AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Science and Technology

Assignment Cover Sheet

Assignment Title:	Final Term project		
Assignment No:		Date of Submission:	25 May, 2025
Course Title:	Introduction to Data Science		
Course Code:	01812	Section:	A
Semester:	Spring 24-25	Course Teacher:	Abdus Salam

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/NO: **05**

No	Name	ID	Program	Signature
1	Mithi Zaman	21-44753-1	BSc.CSE	
2	Zannatul Ferdows Wafi	21-45924-3	BSc.CSE	
3	Tasfi Islam	21-45504-3	BSc.CSE	
4	Nusrat Jannat	21-45795-3	BSc.CSE	

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Part-1: Text Processing:

- Selected **NPR (National Public Radio)** news portal for extracting news information.
- Collected 500 news articles (100 each from Technology, Science, World, Business, Health).
- Cleaned text by removing emojis, contractions, and special characters.
- Corrected spelling errors and split text into words (tokenization).
- Removed common words and reduced words to base forms.

Part-2: Topic Modeling:

- Created a word-frequency table (DTM) from processed text.
- Used LDA to discover 5 hidden topics in the news data.
- Identified top 10 keywords per topic

PART-1

Web Scraping:

This code scrapes 100 news articles from each of five NPR categories: technology, science, world, business, and health. This code loops through the first five pages of each category, extracts the article titles, descriptions, and publication dates using HTML node selectors, and stores them in vectors. After that, these vectors are combined to create a data frame for every category. Once every article has been collected, the full dataset is saved as a CSV file.

Input:

```
categories <- c("technology", "science", "world", "business", "health")
all_news <- data.frame()

for (category in categories) {
  base_url <- paste0("https://www.npr.org/sections/", category, "/")
  titles_cat <- c(); descriptions_cat <- c(); dates_cat <- c()

  for (i in 1:5) {
    full_url <- paste0(base_url, "?page=", i)
    page <- read_html(GET(full_url))

    titles <- page %>% html_nodes(".item-info h2.title a") %>% html_text(trim = TRUE)
    descriptions <- page %>% html_nodes(".item-info p.teaser a") %>% html_text(trim = TRUE)
    dates <- page %>% html_nodes(".item-info .teaser time") %>% html_attr("datetime")

    titles_cat <- c(titles_cat, titles)
    descriptions_cat <- c(descriptions_cat, descriptions)
    dates_cat <- c(dates_cat, dates)
    Sys.sleep(1)
  }

  category_data <- data.frame(
    Title = titles_cat,
    Description = descriptions_cat,
    Date = dates_cat,
    Category = rep(category, length(titles_cat)),
    stringsAsFactors = FALSE
  )

  all_news <- bind_rows(all_news, head(category_data, 100))
  cat("Scraped 100 articles for:", category, "\n")
}
write.csv(all_news, "/Users/mithizaman/Documents/12/Introduction to data science/code/final.csv", row.names = FALSE)
```

OUTPUT:

Filter

	Title	Description		Date	Category
1	Trump seeks to boost nuclear industry and overhaul s...	May 23, 2025	A series of executive orders aims to ...	2025-05-23	technology
2	Inside a Drone Factory in Ukraine	May 23, 2025	Throughout the more than three yea...	2025-05-23	technology
3	OpenAI forges deal with iPhone designer Jony Ive to ...	May 22, 2025	The $6.5 billion deal brings together...	2025-05-22	technology
4	Raising ethics questions, top Trump meme coin inves...	May 22, 2025	President Trump is hosting an exclu...	2025-05-22	technology
5	A Newark air traffic controller on the moment system...	May 21, 2025	Federal regulators are now limiting t...	2025-05-21	technology
6	The great battery race: China and the U.S. compete ov...	May 21, 2025	The car you drive years in the future ...	2025-05-21	technology
7	Musk to slow down political spending: 'I think I've do...	May 20, 2025	The billionaire executive was Trump'...	2025-05-20	technology
8	Verizon ends DEI policies to get FCC's blessing for its ...	May 19, 2025	It's the latest big company to back a...	2025-05-19	technology
9	Labor watchdog opens investigation into DOGE whistl...	May 16, 2025	DOGE employees demanded the hig...	2025-05-16	technology
10	A tale of murder, artificial intelligence, & forgiveness	May 14, 2025	Should AI give you a voice? Even whe...	2025-05-14	technology
11	Is AI the future of America's foreign policy? Some exp...	May 12, 2025	Large language models like ChatGPT...	2025-05-12	technology
12	Elizabeth Holmes' partner raises millions for new biot...	May 10, 2025	The incarcerated former Silicon Valle...	2025-05-10	technology
13	Google will pay Texas $1.4B to settle claims over user...	May 9, 2025	The agreement settles several claims ...	2025-05-09	technology
14	Trump tightens control of independent agency overse...	May 9, 2025	NPR has learned that rules must now ...	2025-05-09	technology
15	How tech companies could shrink AI's climate footprint	May 9, 2025	Google, Microsoft and Meta have all p...	2025-05-09	technology
16	Here's why Bill Gates is giving away most of his remai...	May 8, 2025	After decades of philanthropy followi...	2025-05-08	technology
17	From apps to gadgets, 'Second Life' considers how te...	May 8, 2025	When Amanda Hess learned her unbo...	2025-05-08	technology
18	Economists warn Trump's research cuts could have di...	May 8, 2025	President Trump has proposed slashi...	2025-05-08	technology

Text Processing Steps:

Text Cleaning (Contractions, Emojis, and Formatting):

This code performs text cleaning and preprocessing on the scraped news data. It starts by expanding contractions (like "can't" to "cannot") using a custom function and a list of common contractions. Next, it replaces emojis and emoticons in the text with placeholder words to standardize them. The main cleaning function converts text to lowercase, removes numbers, punctuation, and any non-letter characters, and also strips away any remaining HTML tags. It then removes extra spaces for neatness. This cleaning process is applied separately to the Title and Description columns of the dataset, resulting in two new columns with cleaned text: Cleaned_Title and Cleaned_Description.

INPUT:

```
contractions <- c(
  "don't" = "do not", "doesn't" = "does not", "didn't" = "did not",
  "can't" = "cannot", "won't" = "will not", "wouldn't" = "would not",
  "isn't" = "is not", "aren't" = "are not", "wasn't" = "was not",
  "weren't" = "were not", "haven't" = "have not", "hasn't" = "has not",
  "hadn't" = "had not", "i'm" = "i am", "we're" = "we are", "they're" = "they are",
  "you're" = "you are", "it's" = "it is", "that's" = "that is", "there's" = "there is",
  "what's" = "what is", "who's" = "who is", "she's" = "she is", "he's" = "he is",
  "i've" = "i have", "you've" = "you have", "they've" = "they have",
  "we've" = "we have", "i'll" = "i will", "you'll" = "you will", "they'll" = "they will",
  "we'll" = "we will", "shouldn't" = "should not", "couldn't" = "could not",
  "mustn't" = "must not", "needn't" = "need not", "mightn't" = "might not",
  "shan't" = "shall not", "let's" = "let us", "who'll" = "who will", "how's" = "how is"
)

expand_contractions <- function(text) {
  for (con in names(contractions)) {
    text <- gsub(paste0("\\b", con, "\\b"), contractions[con], text, ignore.case = TRUE)
  }
  return(text)
}

handle_emojis_emoticons <- function(text) {
  emoji_pattern <- "[\\U0001F600-\\U0001F64F\\U0001F300-\\U0001F5FF\\U0001F680-\\U0001F6FF\\U0001F700-\\U0001F77F\\U0001F900-\\U0001F9FF]+"
  emoticon_pattern <- ":\\)|:-\\)|:\\(|:-\\(|;\\)|;-\\)|:D|:-D|XD|<3"
  text <- str_replace_all(text, emoji_pattern, " emoji ")
  text <- str_replace_all(text, emoticon_pattern, " emoticon ")
  return(text)
}


clean_text <- function(text) {
  text <- handle_emojis_emoticons(text)
  text <- tolower(text)
  text <- expand_contractions(text)
  text <- gsub("[0-9]+", "", text)
  text <- gsub("[[:punct:]]", "", text)
  text <- gsub("[^a-zA-Z\\s]", " ", text)
  text <- read_html(paste0("<body>", text, "</body>")) %>% html_text(trim = TRUE)
  text <- str_squish(text)
  return(text)
}

df <- read.csv("/Users/mithizaman/Documents/12/Introduction to data science/code/final/npr_Updated_Dataset.csv", stringsAsFactors = FALSE)
df$Cleaned_Title <- sapply(df$Title, clean_text)
df$Cleaned_Description <- sapply(df$Description, clean_text)
```

Spell Checking:

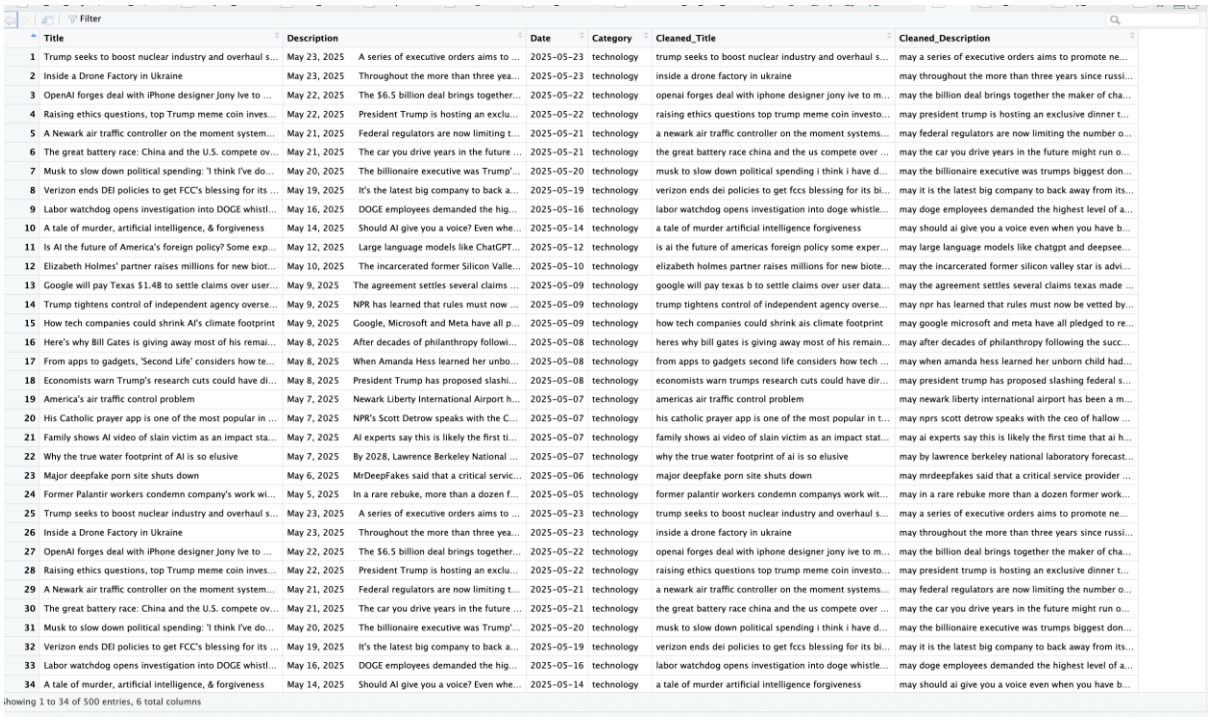
The “spell_check_text()” function checks and corrects spelling in cleaned text using the “hunspell” package. It tokenizes the text, replaces misspelled words with the first suggested correction, and returns the cleaned sentence in lowercase. This function is applied to the all attributes of the cleaned dataset.

