AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Course Title: INTRODUCTION TO DATA SCIENCE

Course Code: CSC4180 Section: E

Semester: Spring 2024-2025 Course Teacher: DR. ABDUS SALAM

Group 6

No	Name	ID	Program
1	SK. Shahed Ali	22-47756-2	BSc in CSE
2	S M Abid Hasan	22-46789-1	BSc in CSE
3	MD. Faysal Kabir	22-46783-1	BSc in CSE
4	Noymul Alam	22-47696-2	BSc in CSE

Faculty use only

FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Text Processing and Topic Modeling

In this project, we first scraped 500 news articles from the bdnews24.com portal across 5 categories: entertainment, business, sports, politics, and technology. Each article includes the title, description, date, and category, and was stored in a CSV file.
For text processing, we applied standard NLP preprocessing steps in R. The news descriptions were first lowercased, followed by the removal of punctuation, numbers, extra whitespaces, and stop words. Contractions like "can't" and "won't" were expanded to their full forms.
After cleaning, the text was tokenized and reconstructed into strings. Then, stemming and lemmatization were applied using corpus-based methods. The cleaned and processed text was saved in a final CSV file for topic modeling.
After preprocessing, Latent Dirichlet Allocation (LDA) was used for topic modeling to uncover hidden themes across the articles. We applied LDA with k = 5 topics on the Document-Term Matrix created from lemmatized text. The top 10 words and most probable word from each topic were extracted for interpretation

- ## Part-1

Scraping News Text from a URL

Code:

```
url <- "https://bdnews24.com/business/766991fff2aa"
webpage <- read_html(url)

title <- html_text(html_node(webpage, "h1"))
description <- html_text(html_node(webpage, "div.details-
brief.dNewsDesc.print-section"))
date_span <- html_nodes(webpage, "div.pub-up p span")[2]
date <- html_text(date_span)
category <- unlist(strsplit(url, "/"))[4]
```

OUTPUT

```
> title <- html_text(html_node(webpage, "h1"))
> title
[1] " India's polished diamond exports hit two-decade low, industry group says "
>
```

```
> description <- html_text(html_node(webpage, "div.details-brief.dNewsDesc.print-section"))
> description
[1] "\n                        \n                        \n                        India's exports of cut and polished
diamonds plummeted to their lowest level in nearly two decades in the 2024/25 fiscal year, which ended in March, on slu
ggish demand from the United States and China, a leading trade body said on Monday.\r\n\r\n \r\n\r\nIndia is the worl
d's largest cutting and polishing hub, handling nine out of every 10 diamonds processed globally. But it is sensitive t
o economic uncertainty - particularly in the US, its biggest market.\r\n\r\n \r\n\r\nCut and polished diamond exports,
which usually account for nearly half of overall gem and jewellery shipments, fell 16.8% to $13.3 billion year-on-year,
the Gems and Jewellery Export Promotion Council (GJEPC) said in a statement.\r\n\r\n \r\n\r\nThe slump dragged down ove
rall gem and jewellery exports by 11.7% to $28.5 billion - a four-year low - from $32.28 billion the previous year.\r\n
\r\n \r\n\r\nThe lower demand for polished diamonds also prompted Indian processors to reduce imports of rough diamonds
by 24.3% to $10.8 billion, the trade body said.\r\n\r\n \r\n\r\nGems and jewellery exports rose by 1% year-on-year in M
arch, however, to $2.56 billion, the GJEPC said, as exporters ramped up shipments ahead of announced US tariffs.\r\n\r
\n \r\n\r\nUS President Donald Trump initially planned to place a 27% tariff on imported Indian goods from April 9 as p
art of duties targeting dozens of countries, but then declared a 90-day pause on the measure.\r\n\r\n \r\n\r\n \r\n\r\n
\"US buyers were loading up in March before the tariffs kicked in. Indian exporters were also rushing to ship out US or
ders first, so they wouldn't get hit with those extra costs,\" said Shaunak Parikh, vice-chairman of GJEPC.\r\n\r\n \r
\n\r\nIndia's gems and jewellery exports are unlikely to recover this year, one major Mumbai-based exporter told Reuter
s, as the US tariffs have roiled global markets and shaken buyer confidence.\n                        "
>
```

```
> date_span <- html_nodes(webpage, "div.pub-up p span")[2]
> date <- html_text(date_span)
> date
[1] "15 Apr 2025, 08:28 AM"
>
```

```
> category <- unlist(strsplit(url, "/"))[4]
> category
[1] "business"
>
```

