liztruss_short)
tweets
```

```
## # A tibble: 6,398 × 5
##    created_at          screen_name text                           favor…¹ r
etwe…²
##    <dttm>              <chr>       <chr>                            <int>
<int>
##  1 2022-09-30 15:54:42 b_johnson   "Vladimir Putin your speech …  175019
21852
##  2 2022-09-21 16:52:27 b_johnson   "I took the oath today in th…   30280
2334
##  3 2022-09-11 18:54:38 b_johnson   "Let us always remember the …   23132
1864
##  4 2022-09-09 17:22:30 b_johnson   "She showed the world not ju…   44327
5413
##  5 2022-09-08 20:21:38 b_johnson   "Statement on the death of H…   54020
5686
##  6 2022-09-08 20:21:13 b_johnson   "Statement on the death of H…   44854
4846
##  7 2022-09-08 20:20:59 b_johnson   "Statement on the death of H…  121657
15404
##  8 2022-09-05 17:49:28 b_johnson   "It's been a privilege to wo…   17870
2002
##  9 2022-09-05 14:13:51 b_johnson   "Congratulations to @trussli…   13174
1653
## 10 2022-09-05 14:13:51 b_johnson   "I have been proud to serve …   39582
3399
## # … with 6,388 more rows, and abbreviated variable names ¹favorite_count,
## #   ²retweet_count
```

```
# some basic cleaning and extraction
# of meta data with regex
tweets <- tweets %>%
  mutate(
    # identify tweets with hashtags
    has_tag = str_detect(text, "#\\w+"),
    # how many at-mentions are there?
    n_at = str_count(text, "(^|\\s)@\\w+"),
    # extract first url
    url = str_extract(text, "(https?://\\S+)"),
    # remove at-mentions, tags, and urls
    clean_text = str_replace_all(text,
                                 "(^|\\s)(@|#|https?://)\\S+", " ") %>%
      str_replace_all("\\W+", " ")
  )
```

```
tweets %>%
  filter(has_tag == TRUE) %>%
  pull(text) %>%
  str_extract_all("#\\w+") %>%
  head(10)
```

```
## [[1]]
## [1] "#StandWithUkraine"
##
## [[2]]
## [1] "#AlevelResultsDay2022"
##
## [[3]]
## [1] "#Farm24"
##
## [[4]]
## [1] "#WEURO2022"
##
## [[5]]
## [1] "#WEURO2022"
##
## [[6]]
## [1] "#WEURO2022"
##
## [[7]]
## [1] "#WEURO2022"
##
## [[8]]
## [1] "#WEURO2022"
##
## [[9]]
## [1] "#EidMubarak"
##
## [[10]]
## [1] "#NHSBirthday"
```

```
tweets %>%
  filter(n_at > 0) %>%
  select(text)
```

```
## # A tibble: 3,634 × 1
##    text
##    <chr>
##  1 "It's been a privilege to work with you @ZelenskyyUa, and I look forwar
d to …
##  2 "Congratulations to @trussliz on her decisive win. I know she has the r
ight …
##  3 "Thank you my friend President @ZelenskyyUa for your kind words.\n\nUK
suppo…
##  4 "RT @RichardMarlesMP: It was an honour to join Prime Minister @BorisJoh
nson …
##  5 "RT @10DowningStreet: This war is only going to end one way. \n\nUkrain
e wil…
##  6 "RT @10DowningStreet: On Ukrainian Independence Day and every day, the
Unite…
##  7 "My thoughts are with Olivia Pratt-Korbel’s family and the people of Li
verpo…
##  8 "RT @10DowningStreet: Congratulations to everyone receiving their A and
T Le…
##  9 "@paula_hudgell @PointsofLight It was fantastic to see you again Tony.
You a…
## 10 "RT @JustinTrudeau: Update: Next week, we’re deploying up to 225 Canadi
an so…
## # … with 3,624 more rows
```

```
tokens <- tweets %>%
  unnest_tokens(word, clean_text) %>%
  select(screen_name, word)

tokens
```

```
## # A tibble: 172,376 × 2
##    screen_name word
##    <chr>       <chr>
##  1 b_johnson   vladimir
##  2 b_johnson   putin
##  3 b_johnson   your
##  4 b_johnson   speech
##  5 b_johnson   is
##  6 b_johnson   a
##  7 b_johnson   fraud
##  8 b_johnson   and
##  9 b_johnson   a
## 10 b_johnson   disgrace
## # … with 172,366 more rows
```