INPUT:

```
spell_check_text <- function(text) {
  if (is.na(text) || !is.character(text) || nchar(text) == 0) return(NA)
  tokens <- unlist(strsplit(text, "\\s+"))
  corrected <- sapply(tokens, function(word) {
    if (!hunspell_check(word)) {
      suggestions <- hunspell_suggest(word)[[1]]
      valid <- suggestions[!grepl("[-\\s]", suggestions)]
      if (length(valid) > 0) return(tolower(valid[1]))
      return("")
    }
    return(tolower(word))
  })
  corrected <- corrected[corrected != ""]
  return(paste(corrected, collapse = " "))
}

df$Spellchecked_Title <- sapply(df$Cleaned_Title, spell_check_text)
df$Spellchecked_Description <- sapply(df$Cleaned_Description, spell_check_text)
write.csv(df, "/Users/mithizaman/Documents/12/Introduction to data science/code/final/npr_cleaned_spellchecked.csv", row.names = FALSE)
```

OUTPUT:



Filter

	Title	Description		Date	Category	Cleaned_Title	Cleaned_Description
1	Trump seeks to boost nuclear industry and overhaul s...	May 23, 2025	A series of executive orders aims to ...	2025-05-23	technology	trump seeks to boost nuclear industry and overhaul s...	may a series of executive orders aims to promote ne...
2	Inside a Drone Factory in Ukraine	May 23, 2025	Throughout the more than three yea...	2025-05-23	technology	inside a drone factory in ukraine	may throughout the more than three years since russi...
3	OpenAI forges deal with iPhone designer Jony Ive to ...	May 22, 2025	The $6.5 billion deal brings together...	2025-05-22	technology	openai forges deal with iphone designer jony ive to m...	may the billion deal brings together the maker of cha...
4	Raising ethics questions, top Trump meme coin inves...	May 22, 2025	President Trump is hosting an exclu...	2025-05-22	technology	raising ethics questions top trump meme coin investo...	may president trump is hosting an exclusive dinner t...
5	A Newark air traffic controller on the moment system...	May 21, 2025	Federal regulators are now limiting t...	2025-05-21	technology	a newark air traffic controller on the moment systems...	may federal regulators are now limiting the number o...
6	The great battery race: China and the U.S. compete ov...	May 21, 2025	The car you drive years in the future ...	2025-05-21	technology	the great battery race china and the us compete over ...	may the car you drive years in the future might run o...
7	Musk to slow down political spending: 'I think I've do...	May 20, 2025	The billionaire executive was Trump'...	2025-05-20	technology	musk to slow down political spending i think i have d...	may the billionaire executive was trumps biggest don...
8	Verizon ends DEI policies to get FCC's blessing for its ...	May 19, 2025	It's the latest big company to back a...	2025-05-19	technology	verizon ends dei policies to get fccs blessing for its bi...	may it is the latest big company to back away from its...
9	Labor watchdog opens investigation into DOGE whistl...	May 16, 2025	DOGE employees demanded the hig...	2025-05-16	technology	labor watchdog opens investigation into doge whistle...	may doge employees demanded the highest level of a...
10	A tale of murder, artificial intelligence, & forgiveness	May 14, 2025	Should AI give you a voice? Even whe...	2025-05-14	technology	a tale of murder artificial intelligence forgiveness	may should ai give you a voice even when you have b...
11	Is AI the future of America's foreign policy? Some exp...	May 12, 2025	Large language models like ChatGPT...	2025-05-12	technology	is ai the future of americas foreign policy some exper...	may large language models like chatgpt and deepsee...
12	Elizabeth Holmes' partner raises millions for new biot...	May 10, 2025	The incarcerated former Silicon Valle...	2025-05-10	technology	elizabeth holmes partner raises millions for new biote...	may the incarcerated former silicon valley star is advi...
13	Google will pay Texas $1.4B to settle claims over user...	May 9, 2025	The agreement settles several claims ...	2025-05-09	technology	google will pay texas b to settle claims over user data...	may the agreement settles several claims texas made ...
14	Trump tightens control of independent agency overse...	May 9, 2025	NPR has learned that rules must now ...	2025-05-09	technology	trump tightens control of independent agency overse...	may npr has learned that rules must now be vetted by...
15	How tech companies could shrink AI's climate footprint	May 9, 2025	Google, Microsoft and Meta have all p...	2025-05-09	technology	how tech companies could shrink ais climate footprint	may google microsoft and meta have all pledged to re...
16	Here's why Bill Gates is giving away most of his remai...	May 8, 2025	After decades of philanthropy followi...	2025-05-08	technology	heres why bill gates is giving away most of his remain...	may after decades of philanthropy following the succ...
17	From apps to gadgets, 'Second Life' considers how te...	May 8, 2025	When Amanda Hess learned her unbo...	2025-05-08	technology	from apps to gadgets second life considers how tech ...	may when amanda hess learned her unborn child had...
18	Economists warn Trump's research cuts could have di...	May 8, 2025	President Trump has proposed slashi...	2025-05-08	technology	economists warn trumps research cuts could have dir...	may president trump has proposed slashing federal s...
19	America's air traffic control problem	May 7, 2025	Newark Liberty International Airport h...	2025-05-07	technology	americas air traffic control problem	may newark liberty international airport has been a m...
20	His Catholic prayer app is one of the most popular in ...	May 7, 2025	NPR's Scott Detrow speaks with the C...	2025-05-07	technology	his catholic prayer app is one of the most popular in t...	may nprs scott detrow speaks with the ceo of hallow ...
21	Family shows AI video of slain victim as an impact sta...	May 7, 2025	AI experts say this is likely the first ti...	2025-05-07	technology	family shows ai video of slain victim as an impact stat...	may ai experts say this is likely the first time that ai h...
22	Why the true water footprint of AI is so elusive	May 7, 2025	By 2028, Lawrence Berkeley National ...	2025-05-07	technology	why the true water footprint of ai is so elusive	may by lawrence berkeley national laboratory forecast...
23	Major deepfake porn site shuts down	May 6, 2025	MrDeepFakes said that a critical servic...	2025-05-06	technology	major deepfake porn site shuts down	may mrdeepfakes said that a critical service provider ...
24	Former Palantir workers condemn company's work wi...	May 5, 2025	In a rare rebuke, more than a dozen f...	2025-05-05	technology	former palantir workers condemn companys work wit...	may in a rare rebuke more than a dozen former work...
25	Trump seeks to boost nuclear industry and overhaul s...	May 23, 2025	A series of executive orders aims to ...	2025-05-23	technology	trump seeks to boost nuclear industry and overhaul s...	may a series of executive orders aims to promote ne...
26	Inside a Drone Factory in Ukraine	May 23, 2025	Throughout the more than three yea...	2025-05-23	technology	inside a drone factory in ukraine	may throughout the more than three years since russi...
27	OpenAI forges deal with iPhone designer Jony Ive to ...	May 22, 2025	The $6.5 billion deal brings together...	2025-05-22	technology	openai forges deal with iphone designer jony ive to m...	may the billion deal brings together the maker of cha...
28	Raising ethics questions, top Trump meme coin inves...	May 22, 2025	President Trump is hosting an exclu...	2025-05-22	technology	raising ethics questions top trump meme coin investo...	may president trump is hosting an exclusive dinner t...
29	A Newark air traffic controller on the moment system...	May 21, 2025	Federal regulators are now limiting t...	2025-05-21	technology	a newark air traffic controller on the moment systems...	may federal regulators are now limiting the number o...
30	The great battery race: China and the U.S. compete ov...	May 21, 2025	The car you drive years in the future ...	2025-05-21	technology	the great battery race china and the us compete over ...	may the car you drive years in the future might run o...
31	Musk to slow down political spending: 'I think I've do...	May 20, 2025	The billionaire executive was Trump'...	2025-05-20	technology	musk to slow down political spending i think i have d...	may the billionaire executive was trumps biggest don...
32	Verizon ends DEI policies to get FCC's blessing for its ...	May 19, 2025	It's the latest big company to back a...	2025-05-19	technology	verizon ends dei policies to get fccs blessing for its bi...	may it is the latest big company to back away from its...
33	Labor watchdog opens investigation into DOGE whistl...	May 16, 2025	DOGE employees demanded the hig...	2025-05-16	technology	labor watchdog opens investigation into doge whistle...	may doge employees demanded the highest level of a...
34	A tale of murder, artificial intelligence, & forgiveness	May 14, 2025	Should AI give you a voice? Even whe...	2025-05-14	technology	a tale of murder artificial intelligence forgiveness	may should ai give you a voice even when you have b...

Showing 1 to 34 of 500 entries, 6 total columns

Tokenization:

This code defines a “format_tokens” function in order to tokenize text into words and sentences and format them as quoted lists. Then it loads the spell-checked dataset and applies the function to the title and description columns to create “tokenized_title” and “tokenized_description” and saves the updated dataset to a new CSV file.

INPUT:

```
format_tokens <- function(text) {
  if (is.na(text) || !is.character(text) || nchar(text) == 0) return(NA)
  sentence_tokens <- tokenize_sentences(text)[[1]]
  formatted_sentences <- sapply(sentence_tokens, function(sentence) {
    word_tokens <- tokenize_words(sentence)[[1]]
    paste0("[", paste0("'", word_tokens, "'", collapse = ", "), "]")
  })
  return(paste(formatted_sentences, collapse = " "))
}

df <- read_csv("/Users/mithizaman/Documents/12/Introduction to data science/code/final/npr_cleaned_spellchecked.csv")
df_tokenized <- df %>%
  mutate(
    tokenized_title = sapply(Spellchecked_Title, format_tokens),
    tokenized_description = sapply(Spellchecked_Description, format_tokens)
  )
write_csv(df_tokenized, "/Users/mithizaman/Documents/12/Introduction to data science/code/final/npr_tokenized_fancy_format.csv")
View(df_tokenized)
```

OUTPUT:

_Title	Cleaned_Description	Spellchecked_Title	Spellchecked_Description	tokenized_title	tokenized_description
eks to boost nuclear industry and overhaul s...	may a series of executive orders aims to promote ne...	trump seeks to boost nuclear industry and overhaul s...	may a series of executive orders aims to promote ne...	['trump', 'seeks', 'to', 'boost', 'nuclear', 'industry', 'and'...	['may', 'a', 'series', 'of', 'executive', 'orders', 'aims', 'to',...
drone factory in ukraine	may throughout the more than three years since russi...	inside a drone factory in ukraine	may throughout the more than three years since inva...	['inside', 'a', 'drone', 'factory', 'in', 'ukraine']	['may', 'throughout', 'the', 'more', 'than', 'three', 'years'...
orges deal with iphone designer jony ive to m...	may the billion deal brings together the maker of cha...	opening forges deal with iphone designer joy vie to m...	may the billion deal brings together the maker of cha...	['opening', 'forges', 'deal', 'with', 'iphone', 'designer', ']...	['may', 'the', 'billion', 'deal', 'brings', 'together', 'the', '...
thics questions top trump meme coin investo...	may president trump is hosting an exclusive dinner t...	raising ethics questions top trump meme coin investo...	may president trump is hosting an exclusive dinner t...	['raising', 'ethics', 'questions', 'top', 'trump', 'meme', 'c...	['may', 'president', 'trump', 'is', 'hosting', 'an', 'exclusiv...
air traffic controller on the moment systems...	may federal regulators are now limiting the number o...	a newark air traffic controller on the moment systems...	may federal regulators are now limiting the number o...	['a', 'newark', 'air', 'traffic', 'controller', 'on', 'the', 'mo...	['may', 'federal', 'regulators', 'are', 'now', 'limiting', 'the...
battery race china and the us compete over ...	may the car you drive years in the future might run o...	the great battery race china and the us compete over ...	may the car you drive years in the future might run o...	['the', 'great', 'battery', 'race', 'china', 'and', 'the', 'us', '...	['may', 'the', 'car', 'you', 'drive', 'years', 'in', 'the', 'futur...
slow down political spending i think i have d...	may the billionaire executive was trumps biggest don...	musk to slow down political spending i think i have d...	may the billionaire executive was trumps biggest don...	['musk', 'to', 'slow', 'down', 'political', 'spending', 'i', 'th...	['may', 'the', 'billionaire', 'executive', 'was', 'trumps', 'b...
nds dei policies to get fccs blessing for its bi...	may it is the latest big company to back away from its...	verizon ends die policies to get fcc blessing for its bill...	may it is the latest big company to back away from its...	['verizon', 'ends', 'die', 'policies', 'to', 'get', 'fcc', 'blessi...	['may', 'it', 'is', 'the', 'latest', 'big', 'company', 'to', 'back...
tchdog opens investigation into doge whistle...	may doge employees demanded the highest level of a...	labor watchdog opens investigation into doge claims ...	may doge employees demanded the highest level of a...	['labor', 'watchdog', 'opens', 'investigation', 'into', 'dog...	['may', 'doge', 'employees', 'demanded', 'the', 'highest'...
murder artificial intelligence forgiveness	may should ai give you a voice even when you have b...	a tale of murder artificial intelligence forgiveness	may should ai give you a voice even when you have b...	['a', 'tale', 'of', 'murder', 'artificial', 'intelligence', 'forgi...	['may', 'should', 'ai', 'give', 'you', 'a', 'voice', 'even', 'wh...
future of americas foreign policy some exper...	may large language models like chatgpt and deepsee...	is ai the future of foreign policy some experts think so	may large language models like chatting and are incr...	['is', 'ai', 'the', 'future', 'of', 'foreign', 'policy', 'some', 'e...	['may', 'large', 'language', 'models', 'like', 'chatting', 'a...
holmes partner raises millions for new biote...	may the incarcerated former silicon valley star is advi...	elizabeth partner raises millions for new biotech testi...	may the incarcerated former silicon valley star is advi...	['elizabeth', 'partner', 'raises', 'millions', 'for', 'new', 'bi...	['may', 'the', 'incarcerated', 'former', 'silicon', 'valley', '...
ill pay texas b to settle claims over user data...	may the agreement settles several claims texas made ...	google will pay b to settle claims over user data collec...	may the agreement settles several claims made again...	['google', 'will', 'pay', 'b', 'to', 'settle', 'claims', 'over', 'u...	['may', 'the', 'agreement', 'settles', 'several', 'claims', '...
ghtens control of independent agency overse...	may npr has learned that rules must now be vetted by...	trump tightens control of independent agency overse...	may npr has learned that rules must now be vetted by...	['trump', 'tightens', 'control', 'of', 'independent', 'agen...	['may', 'npr', 'has', 'learned', 'that', 'rules', 'must', 'now'...
companies could shrink ais climate footprint	may google microsoft and meta have all pledged to re...	how tech companies could shrink aid climate footprint	may google and meta have all pledged to reach at lea...	['how', 'tech', 'companies', 'could', 'shrink', 'aid', 'clima...	['may', 'google', 'and', 'meta', 'have', 'all', 'pledged', 'to'...
y bill gates is giving away most of his remain...	may after decades of philanthropy following the succ...	here why bill gates is giving away most of his remaini...	may after decades of philanthropy following the succ...	['here', 'why', 'bill', 'gates', 'is', 'giving', 'away', 'most', '...	['may', 'after', 'decades', 'of', 'philanthropy', 'following'...
s to gadgets second life considers how tech ...	may when amanda hess learned her unborn child had...	from apps to gadgets second life considers how tech ...	may when amanda learned her unborn child had a ge...	['from', 'apps', 'to', 'gadgets', 'second', 'life', 'considers...	['may', 'when', 'amanda', 'learned', 'her', 'unborn', 'chil...
sts warn trumps research cuts could have dir...	may president trump has proposed slashing federal s...	economists warn trumps research cuts could have dir...	may president trump has proposed slashing federal s...	['economists', 'warn', 'trumps', 'research', 'cuts', 'could...	['may', 'president', 'trump', 'has', 'proposed', 'slashing'...
air traffic control problem	may newark liberty international airport has been a m...	air traffic control problem	may newark liberty international airport has been a m...	['air', 'traffic', 'control', 'problem']	['may', 'newark', 'liberty', 'international', 'airport', 'has',...
ilic prayer app is one of the most popular in t...	may nprs scott detrow speaks with the ceo of hallow ...	his catholic prayer app is one of the most popular in t...	may npr scott dethrone speaks with the ceo of hallow...	['his', 'catholic', 'prayer', 'app', 'is', 'one', 'of', 'the', 'mo...	['may', 'npr', 'scott', 'dethrone', 'speaks', 'with', 'the', 'c...
ows ai video of slain victim as an impact stat...	may ai experts say this is likely the first time that ai h...	family shows ai video of slain victim as an impact stat...	may ai experts say this is likely the first time that ai h...	['family', 'shows', 'ai', 'video', 'of', 'slain', 'victim', 'as', '...	['may', 'ai', 'experts', 'say', 'this', 'is', 'likely', 'the', 'first'...
true water footprint of ai is so elusive	may by lawrence berkeley national laboratory forecast...	why the true water footprint of ai is so elusive	may by lawrence berkeley national laboratory forecast...	['why', 'the', 'true', 'water', 'footprint', 'of', 'ai', 'is', 'so', ...	['may', 'by', 'lawrence', 'berkeley', 'national', 'laborator...
epfake porn site shuts down	may mrdeepfakes said that a critical service provider ...	major deep fake porn site shuts down	may said that a critical service provider terminated se...	['major', 'deep', 'fake', 'porn', 'site', 'shuts', 'down']	['may', 'said', 'that', 'a', 'critical', 'service', 'provider', 'te...
alantir workers condemn companys work wit...	may in a rare rebuke more than a dozen former work...	former palatial workers condemn company work with ...	may in a rare rebuke more than a dozen former work...	['former', 'palatial', 'workers', 'condemn', 'company', '...	['may', 'in', 'a', 'rare', 'rebuke', 'more', 'than', 'a', 'dozen...

Stopword Removal:

This code defines a “process_tokens” function that keeps only meaningful words in tokenized text after removing brackets, quotations, and stopwords. It loads the tokenized dataset, applies the function to the title and description columns to create “title_no_stopwords” and “description_no_stopwords”, then saves the updated data to a new CSV file.

INPUT:

```
process_tokens <- function(text) {
  if (is.na(text) || !is.character(text) || nchar(text) == 0) return(NA)
  text <- gsub("\\[|\\]|'", "", text)
  tokens <- unlist(strsplit(text, ",\\s*"))
  filtered_tokens <- tokens[!tolower(tokens) %in% stopwords("en")]
  paste0("[", paste0("'", filtered_tokens, "'", collapse = ", "), "]")
}

data <- read_csv("/Users/mithizaman/Documents/12/Introduction to data science/code/final/npr_tokenized_fancy_format.csv")
data_filtered <- data %>%
  mutate(
    title_no_stopwords = sapply(tokenized_title, process_tokens),
    description_no_stopwords = sapply(tokenized_description, process_tokens)
  )
write_csv(data_filtered, "/Users/mithizaman/Documents/12/Introduction to data science/code/final/npr_stopwords_removed_fancy_format.csv")

View(data_filtered)
```

OUTPUT:

cked_Title	Spellchecked_Description	tokenized_title	tokenized_description	title_no_stopwords	description_no_stopwords
eks to boost nuclear industry and overhaul s...	may a series of executive orders aims to promote ne...	['trump', 'seeks', 'to', 'boost', 'nuclear', 'industry', 'and'...	['may', 'a', 'series', 'of', 'executive', 'orders', 'aims', 'to',...	['trump', 'seeks', 'boost', 'nuclear', 'industry', 'overhaul...	['may', 'series', 'executive', 'orders', 'aims', 'promote', '...
drone factory in ukraine	may throughout the more than three years since inva...	['inside', 'a', 'drone', 'factory', 'in', 'ukraine']	['may', 'throughout', 'the', 'more', 'than', 'three', 'years'...	['inside', 'drone', 'factory', 'ukraine']	['may', 'throughout', 'three', 'years', 'since', 'invasion', '...
forges deal with iphone designer joy vie to m...	may the billion deal brings together the maker of cha...	['opening', 'forges', 'deal', 'with', 'iphone', 'designer', 'j...	['may', 'the', 'billion', 'deal', 'brings', 'together', 'the', '...	['opening', 'forges', 'deal', 'iphone', 'designer', 'joy', 'vi...	['may', 'billion', 'deal', 'brings', 'together', 'maker', 'cha...
thics questions top trump meme coin investo...	may president trump is hosting an exclusive dinner t...	['raising', 'ethics', 'questions', 'top', 'trump', 'meme', 'c...	['may', 'president', 'trump', 'is', 'hosting', 'an', 'exclusiv...	['raising', 'ethics', 'questions', 'top', 'trump', 'meme', 'c...	['may', 'president', 'trump', 'hosting', 'exclusive', 'dinn...
: air traffic controller on the moment systems...	may federal regulators are now limiting the number o...	['a', 'newark', 'air', 'traffic', 'controller', 'on', 'the', 'mo...	['may', 'federal', 'regulators', 'are', 'now', 'limiting', 'the...	['newark', 'air', 'traffic', 'controller', 'moment', 'systems...	['may', 'federal', 'regulators', 'now', 'limiting', 'number'...
: battery race china and the us compete over ...	may the car you drive years in the future might run o...	['the', 'great', 'battery', 'race', 'china', 'and', 'the', 'us', '...	['may', 'the', 'car', 'you', 'drive', 'years', 'in', 'the', 'futur...	['great', 'battery', 'race', 'china', 'us', 'compete', 'future'...	['may', 'car', 'drive', 'years', 'future', 'might', 'run', 'batt...
slow down political spending i think i have d...	may the billionaire executive was trumps biggest don...	['musk', 'to', 'slow', 'down', 'political', 'spending', 'i', 'th...	['may', 'the', 'billionaire', 'executive', 'was', 'trumps', 'b...	['musk', 'slow', 'political', 'spending', 'think', 'done', 'e...	['may', 'billionaire', 'executive', 'trumps', 'biggest', 'do...
nds die policies to get fcc blessing for its bill...	may it is the latest big company to back away from its...	['verizon', 'ends', 'die', 'policies', 'to', 'get', 'fcc', 'blessi...	['may', 'it', 'is', 'the', 'latest', 'big', 'company', 'to', 'back...	['verizon', 'ends', 'die', 'policies', 'get', 'fcc', 'blessing', '...	['may', 'latest', 'big', 'company', 'back', 'away', 'diversit...
tchdog opens investigation into doge claims ...	may doge employees demanded the highest level of a...	['labor', 'watchdog', 'opens', 'investigation', 'into', 'dog...	['may', 'doge', 'employees', 'demanded', 'the', 'highest'...	['labor', 'watchdog', 'opens', 'investigation', 'doge', 'cla...	['may', 'doge', 'employees', 'demanded', 'highest', 'leve...
murder artificial intelligence forgiveness	may should ai give you a voice even when you have b...	['a', 'tale', 'of', 'murder', 'artificial', 'intelligence', 'forgi...	['may', 'should', 'ai', 'give', 'you', 'a', 'voice', 'even', 'wh...	['tale', 'murder', 'artificial', 'intelligence', 'forgiveness']	['may', 'ai', 'give', 'voice', 'even', 'murdered']
future of foreign policy some experts think so	may large language models like chatting and are incr...	['is', 'ai', 'the', 'future', 'of', 'foreign', 'policy', 'some', 'e...	['may', 'large', 'language', 'models', 'like', 'chatting', 'a...	['ai', 'future', 'foreign', 'policy', 'experts', 'think']	['may', 'large', 'language', 'models', 'like', 'chatting', 'in...
: partner raises millions for new biotech testi...	may the incarcerated former silicon valley star is advi...	['elizabeth', 'partner', 'raises', 'millions', 'for', 'new', 'bi...	['may', 'the', 'incarcerated', 'former', 'silicon', 'valley', '...	['elizabeth', 'partner', 'raises', 'millions', 'new', 'biotech...	['may', 'incarcerated', 'former', 'silicon', 'valley', 'star', '...
ill pay b to settle claims over user data collec...	may the agreement settles several claims made again...	['google', 'will', 'pay', 'b', 'to', 'settle', 'claims', 'over', 'u...	['may', 'the', 'agreement', 'settles', 'several', 'claims', '...	['google', 'pay', 'b', 'settle', 'claims', 'user', 'data', 'colle...	['may', 'agreement', 'settles', 'several', 'claims', 'made',...
ghtens control of independent agency overse...	may npr has learned that rules must now be vetted by...	['trump', 'tightens', 'control', 'of', 'independent', 'agen...	['may', 'npr', 'has', 'learned', 'that', 'rules', 'must', 'now'...	['trump', 'tightens', 'control', 'independent', 'agency', '...	['may', 'npr', 'learned', 'rules', 'must', 'now', 'vetted', 'w...
: companies could shrink aid climate footprint	may google and meta have all pledged to reach at lea...	['how', 'tech', 'companies', 'could', 'shrink', 'aid', 'clima...	['may', 'google', 'and', 'meta', 'have', 'all', 'pledged', 'to'...	['tech', 'companies', 'shrink', 'aid', 'climate', 'footprint']	['may', 'google', 'meta', 'pledged', 'reach', 'least', 'net', ...
r bill gates is giving away most of his remaini...	may after decades of philanthropy following the succ...	['here', 'why', 'bill', 'gates', 'is', 'giving', 'away', 'most', '...	['may', 'after', 'decades', 'of', 'philanthropy', 'following'...	['bill', 'gates', 'giving', 'away', 'remaining', 'fortune', 'ch...	['may', 'decades', 'philanthropy', 'following', 'success', ...
ps to gadgets second life considers how tech ...	may when amanda learned her unborn child had a ge...	['from', 'apps', 'to', 'gadgets', 'second', 'life', 'considers...	['may', 'when', 'amanda', 'learned', 'her', 'unborn', 'chil'...	['apps', 'gadgets', 'second', 'life', 'considers', 'tech', 'ch...	['may', 'amanda', 'learned', 'unborn', 'child', 'genetic', '...
sts warn trumps research cuts could have dir...	may president trump has proposed slashing federal s...	['economists', 'warn', 'trumps', 'research', 'cuts', 'could...	['may', 'president', 'trump', 'has', 'proposed', 'slashing'...	['economists', 'warn', 'trumps', 'research', 'cuts', 'dire', ...	['may', 'president', 'trump', 'proposed', 'slashing', 'fed...
: control problem	may newark liberty international airport has been a m...	['air', 'traffic', 'control', 'problem']	['may', 'newark', 'liberty', 'international', 'airport', 'has',...	['air', 'traffic', 'control', 'problem']	['may', 'newark', 'liberty', 'international', 'airport', 'mes...
ilic prayer app is one of the most popular in t...	may npr scott dethrone speaks with the ceo of hallow...	['his', 'catholic', 'prayer', 'app', 'is', 'one', 'of', 'the', 'mo...	['may', 'npr', 'scott', 'dethrone', 'speaks', 'with', 'the', 'c...	['catholic', 'prayer', 'app', 'one', 'popular', 'world', 'take...	['may', 'npr', 'scott', 'dethrone', 'speaks', 'ceo', 'hallow',...
ows ai video of slain victim as an impact stat...	may ai experts say this is likely the first time that ai h...	['family', 'shows', 'ai', 'video', 'of', 'slain', 'victim', 'as', '...	['may', 'ai', 'experts', 'say', 'this', 'is', 'likely', 'the', 'first'...	['family', 'shows', 'ai', 'video', 'slain', 'victim', 'impact', '...	['may', 'ai', 'experts', 'say', 'likely', 'first', 'time', 'ai', 'us...
true water footprint of ai is so elusive	may by lawrence berkeley national laboratory forecast...	['why', 'the', 'true', 'water', 'footprint', 'of', 'ai', 'is', 'so', ...	['may', 'by', 'lawrence', 'berkeley', 'national', 'laborator...	['true', 'water', 'footprint', 'ai', 'elusive']	['may', 'lawrence', 'berkeley', 'national', 'laboratory', 'fo...
ep fake porn site shuts down	may said that a critical service provider terminated se...	['major', 'deep', 'fake', 'porn', 'site', 'shuts', 'down']	['may', 'said', 'that', 'a', 'critical', 'service', 'provider', 'te...	['major', 'deep', 'fake', 'porn', 'site', 'shuts']	['may', 'said', 'critical', 'service', 'provider', 'terminated'...
alatial workers condemn company work with ...	may in a rare rebuke more than a dozen former work...	['former', 'palatial', 'workers', 'condemn', 'company', '...	['may', 'in', 'a', 'rare', 'rebuke', 'more', 'than', 'a', 'dozen...	['former', 'palatial', 'workers', 'condemn', 'company', '...	['may', 'rare', 'rebuke', 'dozen', 'former', 'workers', 'po...
eks to boost nuclear industry and overhaul s...	may a series of executive orders aims to promote ne...	['trump', 'seeks', 'to', 'boost', 'nuclear', 'industry', 'and'...	['may', 'a', 'series', 'of', 'executive', 'orders', 'aims', 'to',...	['trump', 'seeks', 'boost', 'nuclear', 'industry', 'overhaul...	['may', 'series', 'executive', 'orders', 'aims', 'promote', '...
drone factory in ukraine	may throughout the more than three years since inva...	['inside', 'a', 'drone', 'factory', 'in', 'ukraine']	['may', 'throughout', 'the', 'more', 'than', 'three', 'years'...	['inside', 'drone', 'factory', 'ukraine']	['may', 'throughout', 'three', 'years', 'since', 'invasion', '...
forges deal with iphone designer joy vie to m...	may the billion deal brings together the maker of cha...	['opening', 'forges', 'deal', 'with', 'iphone', 'designer', 'j...	['may', 'the', 'billion', 'deal', 'brings', 'together', 'the', '...	['opening', 'forges', 'deal', 'iphone', 'designer', 'joy', 'vi...	['may', 'billion', 'deal', 'brings', 'together', 'maker', 'cha...
thics questions top trump meme coin investo...	may president trump is hosting an exclusive dinner t...	['raising', 'ethics', 'questions', 'top', 'trump', 'meme', 'c...	['may', 'president', 'trump', 'is', 'hosting', 'an', 'exclusiv...	['raising', 'ethics', 'questions', 'top', 'trump', 'meme', 'c...	['may', 'president', 'trump', 'hosting', 'exclusive', 'dinn...
: air traffic controller on the moment systems...	may federal regulators are now limiting the number o...	['a', 'newark', 'air', 'traffic', 'controller', 'on', 'the', 'mo...	['may', 'federal', 'regulators', 'are', 'now', 'limiting', 'the...	['newark', 'air', 'traffic', 'controller', 'moment', 'systems...	['may', 'federal', 'regulators', 'now', 'limiting', 'number'...
: battery race china and the us compete over ...	may the car you drive years in the future might run o...	['the', 'great', 'battery', 'race', 'china', 'and', 'the', 'us', '...	['may', 'the', 'car', 'you', 'drive', 'years', 'in', 'the', 'futur...	['great', 'battery', 'race', 'china', 'us', 'compete', 'future'...	['may', 'car', 'drive', 'years', 'future', 'might', 'run', 'batt...
slow down political spending i think i have d...	may the billionaire executive was trumps biggest don...	['musk', 'to', 'slow', 'down', 'political', 'spending', 'i', 'th...	['may', 'the', 'billionaire', 'executive', 'was', 'trumps', 'b...	['musk', 'slow', 'political', 'spending', 'think', 'done', 'e...	['may', 'billionaire', 'executive', 'trumps', 'biggest', 'do...
nds die policies to get fcc blessing for its bill...	may it is the latest big company to back away from its...	['verizon', 'ends', 'die', 'policies', 'to', 'get', 'fcc', 'blessi...	['may', 'it', 'is', 'the', 'latest', 'big', 'company', 'to', 'back...	['verizon', 'ends', 'die', 'policies', 'get', 'fcc', 'blessing', '...	['may', 'latest', 'big', 'company', 'back', 'away', 'diversit...
tchdog opens investigation into doge claims ...	may doge employees demanded the highest level of a...	['labor', 'watchdog', 'opens', 'investigation', 'into', 'dog...	['may', 'doge', 'employees', 'demanded', 'the', 'highest'...	['labor', 'watchdog', 'opens', 'investigation', 'doge', 'cla...	['may', 'doge', 'employees', 'demanded', 'highest', 'leve...
murder artificial intelligence forgiveness	may should ai give you a voice even when you have b...	['a', 'tale', 'of', 'murder', 'artificial', 'intelligence', 'forgi...	['may', 'should', 'ai', 'give', 'you', 'a', 'voice', 'even', 'wh...	['tale', 'murder', 'artificial', 'intelligence', 'forgiveness']	['may', 'ai', 'give', 'voice', 'even', 'murdered']
future of foreign policy some experts think so	may large language models like chatting and are incr...	['is', 'ai', 'the', 'future', 'of', 'foreign', 'policy', 'some', 'e...	['may', 'large', 'language', 'models', 'like', 'chatting', 'a...	['ai', 'future', 'foreign', 'policy', 'experts', 'think']	['may', 'large', 'language', 'models', 'like', 'chatting', 'in...
: partner raises millions for new biotech testi...	may the incarcerated former silicon valley star is advi...	['elizabeth', 'partner', 'raises', 'millions', 'for', 'new', 'bi...	['may', 'the', 'incarcerated', 'former', 'silicon', 'valley', '...	['elizabeth', 'partner', 'raises', 'millions', 'new', 'biotech...	['may', 'incarcerated', 'former', 'silicon', 'valley', 'star', '...
ill pay b to settle claims over user data collec...	may the agreement settles several claims made again...	['google', 'will', 'pay', 'b', 'to', 'settle', 'claims', 'over', 'u...	['may', 'the', 'agreement', 'settles', 'several', 'claims', '...	['google', 'pay', 'b', 'settle', 'claims', 'user', 'data', 'colle...	['may', 'agreement', 'settles', 'several', 'claims', 'made',...
ghtens control of independent agency overse...	may npr has learned that rules must now be vetted by...	['trump', 'tightens', 'control', 'of', 'independent', 'agen...	['may', 'npr', 'has', 'learned', 'that', 'rules', 'must', 'now'...	['trump', 'tightens', 'control', 'independent', 'agency', '...	['may', 'npr', 'learned', 'rules', 'must', 'now', 'vetted', 'w...
: companies could shrink aid climate footprint	may google and meta have all pledged to reach at lea...	['how', 'tech', 'companies', 'could', 'shrink', 'aid', 'clima...	['may', 'google', 'and', 'meta', 'have', 'all', 'pledged', 'to'...	['tech', 'companies', 'shrink', 'aid', 'climate', 'footprint']	['may', 'google', 'meta', 'pledged', 'reach', 'least', 'net', ...