The news content extracted from bdnews24.com using the **rvest** package in R. Key elements were programmatically retrieved from the HTML structure of the web page through the use of CSS selectors. The news title was extracted from the **<h1>** tag, the news description was obtained from a ** ** element with the class **details-brief dNewsDesc print-section**, and the publication date was captured from the second **** tag within a ** ** container. The news category was identified by parsing the **URL** and selecting the relevant segment.

Saving News Text to CSV

CODE

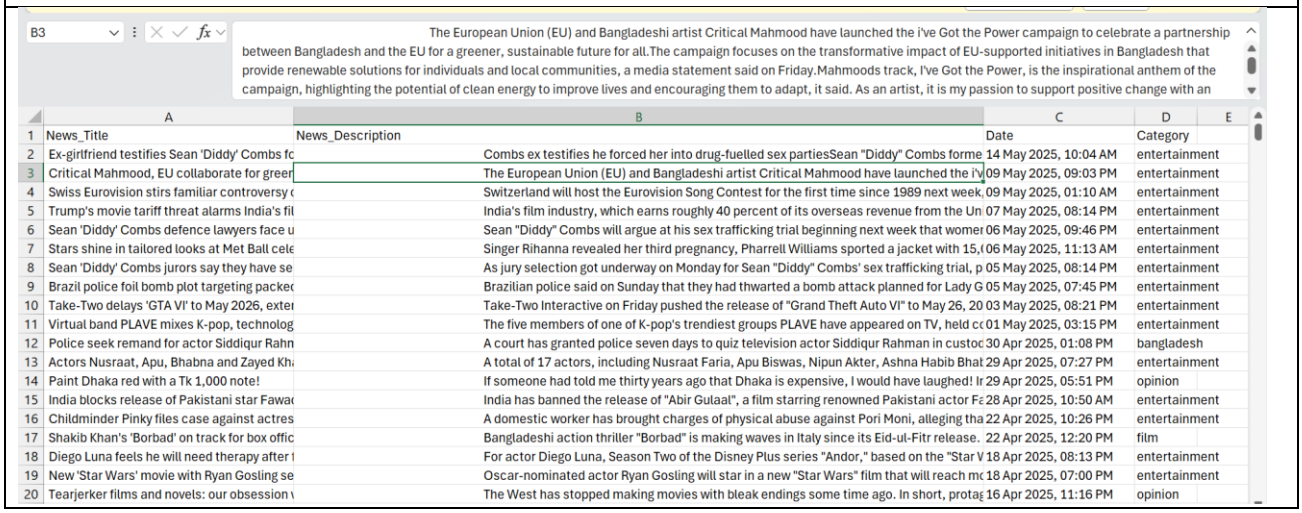
```
news_data <- data.frame(
  News_Title = title,
  News_Description = description,
  Date = date,
  Category = category,
  stringsAsFactors = FALSE
)

write.table(news_data, "ids_final_project_group_06_news_raw.csv",
            sep = ",", row.names = FALSE,
            col.names = FALSE, append = TRUE,
            quote = TRUE, fileEncoding = "UTF-8", qmethod = "double")
```