```
# english stop words
data(stop_words)

stop_words
```

```
## # A tibble: 1,149 × 2
##    word        lexicon
##    <chr>       <chr>
##   1 a           SMART
##   2 a's         SMART
##   3 able        SMART
##   4 about       SMART
##   5 above       SMART
##   6 according   SMART
##   7 accordingly SMART
##   8 across      SMART
##   9 actually    SMART
##  10 after       SMART
## # … with 1,139 more rows
```

```
tokens <- tokens %>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

```
tokens
```

```
## # A tibble: 79,630 × 2
##    screen_name word
##    <chr>       <chr>
##  1 b_johnson   vladimir
##  2 b_johnson   putin
##  3 b_johnson   speech
##  4 b_johnson   fraud
##  5 b_johnson   disgrace
##  6 b_johnson   world
##  7 b_johnson   accept
##  8 b_johnson   sham
##  9 b_johnson   referendums
## 10 b_johnson   cruel
## # … with 79,620 more rows
```

```
tokens %>%
  count(word, screen_name, sort = TRUE) %>%
  filter(screen_name == "b_johnson")
```

```
## # A tibble: 6,592 × 3
##    word      screen_name     n
##    <chr>     <chr>       <int>
##  1 rt        b_johnson     631
##  2 uk        b_johnson     518
##  3 people    b_johnson     512
##  4 country   b_johnson     302
##  5 nhs       b_johnson     270
##  6 world     b_johnson     249
##  7 support   b_johnson     231
##  8 forward   b_johnson     198
##  9 fantastic b_johnson     196
## 10 lives     b_johnson     195
## # … with 6,582 more rows
```

```
tokens %>%
  count(word, screen_name, sort = TRUE) %>%
  filter(screen_name == "liztruss")
```

```
## # A tibble: 5,830 × 3
##    word     screen_name     n
##    <chr>    <chr>       <int>
##  1 rt       liztruss     1462
##  2 trade    liztruss      926
##  3 uk       liztruss      833
##  4 ukraine  liztruss      529
##  5 amp      liztruss      425
##  6 support  liztruss      319
##  7 russia   liztruss      309
##  8 minister liztruss      239
##  9 deal     liztruss      230
## 10 putin    liztruss      229
## # … with 5,820 more rows
```

```
b_johnson_tokens <- b_johnson_tweets %>%
  mutate(
    # identify tweets with hashtags
    has_tag = str_detect(text, "#\\w+"),
    # how many at-mentions are there?
    n_at = str_count(text, "(^|\\s)@\\w+"),
    # extract first url
    url = str_extract(text, "(https?://\\S+)"),
    # remove at-mentions, tags, and urls
    clean_text = str_replace_all(text,
                                 "(^|\\s)(@|#|https?://)\\S+", " ") %>%
      str_replace_all("\\W+", " ")
  ) %>%
  unnest_tokens(word, clean_text)

liztruss_tokens <- liztruss %>%
  mutate(
    # identify tweets with hashtags
    has_tag = str_detect(text, "#\\w+"),
    # how many at-mentions are there?