Showing 1 to 38 of 500 entries, 12 total columns

Stemming and Lemmatization:

This code first loaded the cleaned and tokenized dataset from a CSV file then removes stopwords from tokenized text.

Next, it applies **stemming** to the tokenized news titles and descriptions in the dataset. Stemming reduces words to their root forms, and the stemmed tokens are saved in new columns “title_stem_only” and “description_stem_only”.

After that, **lemmatization** is performed separately on the original tokens (without stopwords). Lemmatization converts words to their dictionary base forms, producing “title_lemma_only” and “description_lemma_only” columns.

INPUT:

```
clean_and_split <- function(text) {
  cleaned <- gsub("\\[|\\]|'", "", text)
  tokens <- unlist(strsplit(cleaned, ",\\s*"))
  tokens <- tokens[!tolower(tokens) %in% stopwords("en")]
  return(tokens)
}


data_stemmed <- data %>%
  mutate(
    title_stem_only = sapply(title_no_stopwords, function(text) {
      tokens <- clean_and_split(text)
      stemmed <- wordStem(tokens, language = "en")
      paste0("[", paste0("'", stemmed, "'", collapse = ", "), "]")
    }),
    description_stem_only = sapply(description_no_stopwords, function(text) {
      tokens <- clean_and_split(text)
      stemmed <- wordStem(tokens, language = "en")
      paste0("[", paste0("'", stemmed, "'", collapse = ", "), "]")
    })
  )


data_lemma_only <- data_stemmed %>%
  mutate(
    title_lemma_only = sapply(title_no_stopwords, function(text) {
      tokens <- clean_and_split(text)
      lemmatized <- lemmatize_words(tokens)
      paste0("[", paste0("'", lemmatized, "'", collapse = ", "), "]")
    }),
    description_lemma_only = sapply(description_no_stopwords, function(text) {
      tokens <- clean_and_split(text)
      lemmatized <- lemmatize_words(tokens)
      paste0("[", paste0("'", lemmatized, "'", collapse = ", "), "]")
    })
  )

write_csv(data_lemma_only, "/Users/mithizaman/Documents/12/Introduction to data science/code/final/npr_stem_lemma_se

View(data_lemma_only)
```

OUTPUT:

_stopwords	description_no_stopwords	title_stem_only	description_stem_only	title_lemma_only	description_lemma_only
'seeks', 'boost', 'nuclear', 'industry', 'overhaul...	['may', 'series', 'executive', 'orders', 'aims', 'promote', '...	['trump', 'seek', 'boost', 'nuclear', 'industri', 'overhaul', ...	['may', 'seri', 'execut', 'order', 'aim', 'promot', 'new', 'ki...	['trump', 'seek', 'boost', 'nuclear', 'industry', 'overhaul',...	['may', 'series', 'executive', 'order', 'aim', 'promote', 'ne...
'drone', 'factory', 'ukraine']	['may', 'throughout', 'three', 'years', 'since', 'invasion', '...	['insid', 'drone', 'factori', 'ukrain']	['may', 'throughout', 'three', 'year', 'sinc', 'invas', 'ukrai...	['inside', 'drone', 'factory', 'ukraine']	['may', 'throughout', 'three', 'year', 'since', 'invasion', '...
j', 'forges', 'deal', 'iphone', 'designer', 'joy', 'vi...	['may', 'billion', 'deal', 'brings', 'together', 'maker', 'cha...	['open', 'forg', 'deal', 'iphon', 'design', 'joy', 'vie', 'make...	['may', 'billion', 'deal', 'bring', 'togeth', 'maker', 'chat', '...	['open', 'forge', 'deal', 'iphone', 'designer', 'joy', 'vie', '...	['may', 'billion', 'deal', 'bring', 'together', 'maker', 'chat'...
, 'ethics', 'questions', 'top', 'trump', 'meme', 'c...	['may', 'president', 'trump', 'hosting', 'exclusive', 'dinn...	['rais', 'ethic', 'question', 'top', 'trump', 'meme', 'coin', '...	['may', 'presid', 'trump', 'host', 'exclus', 'dinner', 'tonig...	['raise', 'ethic', 'question', 'top', 'trump', 'meme', 'coin',...	['may', 'president', 'trump', 'host', 'exclusive', 'dinner', ...
', 'air', 'traffic', 'controller', 'moment', 'systems...	['may', 'federal', 'regulators', 'now', 'limiting', 'number'...	['newark', 'air', 'traffic', 'control', 'moment', 'system', '...	['may', 'feder', 'regul', 'now', 'limit', 'number', 'flight', '...	['newark', 'air', 'traffic', 'controller', 'moment', 'system',...	['may', 'federal', 'regulator', 'now', 'limit', 'numb', 'fligh...
battery', 'race', 'china', 'us', 'compete', 'future'...	['may', 'car', 'drive', 'years', 'future', 'might', 'run', 'batt...	['great', 'batteri', 'race', 'china', 'us', 'compet', 'futur', 'e...	['may', 'car', 'drive', 'year', 'futur', 'might', 'run', 'batteri...	['great', 'battery', 'race', 'china', 'us', 'compete', 'future'...	['may', 'car', 'drive', 'year', 'future', 'may', 'run', 'battery...
'slow', 'political', 'spending', 'think', 'done', 'e...	['may', 'billionaire', 'executive', 'trumps', 'biggest', 'do...	['musk', 'slow', 'polit', 'spend', 'think', 'done', 'enough']	['may', 'billionair', 'execut', 'trump', 'biggest', 'donor', '...	['musk', 'slow', 'political', 'spend', 'think', 'do', 'enough']	['may', 'billionaire', 'executive', 'trump', 'big', 'donor', '...
', 'ends', 'die', 'policies', 'get', 'fcc', 'blessing', '...	['may', 'latest', 'big', 'company', 'back', 'away', 'diversit'...	['verizon', 'end', 'die', 'polici', 'get', 'fcc', 'bless', 'billion...	['may', 'latest', 'big', 'compani', 'back', 'away', 'divers', '...	['verizon', 'end', 'die', 'policy', 'get', 'fcc', 'bless', 'billio...	['may', 'late', 'big', 'company', 'back', 'away', 'diversity',...
watchdog', 'opens', 'investigation', 'doge', 'cla...	['may', 'doge', 'employees', 'demanded', 'highest', 'leve...	['labor', 'watchdog', 'open', 'investig', 'doge', 'claim', 'n...	['may', 'doge', 'employe', 'demand', 'highest', 'level', 'a...	['labor', 'watchdog', 'open', 'investigation', 'doge', 'clai...	['may', 'doge', 'employee', 'demand', 'high', 'level', 'acc...
'urder', 'artificial', 'intelligence', 'forgiveness']	['may', 'ai', 'give', 'voice', 'even', 'murdered']	['tale', 'murder', 'artifici', 'intellig', 'forgiv']	['may', 'ai', 'give', 'voic', 'even', 'murder']	['tale', 'murder', 'artificial', 'intelligence', 'forgiveness']	['may', 'ai', 'give', 'voice', 'even', 'murder']
'are', 'foreign', 'policy', 'experts', 'think']	['may', 'large', 'language', 'models', 'like', 'chatting', 'in...	['ai', 'futur', 'foreign', 'polici', 'expert', 'think']	['may', 'larg', 'languag', 'model', 'like', 'chat', 'increas', '...	['ai', 'future', 'foreign', 'policy', 'expert', 'think']	['may', 'large', 'language', 'model', 'like', 'chat', 'increas...
th', 'partner', 'raises', 'millions', 'new', 'biotech...	['may', 'incarcerated', 'former', 'silicon', 'valley', 'star', '...	['elizabeth', 'partner', 'rais', 'million', 'new', 'biotech', 't...	['may', 'incarcer', 'former', 'silicon', 'valley', 'star', 'advi...	['elizabeth', 'partner', 'raise', 'million', 'new', 'biotech', ...	['may', 'incarcerate', 'former', 'silicon', 'valley', 'star', 'a...
, 'pay', 'b', 'settle', 'claims', 'user', 'data', 'colle...	['may', 'agreement', 'settles', 'several', 'claims', 'made',...	['googl', 'pay', 'b', 'settl', 'claim', 'user', 'data', 'collect']	['may', 'agreement', 'settl', 'sever', 'claim', 'made', 'sear...	['google', 'pay', 'b', 'settle', 'claim', 'user', 'datum', 'coll...	['may', 'agreement', 'settle', 'several', 'claim', 'make', 's...
'tightens', 'control', 'independent', 'agency', '...	['may', 'npr', 'learned', 'rules', 'must', 'now', 'vetted', 'w...	['trump', 'tighten', 'control', 'independ', 'agenc', 'overs...	['may', 'npr', 'learn', 'rule', 'must', 'now', 'vet', 'white', '...	['trump', 'tighten', 'control', 'independent', 'agency', 'o...	['may', 'npr', 'learn', 'rule', 'must', 'now', 'vet', 'white', '...
ompanies', 'shrink', 'aid', 'climate', 'footprint']	['may', 'google', 'meta', 'pledged', 'reach', 'least', 'net', ...	['tech', 'compani', 'shrink', 'aid', 'climat', 'footprint']	['may', 'googl', 'meta', 'pledg', 'reach', 'least', 'net', 'zer...	['tech', 'company', 'shrink', 'aid', 'climate', 'footprint']	['may', 'google', 'meta', 'pledge', 'reach', 'less', 'net', 'z...
ites', 'giving', 'away', 'remaining', 'fortune', 'ch...	['may', 'decades', 'philanthropy', 'following', 'success', ...	['bill', 'gate', 'give', 'away', 'remain', 'fortun', 'charit']	['may', 'decad', 'philanthropi', 'follow', 'success', 'bill', '...	['bill', 'gate', 'give', 'away', 'remain', 'fortune', 'charity']	['may', 'decade', 'philanthropy', 'follow', 'success', 'bill'...
gadgets', 'second', 'life', 'considers', 'tech', 'ch...	['may', 'amanda', 'learned', 'unborn', 'child', 'genetic', '...	['app', 'gadget', 'second', 'life', 'consid', 'tech', 'chang', ...	['may', 'amanda', 'learn', 'unborn', 'child', 'genet', 'con...	['app', 'gadget', '2', 'life', 'consider', 'tech', 'change', 'b...	['may', 'amanda', 'learn', 'unborn', 'child', 'genetic', 'co...
iists', 'warn', 'trumps', 'research', 'cuts', 'dire', ...	['may', 'president', 'trump', 'proposed', 'slashing', 'fed...	['economist', 'warn', 'trump', 'research', 'cut', 'dire', 'co...	['may', 'presid', 'trump', 'propos', 'slash', 'feder', 'scien...	['economist', 'warn', 'trump', 'research', 'cut', 'dire', 'co...	['may', 'president', 'trump', 'propose', 'slash', 'federal', ...
ffic', 'control', 'problem']	['may', 'newark', 'liberty', 'international', 'airport', 'mes...	['air', 'traffic', 'control', 'problem']	['may', 'newark', 'liberti', 'intern', 'airport', 'mess', 'past...	['air', 'traffic', 'control', 'problem']	['may', 'newark', 'liberty', 'international', 'airport', 'mes...
', 'prayer', 'app', 'one', 'popular', 'world', 'take...	['may', 'npr', 'scott', 'dethrone', 'speaks', 'ceo', 'hallow',...	['cathol', 'prayer', 'app', 'one', 'popular', 'world', 'take', ...	['may', 'npr', 'scott', 'dethron', 'speak', 'ceo', 'hallow', 'c...	['catholic', 'prayer', 'app', 'one', 'popular', 'world', 'take...	['may', 'npr', 'scott', 'dethrone', 'speak', 'ceo', 'hallow', ...
'shows', 'ai', 'video', 'slain', 'victim', 'impact', '...	['may', 'ai', 'experts', 'say', 'likely', 'first', 'time', 'ai', 'us...	['famili', 'show', 'ai', 'video', 'slain', 'victim', 'impact', 'st...	['may', 'ai', 'expert', 'say', 'like', 'first', 'time', 'ai', 'use', ...	['family', 'show', 'ai', 'video', 'slay', 'victim', 'impact', 'st...	['may', 'ai', 'expert', 'say', 'likely', 'first', 'time', 'ai', 'use'...
ater', 'footprint', 'ai', 'elusive']	['may', 'lawrence', 'berkeley', 'national', 'laboratory', 'fo...	['true', 'water', 'footprint', 'ai', 'elus']	['may', 'lawrenc', 'berkeley', 'nation', 'laboratori', 'forec...	['true', 'water', 'footprint', 'ai', 'elusive']	['may', 'lawrence', 'berkeley', 'national', 'laboratory', 'fo...
'deep', 'fake', 'porn', 'site', 'shuts']	['may', 'said', 'critical', 'service', 'provider', 'terminated'...	['major', 'deep', 'fake', 'porn', 'site', 'shut']	['may', 'said', 'critic', 'servic', 'provid', 'termin', 'servic', ...	['major', 'deep', 'fake', 'porn', 'site', 'shut']	['may', 'say', 'critical', 'service', 'provider', 'terminate', '...
, 'palatial', 'workers', 'condemn', 'company', '...	['may', 'rare', 'rebuke', 'dozen', 'former', 'workers', 'po...	['former', 'palati', 'worker', 'condemn', 'compani', 'wor...	['may', 'rare', 'rebuk', 'dozen', 'former', 'worker', 'powe...	['former', 'palatial', 'worker', 'condemn', 'company', 'w...	['may', 'rare', 'rebuke', 'dozen', 'former', 'worker', 'pow...
'seeks', 'boost', 'nuclear', 'industry', 'overhaul...	['may', 'series', 'executive', 'orders', 'aims', 'promote', '...	['trump', 'seek', 'boost', 'nuclear', 'industri', 'overhaul', ...	['may', 'seri', 'execut', 'order', 'aim', 'promot', 'new', 'ki...	['trump', 'seek', 'boost', 'nuclear', 'industry', 'overhaul',...	['may', 'series', 'executive', 'order', 'aim', 'promote', 'ne...
'drone', 'factory', 'ukraine']	['may', 'throughout', 'three', 'years', 'since', 'invasion', '...	['insid', 'drone', 'factori', 'ukrain']	['may', 'throughout', 'three', 'year', 'sinc', 'invas', 'ukrai...	['inside', 'drone', 'factory', 'ukraine']	['may', 'throughout', 'three', 'year', 'since', 'invasion', '...
j', 'forges', 'deal', 'iphone', 'designer', 'joy', 'vi...	['may', 'billion', 'deal', 'brings', 'together', 'maker', 'cha...	['open', 'forg', 'deal', 'iphon', 'design', 'joy', 'vie', 'make...	['may', 'billion', 'deal', 'bring', 'togeth', 'maker', 'chat', '...	['open', 'forge', 'deal', 'iphone', 'designer', 'joy', 'vie', '...	['may', 'billion', 'deal', 'bring', 'together', 'maker', 'chat'...
, 'ethics', 'questions', 'top', 'trump', 'meme', 'c...	['may', 'president', 'trump', 'hosting', 'exclusive', 'dinn...	['rais', 'ethic', 'question', 'top', 'trump', 'meme', 'coin', '...	['may', 'presid', 'trump', 'host', 'exclus', 'dinner', 'tonig...	['raise', 'ethic', 'question', 'top', 'trump', 'meme', 'coin',...	['may', 'president', 'trump', 'host', 'exclusive', 'dinner', ...
', 'air', 'traffic', 'controller', 'moment', 'systems...	['may', 'federal', 'regulators', 'now', 'limiting', 'number'...	['newark', 'air', 'traffic', 'control', 'moment', 'system', '...	['may', 'feder', 'regul', 'now', 'limit', 'number', 'flight', '...	['newark', 'air', 'traffic', 'controller', 'moment', 'system',...	['may', 'federal', 'regulator', 'now', 'limit', 'numb', 'fligh...
battery', 'race', 'china', 'us', 'compete', 'future'...	['may', 'car', 'drive', 'years', 'future', 'might', 'run', 'batt...	['great', 'batteri', 'race', 'china', 'us', 'compet', 'futur', 'e...	['may', 'car', 'drive', 'year', 'futur', 'might', 'run', 'batteri...	['great', 'battery', 'race', 'china', 'us', 'compete', 'future'...	['may', 'car', 'drive', 'year', 'future', 'may', 'run', 'battery...
'slow', 'political', 'spending', 'think', 'done', 'e...	['may', 'billionaire', 'executive', 'trumps', 'biggest', 'do...	['musk', 'slow', 'polit', 'spend', 'think', 'done', 'enough']	['may', 'billionair', 'execut', 'trump', 'biggest', 'donor', '...	['musk', 'slow', 'political', 'spend', 'think', 'do', 'enough']	['may', 'billionaire', 'executive', 'trump', 'big', 'donor', '...
', 'ends', 'die', 'policies', 'get', 'fcc', 'blessing', '...	['may', 'latest', 'big', 'company', 'back', 'away', 'diversit'...	['verizon', 'end', 'die', 'polici', 'get', 'fcc', 'bless', 'billion...	['may', 'latest', 'big', 'compani', 'back', 'away', 'divers', '...	['verizon', 'end', 'die', 'policy', 'get', 'fcc', 'bless', 'billio...	['may', 'late', 'big', 'company', 'back', 'away', 'diversity',...
watchdog', 'opens', 'investigation', 'doge', 'cla...	['may', 'doge', 'employees', 'demanded', 'highest', 'leve...	['labor', 'watchdog', 'open', 'investig', 'doge', 'claim', 'n...	['may', 'doge', 'employe', 'demand', 'highest', 'level', 'a...	['labor', 'watchdog', 'open', 'investigation', 'doge', 'clai...	['may', 'doge', 'employee', 'demand', 'high', 'level', 'acc...
'urder', 'artificial', 'intelligence', 'forgiveness']	['may', 'ai', 'give', 'voice', 'even', 'murdered']	['tale', 'murder', 'artifici', 'intellig', 'forgiv']	['may', 'ai', 'give', 'voic', 'even', 'murder']	['tale', 'murder', 'artificial', 'intelligence', 'forgiveness']	['may', 'ai', 'give', 'voice', 'even', 'murder']
'are', 'foreign', 'policy', 'experts', 'think']	['may', 'large', 'language', 'models', 'like', 'chatting', 'in...	['ai', 'futur', 'foreign', 'polici', 'expert', 'think']	['may', 'larg', 'languag', 'model', 'like', 'chat', 'increas', '...	['ai', 'future', 'foreign', 'policy', 'expert', 'think']	['may', 'large', 'language', 'model', 'like', 'chat', 'increas...
th', 'partner', 'raises', 'millions', 'new', 'biotech...	['may', 'incarcerated', 'former', 'silicon', 'valley', 'star', '...	['elizabeth', 'partner', 'rais', 'million', 'new', 'biotech', 't...	['may', 'incarcer', 'former', 'silicon', 'valley', 'star', 'advi...	['elizabeth', 'partner', 'raise', 'million', 'new', 'biotech', ...	['may', 'incarcerate', 'former', 'silicon', 'valley', 'star', 'a...
, 'pay', 'b', 'settle', 'claims', 'user', 'data', 'colle...	['may', 'agreement', 'settles', 'several', 'claims', 'made',...	['googl', 'pay', 'b', 'settl', 'claim', 'user', 'data', 'collect']	['may', 'agreement', 'settl', 'sever', 'claim', 'made', 'sear...	['google', 'pay', 'b', 'settle', 'claim', 'user', 'datum', 'coll...	['may', 'agreement', 'settle', 'several', 'claim', 'make', 's...
'tightens', 'control', 'independent', 'agency', '...	['may', 'npr', 'learned', 'rules', 'must', 'now', 'vetted', 'w...	['trump', 'tighten', 'control', 'independ', 'agenc', 'overs...	['may', 'npr', 'learn', 'rule', 'must', 'now', 'vet', 'white', '...	['trump', 'tighten', 'control', 'independent', 'agency', 'o...	['may', 'npr', 'learn', 'rule', 'must', 'now', 'vet', 'white', '...
ompanies', 'shrink', 'aid', 'climate', 'footprint']	['may', 'qoogle', 'meta', 'pledged', 'reach', 'least', 'net', ...	['tech', 'compani', 'shrink', 'aid', 'climat', 'footprint']	['may', 'qoogl', 'meta', 'pledq', 'reach', 'least', 'net', 'zer...	['tech', 'company', 'shrink', 'aid', 'climate', 'footprint']	['may', 'qoogle', 'meta', 'pledge', 'reach', 'less', 'net', 'z...

Showing 1 to 38 of 500 entries, 16 total columns

PART-2

Cleaning and Finalizing Lemmatized Descriptions:

This code defines a cleaning function to remove brackets and quotes from the lemmatized description text and convert it to lowercase for uniformity. The cleaning function is applied to the "description_lemma_only" column, creating a new cleaned column "clean_description_lemma".