OUTPUT

```
> news_data
                                                                News_Title
1  India's polished diamond exports hit two-decade low, industry group says

News_Description
1 \n                          \n                          \n                          India's exports of cut and polished dia
monds plummeted to their lowest level in nearly two decades in the 2024/25 fiscal year, which ended in March, on sluggi
sh demand from the United States and China, a leading trade body said on Monday.\r\n\r\n \r\n\r\nIndia is the world's l
argest cutting and polishing hub, handling nine out of every 10 diamonds processed globally. But it is sensitive to eco
nomic uncertainty - particularly in the US, its biggest market.\r\n\r\n \r\n\r\nCut and polished diamond exports, which
usually account for nearly half of overall gem and jewellery shipments, fell 16.8% to $13.3 billion year-on-year, the G
ems and Jewellery Export Promotion Council (GJEPC) said in a statement.\r\n\r\n \r\n\r\nThe slump dragged down overall
gem and jewellery exports by 11.7% to $28.5 billion - a four-year low - from $32.28 billion the previous year.\r\n\r\n
\r\n\r\nThe lower demand for polished diamonds also prompted Indian processors to reduce imports of rough diamonds by 2
4.3% to $10.8 billion, the trade body said.\r\n\r\n \r\n\r\nGems and jewellery exports rose by 1% year-on-year in Marc
h, however, to $2.56 billion, the GJEPC said, as exporters ramped up shipments ahead of announced US tariffs.\r\n\r\n
\r\n\r\nUS President Donald Trump initially planned to place a 27% tariff on imported Indian goods from April 9 as part
of duties targeting dozens of countries, but then declared a 90-day pause on the measure.\r\n\r\n \r\n\r\n \r\n\r\n"US
buyers were loading up in March before the tariffs kicked in. Indian exporters were also rushing to ship out US orders
first, so they wouldn't get hit with those extra costs," said Shaunak Parikh, vice-chairman of GJEPC.\r\n\r\n \r\n\r\nI
ndia's gems and jewellery exports are unlikely to recover this year, one major Mumbai-based exporter told Reuters, as t
he US tariffs have roiled global markets and shaken buyer confidence.\n
                  Date Category
1 15 Apr 2025, 08:28 AM business
>
```



B3 — The European Union (EU) and Bangladeshi artist Critical Mahmood have launched the i've Got the Power campaign to celebrate a partnership between Bangladesh and the EU for a greener, sustainable future for all.The campaign focuses on the transformative impact of EU-supported initiatives in Bangladesh that provide renewable solutions for individuals and local communities, a media statement said on Friday.Mahmoods track, I've Got the Power, is the inspirational anthem of the campaign, highlighting the potential of clean energy to improve lives and encouraging them to adapt, it said. As an artist, it is my passion to support positive change with an