    n_at = str_count(text, "(^|\\s)@\\w+"),
    # extract first url
    url = str_extract(text, "(https?://\\S+)"),
    # remove at-mentions, tags, and urls
    clean_text = str_replace_all(text,
                                 "(^|\\s)(@|#|https?://)\\S+", " ") %>%
      str_replace_all("\\W+", " ")
  ) %>%
  unnest_tokens(word, clean_text)
```

```
data(stop_words)

clean_tokens1 <-
  b_johnson_tokens %>%
    select(word) %>%
    anti_join(stop_words)

## Joining, by = "word"

clean_tokens1 %>%
  count(word, sort = TRUE) %>%
    head(10)

## # A tibble: 10 × 2
##    word          n
##    <chr>     <int>
##  1 rt          631
##  2 uk          518
##  3 people      512
##  4 country     302
##  5 nhs         270
##  6 world       249
##  7 support     231
##  8 forward     198
##  9 fantastic   196
## 10 lives       195

clean_tokens2 <-
  liztruss_tokens %>%
    select(word) %>%
    anti_join(stop_words)

## Joining, by = "word"

clean_tokens2 %>%
  count(word, sort = TRUE)%>%
    head(10)

## # A tibble: 10 × 2
##    word         n
##    <chr>    <int>
##  1 rt        1462
##  2 trade      926
##  3 uk         833
##  4 ukraine    529
##  5 amp        425
##  6 support    319
##  7 russia     309
##  8 minister   239
##  9 deal       230
## 10 putin      229
```

```
library(RVerbalExpressions)
regex <-
  rx_with_any_case() %>%
  rx_either_of("ukraine", "russia")

b_johnson_tweets %>%
  filter(str_detect(text, regex)) %>%
  select(text) %>%
  head(3)

## # A tibble: 3 × 1
##   text
##   <chr>
## 1 "Vladimir Putin your speech is a fraud and a disgrace. The world must ne
ver a…
## 2 "It's been a privilege to work with you @ZelenskyyUa, and I look forward
to s…
## 3 "I have been proud to serve as leader of the Conservative Party for the
last …

liztruss %>%
  filter(str_detect(text, regex)) %>%
  select(text) %>%
  head(3)

## # A tibble: 3 × 1
##   text
##   <chr>
## 1 "Today 143 countries stand united in condemnation of Putin’s illegal att
empts…
## 2 "The overwhelming international support for Ukraine’s struggle stands in
star…
## 3 "The appalling attacks on civilian areas in Kyiv and elsewhere are a cle
ar si…

all <- bind_rows(b_johnson_tweets, liztruss)

regex <-
  rx_with_any_case() %>%
  rx_either_of("angry ", "anger ", "happy ", "happiness")

all %>%
  filter(str_detect(text, regex)) %>%
  select(text) %>%
  head(3)

## # A tibble: 3 × 1
##   text
##   <chr>
## 1 "A very happy 75th Independence Day to the people of Pakistan. \n \nIn 7
5 yea…
```

```
## 2 "RT @10DowningStreet: Wishing Muslims here in the UK and around the worl
d a v…
## 3 "Marching for Pride in 1972 was an incredibly brave thing to do.\n\nToda
y wil…
```