INPUT:

```
lemma_df <- read_csv("/Users/mithizaman/Documents/12/Introduction to data science/code/final/npr_stem_lemma_separate.csv")


clean_column <- function(text) {
  text <- gsub("\\[|\\]|'", "", text)  # Remove brackets and quotes
  text <- tolower(text)                # Convert to lowercase
  return(text)
}


lemma_df <- lemma_df %>%
  mutate(clean_description_lemma = sapply(description_lemma_only, clean_column))


View(lemma_df)


write_csv(lemma_df, "/Users/mithizaman/Documents/12/Introduction to data science/code/final/npr_lemma_cleaned_final.csv")
```

OUTPUT:

ion_no_stopwords	title_stem_only	description_stem_only	title_lemma_only	description_lemma_only	clean_description_lemma
eries', 'executive', 'orders', 'aims', 'promote', '...	['trump', 'seek', 'boost', 'nuclear', 'industri', 'overhaul', ...	['may', 'seri', 'execut', 'order', 'aim', 'promot', 'new', 'ki...	['trump', 'seek', 'boost', 'nuclear', 'industry', 'overhaul',...	['may', 'series', 'executive', 'order', 'aim', 'promote', 'ne...	may, series, executive, order, aim, promote, new, kin...
hroughout', 'three', 'years', 'since', 'invasion', '...	['insid', 'drone', 'factori', 'ukrain']	['may', 'throughout', 'three', 'year', 'sinc', 'invas', 'ukrai...	['inside', 'drone', 'factory', 'ukraine']	['may', 'throughout', 'three', 'year', 'since', 'invasion', '...	may, throughout, three, year, since, invasion, ukraine...
illion', 'deal', 'brings', 'together', 'maker', 'cha...	['open', 'forg', 'deal', 'iphon', 'design', 'joy', 'vie', 'make...	['may', 'billion', 'deal', 'bring', 'togeth', 'maker', 'chat', '...	['open', 'forge', 'deal', 'iphone', 'designer', 'joy', 'vie', '...	['may', 'billion', 'deal', 'bring', 'together', 'maker', 'chat'...	may, billion, deal, bring, together, maker, chat, one, ...
resident', 'trump', 'hosting', 'exclusive', 'dinn...	['rais', 'ethic', 'question', 'top', 'trump', 'meme', 'coin', '...	['may', 'presid', 'trump', 'host', 'exclus', 'dinner', 'tonig...	['raise', 'ethic', 'question', 'top', 'trump', 'meme', 'coin',...	['may', 'president', 'trump', 'host', 'exclusive', 'dinner', ...	may, president, trump, host, exclusive, dinner, tonigh...
ederal', 'regulators', 'now', 'limiting', 'number'...	['newark', 'air', 'traffic', 'control', 'moment', 'system', '...	['may', 'feder', 'regul', 'now', 'limit', 'number', 'flight', '...	['newark', 'air', 'traffic', 'controller', 'moment', 'system',...	['may', 'federal', 'regulator', 'now', 'limit', 'numb', 'fligh...	may, federal, regulator, now, limit, numb, flight, new...
air', 'drive', 'years', 'future', 'might', 'run', 'batt...	['great', 'batteri', 'race', 'china', 'us', 'compet', 'futur', 'e...	['may', 'car', 'drive', 'year', 'futur', 'might', 'run', 'batteri...	['great', 'battery', 'race', 'china', 'us', 'compete', 'future'...	['may', 'car', 'drive', 'year', 'future', 'may', 'run', 'battery...	may, car, drive, year, future, may, run, battery, invent...
illionaire', 'executive', 'trumps', 'biggest', 'do...	['musk', 'slow', 'polit', 'spend', 'think', 'done', 'enough']	['may', 'billionair', 'execut', 'trump', 'biggest', 'donor', '...	['musk', 'slow', 'political', 'spend', 'think', 'do', 'enough']	['may', 'billionaire', 'executive', 'trump', 'big', 'donor', '...	may, billionaire, executive, trump, big, donor, now, h...
atest', 'big', 'company', 'back', 'away', 'diversit...	['verizon', 'end', 'die', 'polici', 'get', 'fcc', 'bless', 'billion...	['may', 'latest', 'big', 'compani', 'back', 'away', 'divers', '...	['verizon', 'end', 'die', 'policy', 'get', 'fcc', 'bless', 'billio...	['may', 'late', 'big', 'company', 'back', 'away', 'diversity',...	may, late, big, company, back, away, diversity, pledge...
oge', 'employees', 'demanded', 'highest', 'leve...	['labor', 'watchdog', 'open', 'investig', 'doge', 'claim', 'n...	['may', 'doge', 'employe', 'demand', 'highest', 'level', 'a...	['labor', 'watchdog', 'open', 'investigation', 'doge', 'clai...	['may', 'doge', 'employee', 'demand', 'high', 'level', 'acc...	may, doge, employee, demand, high, level, access, la...
i', 'give', 'voice', 'even', 'murdered']	['tale', 'murder', 'artifici', 'intelligi', 'forgiv']	['may', 'ai', 'give', 'voic', 'even', 'murder']	['tale', 'murder', 'artificial', 'intelligence', 'forgiveness']	['may', 'ai', 'give', 'voice', 'even', 'murder']	may, ai, give, voice, even, murder
arge', 'language', 'models', 'like', 'chatting', 'in...	['ai', 'futur', 'foreign', 'polici', 'expert', 'think']	['may', 'larg', 'languag', 'model', 'like', 'chat', 'increas', '...	['ai', 'future', 'foreign', 'policy', 'expert', 'think']	['may', 'large', 'language', 'model', 'like', 'chat', 'increas...	may, large, language, model, like, chat, increasingly, l...
ncarcerated', 'former', 'silicon', 'valley', 'star', '...	['elizabeth', 'partner', 'rais', 'million', 'new', 'biotech', 't...	['may', 'incarcer', 'former', 'silicon', 'valley', 'star', 'advi...	['elizabeth', 'partner', 'raise', 'million', 'new', 'biotech', ...	['may', 'incarcerate', 'former', 'silicon', 'valley', 'star', 'a...	may, incarcerate, former, silicon, valley, star, advise, ...
greement', 'settles', 'several', 'claims', 'made',...	['googl', 'pay', 'b', 'settl', 'claim', 'user', 'data', 'collect']	['may', 'agreement', 'settl', 'sever', 'claim', 'made', 'sear...	['google', 'pay', 'b', 'settle', 'claim', 'user', 'datum', 'coll...	['may', 'agreement', 'settle', 'several', 'claim', 'make', 's...	may, agreement, settle, several, claim, make, search, ...
pr', 'learned', 'rules', 'must', 'now', 'vetted', 'w...	['trump', 'tighten', 'control', 'independ', 'agenc', 'overs...	['may', 'npr', 'learn', 'rule', 'must', 'now', 'vet', 'white', '...	['trump', 'tighten', 'control', 'independent', 'agency', 'o...	['may', 'npr', 'learn', 'rule', 'must', 'now', 'vet', 'white', '...	may, npr, learn, rule, must, now, vet, white, house, a...
oogle', 'meta', 'pledged', 'reach', 'least', 'net', ...	['tech', 'compani', 'shrink', 'aid', 'climat', 'footprint']	['may', 'googl', 'meta', 'pledg', 'reach', 'least', 'net', 'zer...	['tech', 'company', 'shrink', 'aid', 'climate', 'footprint']	['may', 'google', 'meta', 'pledge', 'reach', 'less', 'net', 'z...	may, google, meta, pledge, reach, less, net, zero, car...
ecades', 'philanthropy', 'following', 'success', ...	['bill', 'gate', 'give', 'away', 'remain', 'fortun', 'charit']	['may', 'decad', 'philanthropi', 'follow', 'success', 'bill', '...	['bill', 'gate', 'give', 'away', 'remain', 'fortune', 'charity']	['may', 'decade', 'philanthropy', 'follow', 'success', 'bill'...	may, decade, philanthropy, follow, success, bill, gate,...
manda', 'learned', 'unborn', 'child', 'genetic', '...	['app', 'gadget', 'second', 'life', 'consid', 'tech', 'chang', ...	['may', 'amanda', 'learn', 'unborn', 'child', 'genet', 'con...	['app', 'gadget', '2', 'life', 'consider', 'tech', 'change', 'b...	['may', 'amanda', 'learn', 'unborn', 'child', 'genetic', 'co...	may, amanda, learn, unborn, child, genetic, condition...
resident', 'trump', 'proposed', 'slashing', 'fed...	['economist', 'warn', 'trump', 'research', 'cut', 'dire', 'co...	['may', 'presid', 'trump', 'propos', 'slash', 'feder', 'scien...	['economist', 'warn', 'trump', 'research', 'cut', 'dire', 'co...	['may', 'president', 'trump', 'propose', 'slash', 'federal', ...	may, president, trump, propose, slash, federal, scient...
ewark', 'liberty', 'international', 'airport', 'mes...	['air', 'traffic', 'control', 'problem']	['may', 'newark', 'liberti', 'intern', 'airport', 'mess', 'past...	['air', 'traffic', 'control', 'problem']	['may', 'newark', 'liberty', 'international', 'airport', 'mes...	may, newark, liberty, international, airport, mess, pas...
pr', 'scott', 'dethrone', 'speaks', 'ceo', 'hallow',...	['cathol', 'prayer', 'app', 'one', 'popular', 'world', 'take', ...	['may', 'npr', 'scott', 'dethron', 'speak', 'ceo', 'hallow', 'c...	['catholic', 'prayer', 'app', 'one', 'popular', 'world', 'take...	['may', 'npr', 'scott', 'dethrone', 'speak', 'ceo', 'hallow', ...	may, npr, scott, dethrone, speak, ceo, hallow, catholi...
i', 'experts', 'say', 'likely', 'first', 'time', 'ai', 'us...	['famili', 'show', 'ai', 'video', 'slain', 'victim', 'impact', 'st...	['may', 'ai', 'expert', 'say', 'like', 'first', 'time', 'ai', 'use', ...	['family', 'show', 'ai', 'video', 'slay', 'victim', 'impact', 'st...	['may', 'ai', 'expert', 'say', 'likely', 'first', 'time', 'ai', 'use'...	may, ai, expert, say, likely, first, time, ai, use, us, crea...
awrence', 'berkeley', 'national', 'laboratory', 'fo...	['true', 'water', 'footprint', 'ai', 'elus']	['may', 'lawrenc', 'berkeley', 'nation', 'laboratori', 'forec...	['true', 'water', 'footprint', 'ai', 'elusive']	['may', 'lawrence', 'berkeley', 'national', 'laboratory', 'fo...	may, lawrence, berkeley, national, laboratory, forecas...
aid', 'critical', 'service', 'provider', 'terminated'...	['major', 'deep', 'fake', 'porn', 'site', 'shut']	['may', 'said', 'critic', 'servic', 'provid', 'termin', 'servic', ...	['major', 'deep', 'fake', 'porn', 'site', 'shut']	['may', 'say', 'critical', 'service', 'provider', 'terminate', '...	may, say, critical, service, provider, terminate, service...
are', 'rebuke', 'dozen', 'former', 'workers', 'po...	['former', 'palati', 'worker', 'condemn', 'compani', 'wor...	['may', 'rare', 'rebuk', 'dozen', 'former', 'worker', 'powe...	['former', 'palatial', 'worker', 'condemn', 'company', 'w...	['may', 'rare', 'rebuke', 'dozen', 'former', 'worker', 'pow...	may, rare, rebuke, dozen, former, worker, powerful, d...
eries', 'executive', 'orders', 'aims', 'promote', '...	['trump', 'seek', 'boost', 'nuclear', 'industri', 'overhaul', ...	['may', 'seri', 'execut', 'order', 'aim', 'promot', 'new', 'ki...	['trump', 'seek', 'boost', 'nuclear', 'industry', 'overhaul',...	['may', 'series', 'executive', 'order', 'aim', 'promote', 'ne...	may, series, executive, order, aim, promote, new, kin...
hroughout', 'three', 'years', 'since', 'invasion', '...	['insid', 'drone', 'factori', 'ukrain']	['may', 'throughout', 'three', 'year', 'sinc', 'invas', 'ukrai...	['inside', 'drone', 'factory', 'ukraine']	['may', 'throughout', 'three', 'year', 'since', 'invasion', '...	may, throughout, three, year, since, invasion, ukraine...
illion', 'deal', 'brings', 'together', 'maker', 'cha...	['open', 'forg', 'deal', 'iphon', 'design', 'joy', 'vie', 'make...	['may', 'billion', 'deal', 'bring', 'togeth', 'maker', 'chat', '...	['open', 'forge', 'deal', 'iphone', 'designer', 'joy', 'vie', '...	['may', 'billion', 'deal', 'bring', 'together', 'maker', 'chat'...	may, billion, deal, bring, together, maker, chat, one, ...
resident', 'trump', 'hosting', 'exclusive', 'dinn...	['rais', 'ethic', 'question', 'top', 'trump', 'meme', 'coin', '...	['may', 'presid', 'trump', 'host', 'exclus', 'dinner', 'tonig...	['raise', 'ethic', 'question', 'top', 'trump', 'meme', 'coin',...	['may', 'president', 'trump', 'host', 'exclusive', 'dinner', ...	may, president, trump, host, exclusive, dinner, tonigh...
ederal', 'regulators', 'now', 'limiting', 'number'...	['newark', 'air', 'traffic', 'control', 'moment', 'system', '...	['may', 'feder', 'regul', 'now', 'limit', 'number', 'flight', '...	['newark', 'air', 'traffic', 'controller', 'moment', 'system',...	['may', 'federal', 'regulator', 'now', 'limit', 'numb', 'fligh...	may, federal, regulator, now, limit, numb, flight, new...
air', 'drive', 'years', 'future', 'might', 'run', 'batt...	['great', 'batteri', 'race', 'china', 'us', 'compet', 'futur', 'e...	['may', 'car', 'drive', 'year', 'futur', 'might', 'run', 'batteri...	['great', 'battery', 'race', 'china', 'us', 'compete', 'future'...	['may', 'car', 'drive', 'year', 'future', 'may', 'run', 'battery...	may, car, drive, year, future, may, run, battery, invent...
illionaire', 'executive', 'trumps', 'biggest', 'do...	['musk', 'slow', 'polit', 'spend', 'think', 'done', 'enough']	['may', 'billionair', 'execut', 'trump', 'biggest', 'donor', '...	['musk', 'slow', 'political', 'spend', 'think', 'do', 'enough']	['may', 'billionaire', 'executive', 'trump', 'big', 'donor', '...	may, billionaire, executive, trump, big, donor, now, h...
atest', 'big', 'company', 'back', 'away', 'diversit...	['verizon', 'end', 'die', 'polici', 'get', 'fcc', 'bless', 'billion...	['may', 'latest', 'big', 'compani', 'back', 'away', 'divers', '...	['verizon', 'end', 'die', 'policy', 'get', 'fcc', 'bless', 'billio...	['may', 'late', 'big', 'company', 'back', 'away', 'diversity',...	may, late, big, company, back, away, diversity, pledge...
oge', 'employees', 'demanded', 'highest', 'leve...	['labor', 'watchdog', 'open', 'investig', 'doge', 'claim', 'n...	['may', 'doge', 'employe', 'demand', 'highest', 'level', 'a...	['labor', 'watchdog', 'open', 'investigation', 'doge', 'clai...	['may', 'doge', 'employee', 'demand', 'high', 'level', 'acc...	may, doge, employee, demand, high, level, access, la...
i', 'give', 'voice', 'even', 'murdered']	['tale', 'murder', 'artifici', 'intelligi', 'forgiv']	['may', 'ai', 'give', 'voic', 'even', 'murder']	['tale', 'murder', 'artificial', 'intelligence', 'forgiveness']	['may', 'ai', 'give', 'voice', 'even', 'murder']	may, ai, give, voice, even, murder
arge', 'language', 'models', 'like', 'chatting', 'in...	['ai', 'futur', 'foreign', 'polici', 'expert', 'think']	['may', 'larg', 'languag', 'model', 'like', 'chat', 'increas', '...	['ai', 'future', 'foreign', 'policy', 'expert', 'think']	['may', 'large', 'language', 'model', 'like', 'chat', 'increas...	may, large, language, model, like, chat, increasingly, l...
ncarcerated', 'former', 'silicon', 'valley', 'star', '...	['elizabeth', 'partner', 'rais', 'million', 'new', 'biotech', 't...	['may', 'incarcer', 'former', 'silicon', 'valley', 'star', 'advi...	['elizabeth', 'partner', 'raise', 'million', 'new', 'biotech', ...	['may', 'incarcerate', 'former', 'silicon', 'valley', 'star', 'a...	may, incarcerate, former, silicon, valley, star, advise, ...
greement', 'settles', 'several', 'claims', 'made',...	['googl', 'pay', 'b', 'settl', 'claim', 'user', 'data', 'collect']	['may', 'agreement', 'settl', 'sever', 'claim', 'made', 'sear...	['google', 'pay', 'b', 'settle', 'claim', 'user', 'datum', 'coll...	['may', 'agreement', 'settle', 'several', 'claim', 'make', 's...	may, agreement, settle, several, claim, make, search, ...
pr', 'learned', 'rules', 'must', 'now', 'vetted', 'w...	['trump', 'tighten', 'control', 'independ', 'agenc', 'overs...	['may', 'npr', 'learn', 'rule', 'must', 'now', 'vet', 'white', '...	['trump', 'tighten', 'control', 'independent', 'agency', 'o...	['may', 'npr', 'learn', 'rule', 'must', 'now', 'vet', 'white', '...	may, npr, learn, rule, must, now, vet, white, house, a...
oogle', 'meta', 'pledged', 'reach', 'least', 'net', ...	['tech', 'compani', 'shrink', 'aid', 'climat', 'footprint']	['may', 'googl', 'meta', 'pledg', 'reach', 'least', 'net', 'zer...	['tech', 'company', 'shrink', 'aid', 'climate', 'footprint']	['may', 'google', 'meta', 'pledge', 'reach', 'less', 'net', 'z...	may, google, meta, pledge, reach, less, net, zero, car...

