

## Source code

```
# Load packages for web scraping from times of India website and create chatbot
library(rvest)
library(dplyr)

##
## 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(NLP)
library(tm)
library(DSL)
library(tm.plugin.dc)

# Specify the URL of the webpage you want to scrape
url <- "https://timesofindia.indiatimes.com/india/tamil-nadu"

# Read the HTML content of the webpage
webpage <- read_html(url)

# Extract the headlines from the webpage
scrape_headlines <- function(url) {
  # Read HTML content
  webpage <- read_html(url)
  # Extract headlines
  headlines <- webpage %>%
    html_nodes(".w_tle") %>%
    html_text()
  # Clean headlines
  cleaned_headlines <- str_trim(headlines)

  return(cleaned_headlines)
}
headlines <- scrape_headlines(url)
# Save headlines to CSV
write.csv(headlines, file = "times_of_india_headlines.csv", row.names =
FALSE)
# Print the headlines
head(headlines)
```

```
## [1] "NDA govt is not in unison with people: Kamal Haasan"
## [2] "Dr. Anita Ratnam's Arangham Dance Theatre presented 'Naachiyar Next'"
## [3] "Rashmika Mandanna & Vijay Deverakonda caught holidaying together
again?"
## [4] "Actress Dharshana Ashokan quits Tamil daily soap 'Kanaa'"
## [5] "Tamil Nadu: Parties will reap what they sow"
## [6] "Vaiko: Congress, Modi betrayed Tamil Nadu"
```

*# Preprocessing steps (you can add more preprocessing steps here)*

*# Example: Convert to lowercase*

```
headlines <- tolower(headlines)
```

```
head(headlines)
```

```
## [1] "nda govt is not in unison with people: kamal haasan"
## [2] "dr. anita ratnam's arangham dance theatre presented 'naachiyar next'"
## [3] "rashmika mandanna & vijay deverakonda caught holidaying together
again?"
## [4] "actress dharshana ashokan quits tamil daily soap 'kanaa'"
## [5] "tamil nadu: parties will reap what they sow"
## [6] "vaiko: congress, modi betrayed tamil nadu"
```

*# Example: Remove punctuation*

```
headlines <- gsub("[[:punct:]]", "", headlines)
```

```
head(headlines)
```

```
## [1] "nda govt is not in unison with people kamal haasan"
## [2] "dr anita ratnams arangham dance theatre presented naachiyar next"
## [3] "rashmika mandanna vijay deverakonda caught holidaying together
again"
## [4] "actress dharshana ashokan quits tamil daily soap kanaa"
## [5] "tamil nadu parties will reap what they sow"
## [6] "vaiko congress modi betrayed tamil nadu"
```

*# Example: Remove numbers*

```
headlines <- gsub("\\d+", "", headlines)
```

```
head(headlines)
```

```
## [1] "nda govt is not in unison with people kamal haasan"
## [2] "dr anita ratnams arangham dance theatre presented naachiyar next"
## [3] "rashmika mandanna vijay deverakonda caught holidaying together
again"
## [4] "actress dharshana ashokan quits tamil daily soap kanaa"
## [5] "tamil nadu parties will reap what they sow"
## [6] "vaiko congress modi betrayed tamil nadu"
```

*# Example: Remove extra whitespaces*

```
headlines <- gsub("\\s+", " ", headlines)
```

```
head(headlines)
```

```
## [1] "nda govt is not in unison with people kamal haasan"
## [2] "dr anita ratnams arangham dance theatre presented naachiyar next"
## [3] "rashmika mandanna vijay deverakonda caught holidaying together again"
```

```

## [4] "actress dharshana ashokan quits tamil daily soap kanaa"
## [5] "tamil nadu parties will reap what they sow"
## [6] "vaiko congress modi betrayed tamil nadu"

# Save preprocessed headlines to CSV

write.csv(headlines, file = "preprocessed_times_of_india_headlines.csv",
row.names = FALSE)

# Load the data

news_data <- read.csv(file="preprocessed_times_of_india_headlines.csv")

# Define a function to search for news based on a keyword
search_news <- function(keyword) {
  result <- news_data[grep(keyword, news_data, ignore.case = TRUE) ]
  if (nrow(result) == 0) {
    return("No matching news found.")
  } else {
    return(result)
  }
}

# Interactive chat function
chatbot <- function() {
  cat("Welcome to the Times of India News Chatbot!\n")
  repeat {
    cat("Enter a keyword to search for news (type 'exit' to quit): ")
    keyword <- tolower(readline())
    if (keyword == "exit") {
      cat("Goodbye!")
      break
    }
  }
}

```

```
    } else {  
        result <- search_news(keyword)  
        cat("\n")  
        print(result)  
        cat("\n")  
    }  
}  
}
```

```
# Run the chatbot
```

```
chatbot()
```

```
##Welcome to the Times of India News Chatbot!
```

```
##Enter a keyword to search for news (type 'exit' to quit):
```

```
India
```

```
##nda govt is not in unison with people kamal haasan
```

```
##dr anita ratnams arangham dance theatre presented naachiyar next
```

```
##rashmika mandanna vijay deverakonda caught holidaying together again
```

```
##actress dharshana ashokan quits tamil daily soap kanaa
```

```
##tamil nadu parties will reap what they sow
```

```
##vaiko congress modi betrayed tamil nadu
```

```
##fisherman dies as boat sinks in bay of Bengal
```

```
##lok sabha election its the last chance to save democracy chidambaram says
```

```
##nda govt is not in unison with people kamal haasan
```

```
##tamil nadu parties will reap what they sow
```

```
##Enter a keyword to search for news (type 'exit' to quit):
```

```
exit
```

```
##Goodbye!
```