THE CORPUS OBJECT

```
library(quanteda)

## Warning in .recacheSubclasses(def@className, def, env): undefined subclass
## "unpackedMatrix" of class "mMatrix"; definition not updated

## Warning in .recacheSubclasses(def@className, def, env): undefined subclass
## "unpackedMatrix" of class "replValueSp"; definition not updated

## Package version: 3.2.3
## Unicode version: 13.0
## ICU version: 69.1

## Parallel computing: 8 of 8 threads used.

## See https://quanteda.io for tutorials and examples.

# rename user id and add row_id
# all <- all %>%
#   mutate(user_id = id) %>%
#   select(-id) %>%
#   rowid_to_column("id")

# Error in rowid_to_column(., "id") :
#  could not find function "rowid_to_column"

tweets_corpus <- corpus(all,
                        docid_field = "id",
                        text_field = "text")

head(tweets_corpus)

## Corpus consisting of 6 documents and 41 docvars.
## 1575846604182982656 :
## "Vladimir Putin your speech is a fraud and a disgrace. The wo..."
##
## 1572599645519446016 :
## "I took the oath today in the House of Commons. God Save the ..."
##
## 1569006516581142528 :
## "Let us always remember the victims of terror on 9/11. Let us..."
##
## 1568258550920626176 :
## "She showed the world not just how to reign over a people, sh..."
##
## 1567941244176809984 :
```

```
## "Statement on the death of Her Majesty Queen Elizabeth II (3/..."
##
## 1567941141722533888 :
## "Statement on the death of Her Majesty Queen Elizabeth II (2/..."

## Corpus consisting of 6398 documents, showing 10 documents:
##
##                  Text Types Tokens Sentences          created_at
##  1575846604182982656    40     50         3 2022-09-30 15:54:42
##  1572599645519446016    15     17         2 2022-09-21 16:52:27
##  1569006516581142528    35     39         2 2022-09-11 18:54:38
##  1568258550920626176    36     54         2 2022-09-09 17:22:30
##  1567941244176809984    15     16         1 2022-09-08 20:21:38
##  1567941141722533888    16     16         1 2022-09-08 20:21:13
##  1567941080796078080    16     16         1 2022-09-08 20:20:59
##  1566815789281312768    41     52         3 2022-09-05 17:49:28
##   1.566761527558e+18    43     54         3 2022-09-05 14:13:51
##  1566761525293027328    36     43         1 2022-09-05 14:13:51
##               id_str
##  1575846604182982656
##  1572599645519446021
##  1569006516581142529
##  1568258550920626177
##  1567941244176809987
##  1567941141722533889
##  1567941080796078085
##  1566815789281312769
##  1566761527557996546
##  1566761525293027329
##
full_text
##        Vladimir Putin your speech is a fraud and a disgrace. The world mu
st never accept your sham referendums or your cruel and illegal attempt to co
lonise Ukraine. We stand with the people of Ukraine and will support them wit
hout flinching until their country is whole and free.
##
I took the oath today in the House of Commons. God Save the King GB https://t
.co/3FMPMRtR0N
##
Let us always remember the victims of terror on 9/11. Let us honour their mem
ory by standing strong with the United States and all our allies against thos
e who would undermine our democratic values USGB
##       She showed the world not just how to reign over a people, she showe
d the world how to give, how to love and how to serve. \n\nIt was that indomi
tability, that humour, that work ethic, that sense of history which together
made her Elizabeth the Great. https://t.co/PlACJiVb6j
#
#
Statement on the death of Her Majesty Queen Elizabeth II (3/3) https://t.co/y
HPiUfBWlH
```

```
#
#
Statement on the death of Her Majesty Queen Elizabeth II (2/3) https://t.co/e
1urjObgld
##
Statement on the death of Her Majesty Queen Elizabeth II (1/3) https://t.co/k
DN6cW8Njp
##  It's been a privilege to work with you @ZelenskyyUa, and I look forward t
o staying friends. The UK will continue to back Ukraine every step of the way
, because we know that your security is our security, and your freedom is our
freedom.\n\nSlava Ukraini UA https://t.co/knjrR9sGDj
##           Congratulations to @trussliz on her decisive win. I know she has
the right plan to tackle the cost of living crisis, unite our party and conti
nue the great work of uniting and levelling up our country. Now is the time f
or all Conservatives to get behind her 100 per cent.
##                                             I have been proud to serve as 
leader of the Conservative Party for the last three years, winning the bigges
t majority for decades, getting Brexit done, overseeing the fastest vaccine r
ollout in Europe and giving vital support to Ukraine.
##  truncated display_text_range
##      FALSE                273
##      FALSE                 67
##      FALSE                202
##      FALSE                248
##      FALSE                 62
##      FALSE                 62
##      FALSE                 62
##      FALSE                254
##      FALSE                271
##      FALSE                237
```

