WEBSCRAPING AND PREPROCESSING

23CSEG28

```
library(rvest)
library(stringr)

url <- "https://en.wikipedia.org/wiki/India"

scrape_headlines <- function(url) {
  contents <- read_html(url)
  headlines <- contents %>%
    html_nodes(".mw-headline") %>%
    html_text()
  return(headlines)
}

headlines <- scrape_headlines(url)

# Save headlines to CSV
write.csv(headlines, file = "india_wikipedia_headlines.csv", row.names = FALS
E)
print(headlines)

##  [1] "Etymology"
##  [2] "History"
##  [3] "Ancient India"
##  [4] "Medieval India"
##  [5] "Early modern India"
##  [6] "Modern India"
##  [7] "Geography"
##  [8] "Biodiversity"
##  [9] "Politics and government"
## [10] "Politics"
## [11] "Government"
## [12] "Administrative divisions"
## [13] "States"
## [14] "Union territories"
## [15] "Foreign and strategic relations"
## [16] "Military"
## [17] "Economy"
## [18] "Industries"
## [19] "Energy"
## [20] "Socio-economic challenges"
## [21] "Demographics, languages and religion"
## [22] "Culture"
## [23] "Visual art"
## [24] "Architecture"
## [25] "Literature"
## [26] "Performing arts and media"
## [27] "Society"
## [28] "Education"
## [29] "Clothing"
## [30] "Cuisine"
## [31] "Sports and recreation"
## [32] "See also"
## [33] "Notes"
## [34] "References"
## [35] "Bibliography"
## [36] "External links"

# Preprocessing
# Convert to lowercase
lower_content <- tolower(headlines)
head(lower_content)

## [1] "etymology"          "history"            "ancient india"
## [4] "medieval india"     "early modern india" "modern india"

# Remove punctuation
rem_punct <- gsub("[[:punct:]]", "", lower_content)
head(rem_punct)

## [1] "etymology"          "history"            "ancient india"
## [4] "medieval india"     "early modern india" "modern india"

# Remove numbers
rem_num <- gsub("\\d+", "", rem_punct)
head(rem_num)

## [1] "etymology"          "history"            "ancient india"
## [4] "medieval india"     "early modern india" "modern india"

# Remove extra whitespaces
cleaned_headlines <- gsub("\\s+", " ", rem_num)

head(cleaned_headlines)

## [1] "etymology"          "history"            "ancient india"
## [4] "medieval india"     "early modern india" "modern india"
```