	A	B	C	D
1	News_Title	News_Description	Date	Category
2	Ex-girlfriend testifies Sean 'Diddy' Combs fc	Combs ex testifies he forced her into drug-fuelled sex partiesSean "Diddy" Combs forme	14 May 2025, 10:04 AM	entertainment
3	Critical Mahmood, EU collaborate for greer	The European Union (EU) and Bangladeshi artist Critical Mahmood have launched the i'v	09 May 2025, 09:03 PM	entertainment
4	Swiss Eurovision stirs familiar controversy (	Switzerland will host the Eurovision Song Contest for the first time since 1989 next week,	09 May 2025, 01:10 AM	entertainment
5	Trump's movie tariff threat alarms India's fil	India's film industry, which earns roughly 40 percent of its overseas revenue from the Un	07 May 2025, 08:14 PM	entertainment
6	Sean 'Diddy' Combs defence lawyers face u	Sean "Diddy" Combs will argue at his sex trafficking trial beginning next week that womer	06 May 2025, 09:46 PM	entertainment
7	Stars shine in tailored looks at Met Ball cele	Singer Rihanna revealed her third pregnancy, Pharrell Williams sported a jacket with 15,(	06 May 2025, 11:13 AM	entertainment
8	Sean 'Diddy' Combs jurors say they have se	As jury selection got underway on Monday for Sean "Diddy" Combs' sex trafficking trial, p	05 May 2025, 08:14 PM	entertainment
9	Brazil police foil bomb plot targeting packec	Brazilian police said on Sunday that they had thwarted a bomb attack planned for Lady G	05 May 2025, 07:45 PM	entertainment
10	Take-Two delays 'GTA VI' to May 2026, exter	Take-Two Interactive on Friday pushed the release of "Grand Theft Auto VI" to May 26, 20	03 May 2025, 08:21 PM	entertainment
11	Virtual band PLAVE mixes K-pop, technolog	The five members of one of K-pop's trendiest groups PLAVE have appeared on TV, held cc	01 May 2025, 03:15 PM	entertainment
12	Police seek remand for actor Siddiqur Rahn	A court has granted police seven days to quiz television actor Siddiqur Rahman in custoc	30 Apr 2025, 01:08 PM	bangladesh
13	Actors Nusraat, Apu, Bhabna and Zayed Kh	A total of 17 actors, including Nusraat Faria, Apu Biswas, Nipun Akter, Ashna Habib Bhat	29 Apr 2025, 07:27 PM	entertainment
14	Paint Dhaka red with a Tk 1,000 note!	If someone had told me thirty years ago that Dhaka is expensive, I would have laughed! Ir	29 Apr 2025, 05:51 PM	opinion
15	India blocks release of Pakistani star Fawa	India has banned the release of "Abir Gulaal", a film starring renowned Pakistani actor Fa	28 Apr 2025, 10:50 AM	entertainment
16	Childminder Pinky files case against actres	A domestic worker has brought charges of physical abuse against Pori Moni, alleging tha	22 Apr 2025, 10:26 PM	entertainment
17	Shakib Khan's 'Borbad' on track for box offic	Bangladeshi action thriller "Borbad" is making waves in Italy since its Eid-ul-Fitr release.	22 Apr 2025, 12:20 PM	film
18	Diego Luna feels he will need therapy after	For actor Diego Luna, Season Two of the Disney Plus series "Andor," based on the "Star V	18 Apr 2025, 08:13 PM	entertainment
19	New 'Star Wars' movie with Ryan Gosling se	Oscar-nominated actor Ryan Gosling will star in a new "Star Wars" film that will reach mc	18 Apr 2025, 07:00 PM	entertainment
20	Tearjerker films and novels: our obsession \	The West has stopped making movies with bleak endings some time ago. In short, protag	16 Apr 2025, 11:16 PM	opinion

This code creates a data frame containing a single scraped news article, including its `News_Title`, `News_Description`, `Date`, and `Category`. It then appends this data to the existing ids_final_project_group_06_news_raw.csv file without rewriting the header, ensuring all articles are stored in one file.

Reading the Collected Dataset & Counting Total Number of Contractions

CODE

```
mydata <-
read.csv("C:/Users/Admin/Desktop/ids_final_project_group_06_news_raw.c
sv", header = TRUE, sep = ",", stringsAsFactors = FALSE)

sum(grepl("\\b(can't|won't|n't|I'm|it's|he's|she's|they're|I've|you've
|we're|you'd)\\b", mydata$News_Description, ignore.case = TRUE))
```

OUTPUT

```
> sum(grepl("\\b(can't|won't|n't|I'm|it's|he's|she's|they're|I've|you've|we're|you'd)
\\b",
+           mydata$News_Description, ignore.case = TRUE))
[1] 151
>
```

The dataset containing the scraped news articles is loaded from a CSV file. A regular expression is used to detect common English contractions within the News_Description column. The output indicated that 151 contractions were present, justifying the need to expand contractions in the text preprocessing phase.