```
# now we can tokenize again
tweet_tokens <- quanteda::tokens(tweets_corpus,
                                 remove_punct = TRUE,   # removes punctuation
s
                                 remove_numbers = TRUE, # removes numbers
                                 remove_symbols = TRUE, # removes symbols (al
so: emojis)
                                 remove_url = TRUE)     # removes urls

head(tweet_tokens, 6)
```

```
## Tokens consisting of 6 documents and 41 docvars.
## 1575846604182982656 :
##  [1] "Vladimir" "Putin"    "your"     "speech"   "is"       "a"
##  [7] "fraud"    "and"      "a"        "disgrace" "The"      "world"
## [ ... and 35 more ]
##
## 1572599645519446016 :
```

```
##  [1] "I"       "took"    "the"     "oath"    "today"   "in"      "the"    
##  [8] "House"   "of"      "Commons" "God"     "Save"   
## [ ... and 3 more ]
## 
## 1569006516581142528 :
##  [1] "Let"      "us"       "always"   "remember" "the"      "victims" 
##  [7] "of"       "terror"   "on"       "Let"      "us"       "honour"  
## [ ... and 23 more ]
## 
## 1568258550920626176 :
##  [1] "She"    "showed" "the"    "world"  "not"    "just"   "how"    "to"    
##  [9] "reign"  "over"   "a"      "people"
## [ ... and 34 more ]
## 
## 1567941244176809984 :
##  [1] "Statement" "on"        "the"       "death"     "of"        "Her"      
##  [7] "Majesty"   "Queen"     "Elizabeth" "II"       
## 
## 1567941141722533888 :
##  [1] "Statement" "on"        "the"       "death"     "of"        "Her"      
##  [7] "Majesty"   "Queen"     "Elizabeth" "II"

# lower key tokens
tweet_tokens_lk <- tweet_tokens %>% tokens_tolower()

# remove stop words
tweet_tokens_nosw <- tweet_tokens_lk %>%
  tokens_remove(stopwords("english"))

head(tweet_tokens_nosw,6)

## Tokens consisting of 6 documents and 41 docvars.
## 1575846604182982656 :
##  [1] "vladimir"    "putin"       "speech"      "fraud"       "disgrace"   
##  [6] "world"       "must"        "never"       "accept"      "sham"       
## [11] "referendums" "cruel"      
## [ ... and 13 more ]
## 
## 1572599645519446016 :
## [1] "took"    "oath"    "today"   "house"   "commons" "god"     "save"   
## [8] "king"    "GB"     
## 
## 1569006516581142528 :
##  [1] "let"      "us"       "always"   "remember" "victims"  "terror"  
##  [7] "let"      "us"       "honour"   "memory"   "standing" "strong"  
## [ ... and 8 more ]
## 
## 1568258550920626176 :
##  [1] "showed"         "world"          "just"           "reign"         
##  [5] "people"         "showed"         "world"          "give"
##  [9] "love"           "serve"          "indomitability" "humour"
## [ ... and 8 more ]
##
## 1567941244176809984 :
## [1] "statement" "death"     "majesty"   "queen"     "elizabeth" "ii"
##
## 1567941141722533888 :
## [1] "statement" "death"     "majesty"   "queen"     "elizabeth" "ii"
```

```
# STEMMING

tweet_tokens_nosw %>% tokens_wordstem() %>%
    head()
```

```
## Tokens consisting of 6 documents and 41 docvars.
## 1575846604182982656 :
##  [1] "vladimir"   "putin"      "speech"     "fraud"      "disgrac"
##  [6] "world"      "must"       "never"      "accept"     "sham"
## [11] "referendum" "cruel"
## [ ... and 13 more ]
##
## 1572599645519446016 :
## [1] "took"   "oath"   "today"  "hous"   "common" "god"    "save"   "king"
## [9] "GB"
##
## 1569006516581142528 :
##  [1] "let"    "us"     "alway"  "rememb" "victim" "terror" "let"    "us"
##  [9] "honour" "memori" "stand"  "strong"
## [ ... and 8 more ]
##
## 1568258550920626176 :
##  [1] "show"    "world"   "just"    "reign"   "peopl"   "show"    "world"
##  [8] "give"    "love"    "serv"    "indomit" "humour"
## [ ... and 8 more ]
##
## 1567941244176809984 :
## [1] "statement" "death"     "majesti"   "queen"     "elizabeth" "ii"
##
## 1567941141722533888 :
## [1] "statement" "death"     "majesti"   "queen"     "elizabeth" "ii"
```

```
tweets <- b_johnson_tweets
corpus <- corpus(tweets,
                 docid_field = "id",
                 text_field = "text")

library(udpipe)

ud_model_en <- udpipe_download_model(language = "english")
```

```
## Downloading udpipe model from https://raw.githubusercontent.com/jwijffels/
udpipe.models.ud.2.5/master/inst/udpipe-ud-2.5-191206/english-ewt-ud-2.5-1912
06.udpipe to C:/Users/KayaBennur/OneDrive - Makersite GmbH/Desktop/english-ew
t-ud-2.5-191206.udpipe

##  - This model has been trained on version 2.5 of data from https://univers
aldependencies.org

##  - The model is distributed under the CC-BY-SA-NC license: https://creativ
ecommons.org/licenses/by-nc-sa/4.0

##  - Visit https://github.com/jwijffels/udpipe.models.ud.2.5 for model licen
se details.