Showing 1 to 38 of 500 entries, 17 total columns

Generating Document-Term Matrix (DTM):

This part takes the cleaned text and turns it into a simple word count table (DTM). It:
Prepares the text (fixes whitespaces, converting to lowercase, removes punctions and numbers).Counts how often each word appears in each document. Saves the final table with the original data as a CSV file (lemma_dtm_combined_final.csv).

INPUT:

```
corpus <- Corpus(VectorSource(lemma_df$clean_description_lemma))

corpus <- tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeNumbers)

dtm <- DocumentTermMatrix(corpus)

dtm_sparse <- removeSparseTerms(dtm, 0.99)

dtm_matrix <- as.matrix(dtm_sparse)
dtm_df <- as.data.frame(dtm_matrix)

lemma_dtm_combined <- bind_cols(lemma_df, dtm_df)

View(lemma_dtm_combined)
write_csv(lemma_dtm_combined, "/Users/mithizaman/Documents/12/Introduction to data science/code/final/lemma_dtm_combined_final.csv")
```

OUTPUT:

aim	body	charge	executive	kind	may	new	nuclear	order	promote	reactor	restructure	safety	series	aerial	churn	come	conflict	cut	drone	edge	eleanor	episode	factory	fly	good	high	innovation
1	1	1	1	1	1	1	2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	4	1	1	1	1	1	1	1	
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	1	1	1	1	1	1	2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	4	1	1	1	1	1	1	1	
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Showing 1 to 38 of 500 entries, 421 total columns

LDA Topic Modeling and Term Visualization:

The tool reads the word counts (DTM) and finds 5 hidden themes. For each topic, it picks the 10 most important words. A bar graph shows these words, making it easy to see what each topic is about. The top words for each topic are saved in (lda_top_terms.csv).

INPUT:

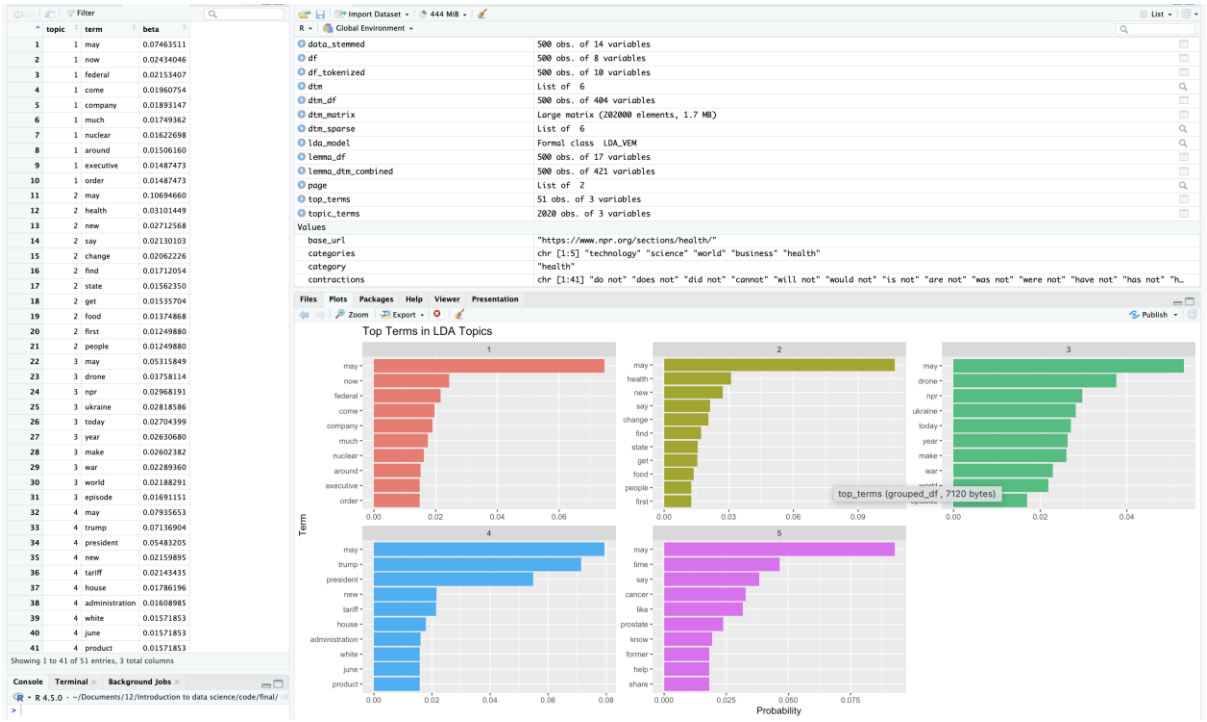
```
set.seed(1234)
num_topics <- 5
lda_model <- LDA(dtm_sparse, k = num_topics, control = list(seed = 1234))

topic_terms <- tidy(lda_model, matrix = "beta")
top_terms <- topic_terms %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  arrange(topic, -beta)
View(top_terms)

top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered() +
  labs(
    title = "Top Terms in LDA Topics",
    x = "Term", y = "Probability"
  )

write.csv(top_terms, "/Users/mithizaman/Documents/12/Introduction to data science/code/final/lda_top_terms.csv", row.names = FALSE)
```

OUTPUT:


Top Terms in LDA Topics
Term
Probability
top_terms (grouped_df , 7120 bytes)