Cleantweets

Noblezada, Camasa, Cabia

2024-12-13

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(tidytext)
library(dplyr)
library(stringr)
library(ggplot2)
library(sentimentr)
```

```r
# Load the dataset
tweetsDF <- read_csv("/cloud/project/ProjectDS/tweetsDF.csv")
```

```
## New names:
## Rows: 58086 Columns: 7
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (4): screenName, text, statusSource, tweetSource dbl (1): ...1 dttm (2):
## created, Created_At_Round
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```r
# Clean the tweet text
tweetsDF <- tweetsDF %>%
  mutate(
    text = text %>%
      iconv(from = "UTF-8", to = "ASCII//TRANSLIT", sub = "") %>% # Remove non-ASCII characters
      tolower() %>% # Convert to lowercase
      str_remove_all("https\\S+") %>% # Remove URLs
      str_remove_all("[#\\n]") %>% # Remove hashtags and newlines
      str_remove_all("[@?]\\S+") %>% # Remove mentions
      str_remove_all("\\?") %>% # Remove question marks
      str_remove_all("\\b\\d{2}\\.\\d{2}\\.\\d{4}\\b") %>% # Remove dates in dd.mm.yyyy format
      str_remove_all("<a href=httptwitter.comdownloadiphone rel=nofollow>twitter for iphone<a>") %>% #
      str_remove_all("<a href=([^>]*?) rel=nofollow>([^<]*?)<a>") %>%
      str_remove_all("<a href=httptwitter.comdownloadandroid rel=nofollow>twitter for android<a>") %>%
      str_remove_all("<a href= rel=nofollow>twitter web app<a>") %>%
      str_remove_all("30102022") %>% # Remove specific date
      str_squish() # Remove extra whitespace
  )

# Function to create chunks of data
create_chunks <- function(df, start_row, end_row) {
  return(df[start_row:end_row, ])
}

# Define chunk size
start_row <- 1
end_row <- 1000

# Extract chunk of data
chunk_data <- create_chunks(tweetsDF, start_row, end_row)

# Print cleaned dataset
print(tweetsDF)
```

```
## # A tibble: 58,086 x 7
##     ...1 screenName   text  created             statusSource Created_At_Round
##    <dbl> <chr>        <chr> <dttm>              <chr>        <dttm>
##  1     1 whourj31     a so~ 2022-10-30 23:59:43 "<a href=\"~ 2022-10-31 00:00:00
##  2     2 nnainot      nah ~ 2022-10-30 23:59:32 "<a href=\"~ 2022-10-31 00:00:00
##  3     3 febry_sri_M  pray~ 2022-10-30 23:59:31 "<a href=\"~ 2022-10-31 00:00:00
##  4     4 telehuntwat~ tran~ 2022-10-30 23:59:28 "<a href=\"~ 2022-10-31 00:00:00
##  5     5 Typing0824   the ~ 2022-10-30 23:59:20 "<a href=\"~ 2022-10-31 00:00:00
##  6     6 niccijsmith  what~ 2022-10-30 23:59:04 "<a href=\"~ 2022-10-31 00:00:00
##  7     7 502SPIDEY    can'~ 2022-10-30 23:58:56 "<a href=\"~ 2022-10-31 00:00:00
##  8     8 maeannesala~ pray~ 2022-10-30 23:58:45 "<a href=\"~ 2022-10-31 00:00:00
##  9     9 bigvirtue1   bigv~ 2022-10-30 23:58:37 "<a href=\"~ 2022-10-31 00:00:00
## 10    10 ashxxy       ther~ 2022-10-30 23:58:31 "<a href=\"~ 2022-10-31 00:00:00
## # i 58,076 more rows
## # i 1 more variable: tweetSource <chr>
```