##  - For a list of all models and their licenses (most models you can downlo
ad with this package have either a CC-BY-SA or a CC-BY-SA-NC license) read th
e documentation at ?udpipe_download_model. For building your own models: visi
t the documentation by typing vignette('udpipe-train', package = 'udpipe')

## Downloading finished, model stored at 'C:/Users/KayaBennur/OneDrive - Make
rsite GmbH/Desktop/english-ewt-ud-2.5-191206.udpipe'
```

```
udpipe_lemmas <- udpipe(tweets$text, object = ud_model_en)

# hack in screen_name
udpipe_lemmas <-
  udpipe_lemmas %>%
  mutate(
    screen_name = if_else(doc_id %in% str_c("doc", 1:1200), "BorisJohnson", "
trussliz")
  )

docs <- udpipe_lemmas %>%
  group_by(doc_id) %>%
  summarize(text = paste(lemma, collapse = " "), screen_name = screen_name) %
>%
  distinct()
```

```
## `summarise()` has grouped output by 'doc_id'. You can override using the
## `.groups` argument.
```

```
lemma_tokens_nosw <- docs %>%
  corpus() %>%
  quanteda::tokens(remove_punct = TRUE,
                   remove_numbers = TRUE,
                   remove_symbols = TRUE,
                   remove_separators = TRUE,
                   remove_url = TRUE) %>%
  tokens_tolower() %>%
  tokens_remove(stopwords("english"))
```

```
library(quanteda.textstats)

## Warning in .recacheSubclasses(def@className, def, env): undefined subclass
## "unpackedMatrix" of class "mMatrix"; definition not updated

## Warning in .recacheSubclasses(def@className, def, env): undefined subclass
## "unpackedMatrix" of class "replValueSp"; definition not updated

lemma_tokens_nosw %>%
  textstat_collocations() %>%
  arrange(desc(count)) %>%
  head(10)

##            collocation count count_nested length    lambda         z
## 982 @10 downingstreet   312            0      2 17.955336  8.974057
## 62             rt @10   309            0      2  9.524099 17.623140
## 6          get brexit   132            0      2  6.362835 28.739223
## 1          watch live   119            0      2  7.528812 34.696541
## 35     prime minister   107            0      2 11.260707 20.695475
## 48   rt @conservative    96            0      2  7.089366 18.580241
## 2           save life    91            0      2  6.354177 33.337274
## 7              let us    88            0      2  8.431877 28.672532
## 8           across uk    78            0      2  4.292237 28.463400
## 12       look forward    77            0      2  7.450723 27.180762

kwic(lemma_tokens_nosw, pattern = phrase("brexit"))

## Keyword-in-context with 159 matches.
##    [doc10, 14]                    win biggest majority decade get | brexit
|
##  [doc1765, 21]                amp ensure northern ireland benefit | brexit
|
##   [doc2252, 9]                     step downing street pledge get | brexit
|
##  [doc2787, 12]                     dexeu hard work dedication get | brexit
|
##  [doc2791, 15]                 come together now make opportunity | brexit
|
##   [doc2810, 1]                                                    | brexit
|
##   [doc2810, 8]                 holiday answer top search question | brexit
|

##  oversee fastest vaccine rollout europe
##  just like every part united
##  unite level country year extraordinary
##  deliver
##  bring let us unleash potential
##  affect holiday answer top search
##  happen leave eu friday
##  deliver change people vote go


# preserve in BOW approach
toks_comp <- tokens_compound(lemma_tokens_nosw,
                             pattern = phrase("get brexit"))

kwic(toks_comp, pattern = phrase("get brexit"))

## Keyword-in-context with 0 matches.

## Keyword-in-context with 0 matches.
kwic(toks_comp, pattern = phrase("get_brexit"))

## Keyword-in-context with 132 matches.
##    [doc10, 13]                   year win biggest majority decade | get_br
exit
##   [doc2252, 8]                   stand step downing street pledge | get_br
exit
##  [doc2787, 11]                everyone dexeu hard work dedication | get_br
exit
##   [doc2872, 6]                          go level unite country go | get_br
exit
##   [doc2875, 2]                                                 go | get_br
exit
##  [doc2888, 11]                          pass mean one step closer | get_br
exit
##  [doc2889, 12]                         just vote pass brexit deal | get_br
exit

##  | oversee fastest vaccine rollout europe
##  | unite level country year extraordinary
##  | deliver
##  | deliver change people vote go
##  | unite fantastic country
##  | GB
##  | 31st january make sure
##  | deliver key priorities british people
##  | deliver change people vote go
##  | unite fantastic country
##  |


#N-GRAMS
tokens_ngrams <- lemma_tokens_nosw %>%
  tokens_ngrams(n = 2:5)

tokens_ngrams

## Tokens consisting of 3,199 documents and 1 docvar.
## doc1 :
```

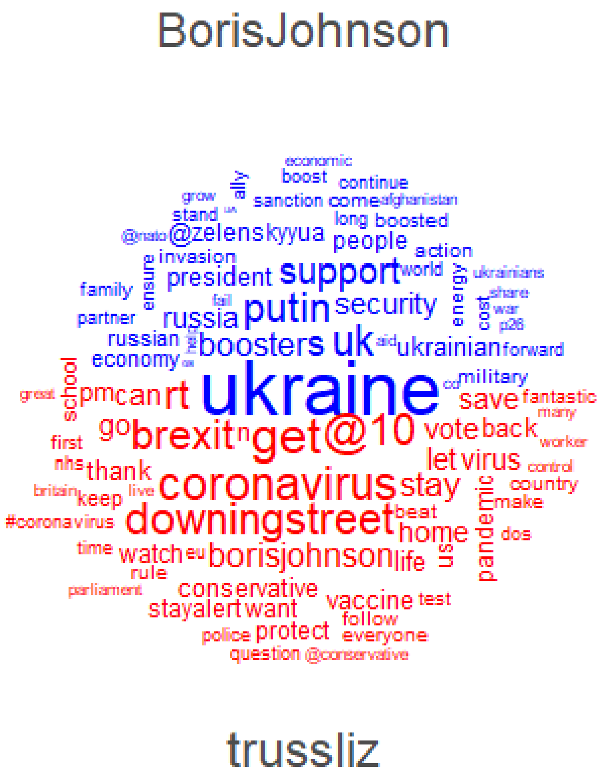
```
##  [1] "vladimy_putin"    "putin_speech"     "speech_fraud"     "fraud_disgr
ace"
##  [5] "disgrace_world"   "world_must"       "must_never"       "never_accep
t"
##  [9] "accept_sham"      "sham_referendum"  "referendum_cruel" "cruel_illeg
al"
## [ ... and 78 more ]
##
## doc10 :
##  [1] "proud_serve"         "serve_leader"        "leader_conservative"
##  [4] "conservative_party"  "party_last"          "last_three"
##  [7] "three_year"          "year_win"            "win_biggest"
## [10] "biggest_majority"    "majority_decade"     "decade_get"
## [ ... and 70 more ]
##
## doc100 :
##  [1] "rt_@trussliz"        "@trussliz_today"     "today_uk"
##  [4] "uk_ambassador"       "ambassador_@nato"    "@nato_davidquarrey"
##  [7] "davidquarrey_sign"   "sign_accession"      "accession_protocols"
## [10] "protocols_sweden"    "sweden_amp"          "amp_finland"
## [ ... and 42 more ]
##
## doc1000 :
##  [1] "sir_david"                    "david_amess"
##  [3] "amess_mp"                     "mp_1952-2021"
##  [5] "sir_david_amess"              "david_amess_mp"
##  [7] "amess_mp_1952-2021"           "sir_david_amess_mp"
##  [9] "david_amess_mp_1952-2021"     "sir_david_amess_mp_1952-2021"
##
## doc1001 :
##  [1] "heart_full"     "full_shock"     "shock_sadness"  "sadness_death"
##  [5] "death_sir"      "sir_david"      "david_amess"    "amess_mp"
##  [9] "mp_one"         "one_kindest"    "kindest_nicest" "nicest_gentle"
## [ ... and 38 more ]
##
## doc1002 :
##  [1] "first_industrial"      "industrial_revolution" "revolution_powere"
##  [4] "powere_steam"          "steam_next"            "next_industrial"
##  [7] "industrial_revolution" "revolution_green"      "green_create"
## [10] "create_new"            "new_job"               "job_sustainable"
## [ ... and 74 more ]
##
## [ reached max_ndoc ... 3,193 more documents ]
```

```
dfm <- dfm(lemma_tokens_nosw)
```

```
dfm
```

```
## Document-feature matrix of: 3,199 documents, 6,684 features (99.76% sparse
) and 1 docvar.
```

```
##          features
## docs      vladimy putin speech fraud disgrace world must never accept sham
##   doc1          1     1      1     1        1     1    1     1      1    1
##   doc10         0     0      0     0        0     0    0     0      0    0
##   doc100        0     0      0     0        0     0    0     0      0    0
##   doc1000       0     0      0     0        0     0    0     0      0    0
##   doc1001       0     0      0     0        0     0    0     0      0    0
##   doc1002       0     0      0     0        0     0    0     0      0    0
## [ reached max_ndoc ... 3,193 more documents, reached max_nfeat ... 6,674 m
ore features ]

topfeatures(dfm, groups = screen_name)

## $BorisJohnson
##      uk       s  people      rt   today ukraine     get support     can 
work 
##     275     273     241     210     196     189     177     160     150 
149 
## 
## $trussliz
##            rt           get           can         today        people 
##           421           416           290           286           275 
##             s            uk           @10 downingstreet          work 
##           275           246           240           240           237

dfm_mentions <- dfm_select(dfm, "@*")

topfeatures(dfm_mentions, groups = screen_name)

## $BorisJohnson
##           @10  @zelenskyyua @conservative         @nato   @rishisunak 
##            72            41            30            22            22 
##        @cop26     @trussliz   @sajidjavid   @pritipatel @nadhimzahawi 
##            18            12            12            10            10 
## 
## $trussliz
##           @10 @conservative   @rishisunak @borisjohnson      @england 
##           240            73            33            16            13 
##    @dhscgovuk        @cop26   @pritipatel           @g7            @g 
##            10             8             8             7             7

# COMPARE
# grouped DFM
tweet_dfm_grouped <- dfm_group(dfm, groups = screen_name)

# wordcloud
library(quanteda.textplots)
textplot_wordcloud(tweet_dfm_grouped,
                   max_words = 100,
                   comparison = TRUE,
                   color = c("blue", "red"))
```

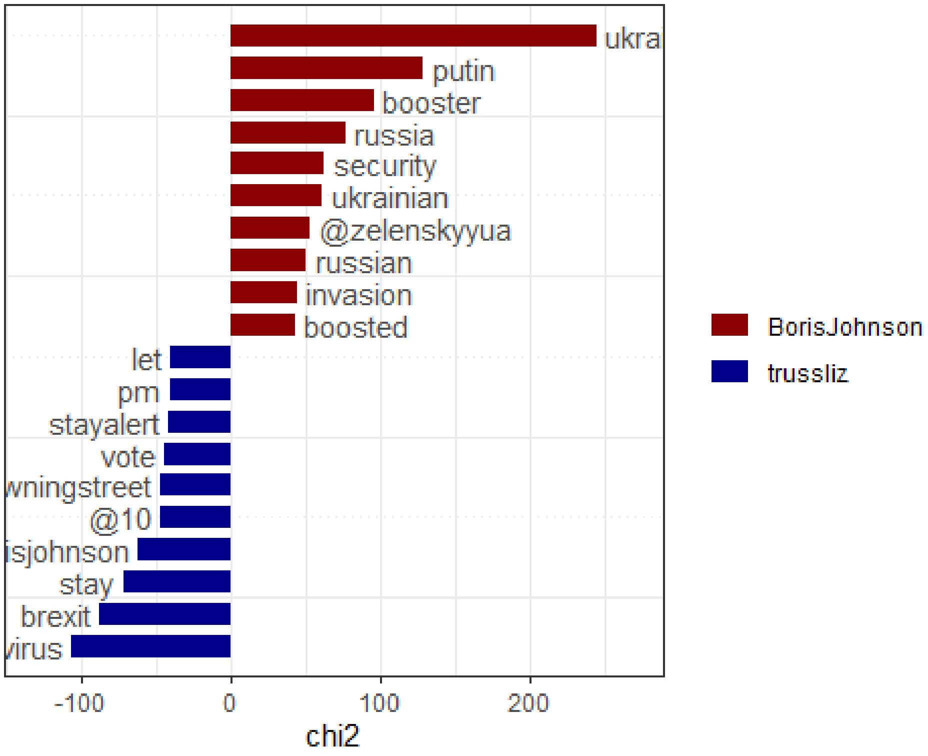


```
# LEXICAL DIVERSITY

# lexical diversity
# quantifies how lexically rich a text is,
# e.g. Type-Token Ratio (TTR), which divides the amount
# of unique tokens through all tokens within a corpus.
# it is useful, for instance, for analysing speakers’ or
# writers’ linguistic skills, or the complexity of ideas expressed
# in documents.
lexdiv <- textstat_lexdiv(tweet_dfm_grouped)
lexdiv

##        document       TTR
## 1 BorisJohnson 0.1923128
## 2     trussliz 0.1464749

# KEYNESS
# keyness: quantifies the uniqueness of words for a corpus as
# compared to another corpus (using chi-squared statistics)
textstat_keyness(tweet_dfm_grouped, target = "BorisJohnson") %>%
  as_tibble()

## # A tibble: 6,684 × 5
##    feature       chi2        p n_target n_reference
##    <chr>        <dbl>    <dbl>    <dbl>       <dbl>
##  1 ukraine      244.  0             189           5
##  2 putin        129.  0              96           1
##  3 booster       95.6 0              72            1
##  4 russia        76.3 0              58            1
##  5 security      62.3 3.00e-15       73           14
##  6 ukrainian     61.2 5.11e-15       47            1
##  7 @zelenskyyua  53.0 3.35e-13       41            1
##  8 russian       49.9 1.61e-12       41            2
##  9 invasion      43.9 3.44e-11       32            0
## 10 boosted       42.5 6.93e-11       31            0
## # … with 6,674 more rows
```

```
textstat_keyness(tweet_dfm_grouped, target = "trussliz") %>%
  as_tibble()
```

```
## # A tibble: 6,684 × 5
##    feature         chi2        p n_target n_reference
##    <chr>          <dbl>    <dbl>    <dbl>       <dbl>
##  1 coronavirus   106.  0              156           3
##  2 brexit         87.2 0              150           9
##  3 stay           71.8 0              109           3
##  4 borisjohnson   61.4 4.55e-15       111           8
##  5 @10            46.9 7.44e-12       240          72
##  6 downingstreet  46.9 7.44e-12       240          72
##  7 vote           43.9 3.52e-11       103          14
##  8 stayalert      41.6 1.12e-10        57           0
##  9 pm             40.9 1.57e-10        82           8
## 10 let            40.0 2.53e-10       120          23
## # … with 6,674 more rows
```

```
# plot
textstat_keyness(tweet_dfm_grouped, target = "BorisJohnson") %>%
  textplot_keyness(n = 10, color = c("darkred", "darkblue"))
```


ukra
putin
booster
russia
security
ukrainian
@zelenskyyua
russian
invasion
boosted
let
pm
stayalert
vote
wningstreet
@10
isjohnson
stay
brexit
virus
-100
0
100
200
chi2
BorisJohnson
trussliz