Expanding Contractions & Corpus-Based Text Cleaning

CODE

```
expand_contractions <- function(text) {
  text <- gsub("\\bcan't\\b", "cannot", text, ignore.case = TRUE)
  text <- gsub("\\bwon't\\b", "will not", text, ignore.case = TRUE)
  text <- gsub("n't\\b", " not", text, ignore.case = TRUE)
  text <- gsub("\\bI'm\\b", "I am", text, ignore.case = TRUE)
  text <- gsub("\\bit's\\b", "it is", text, ignore.case = TRUE)
  text <- gsub("\\bhe's\\b", "he is", text, ignore.case = TRUE)
  text <- gsub("\\bshe's\\b", "she is", text, ignore.case = TRUE)
  text <- gsub("\\bthey're\\b", "they are", text, ignore.case = TRUE)
  text <- gsub("\\bI've\\b", "I have", text, ignore.case = TRUE)
  text <- gsub("\\byou've\\b", "you have", text, ignore.case = TRUE)
  text <- gsub("\\bwe're\\b", "we are", text, ignore.case = TRUE)
  text <- gsub("\\byou'd\\b", "you would", text, ignore.case = TRUE)
  text <- gsub("\\bthat's\\b", "that is", text, ignore.case = TRUE)
  text <- gsub("\\bthere's\\b", "there is", text, ignore.case = TRUE)
  text <- gsub("\\bwho's\\b", "who is", text, ignore.case = TRUE)
  text <- gsub("\\bwhat's\\b", "what is", text, ignore.case = TRUE)
  text <- gsub("\\bdoesn't\\b", "does not", text, ignore.case = TRUE)
  text <- gsub("\\bdidn't\\b", "did not", text, ignore.case = TRUE)
  text <- gsub("\\bcouldn't\\b", "could not", text, ignore.case = TRUE)
  text <- gsub("\\bshouldn't\\b", "should not", text, ignore.case =
TRUE)
  text <- gsub("\\bwouldn't\\b", "would not", text, ignore.case = TRUE)
  return(text)
}

corpus <- VCorpus(VectorSource(mydata$News_Description))
```

```
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, content_transformer(removePunctuation))
corpus <- tm_map(corpus, content_transformer(removeNumbers))
corpus <- tm_map(corpus, removeWords, stopwords("en"))
corpus <- tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, content_transformer(expand_contractions))

Cleaned_Text <- sapply(corpus, as.character)
```

OUTPUT

```
508

" indias exports cut polished diamonds plummeted lowest level nearly two decades fiscal
year ended march sluggish demand united states china leading trade body said monday ind
ia worlds largest cutting polishing hub handling nine every diamonds processed globally
sensitive economic uncertainty particularly us biggest market cut polished diamond expo
rts usually account nearly half overall gem jewellery shipments fell billion yearonyear
gems jewellery export promotion council gjepc said statement slump dragged overall gem
jewellery exports billion fouryear low billion previous year lower demand polished diam
onds also prompted indian processors reduce imports rough diamonds billion trade body s
aid gems jewellery exports rose yearonyear march however billion gjepc said exporters r
amped shipments ahead announced us tariffs us president donald trump initially planned
place tariff imported indian goods april part duties targeting dozens countries declare
d day pause measure us buyers loading march tariffs kicked indian exporters also rushin
g ship us orders first wouldnt get hit extra costs said shaunak parikh vicechairman gje
pc indias gems jewellery exports unlikely recover year one major mumbaibased exporter t
old reuters us tariffs roiled global markets shaken buyer confidence "
>
```

A custom function is defined to expand common English contractions into their full forms using regular expressions. The tm_map() function is used to preprocess the news descriptions by converting them to lowercase, removing punctuation, numbers, stop words, and extra whitespaces. The contraction expansion function is then applied to the cleaned corpus. The resulting cleaned text is extracted and stored in a character vector.

Tokenization and Reconstruction

CODE

```
Tokens <- tokenize_words(Cleaned_Text)
Tokens_Text <- sapply(Tokens, paste, collapse = " ")
```

OUTPUT

```
$`508`
  [1] "indias"       "exports"      "cut"          "polished"     "diamonds"     "plummeted"    "lowest"
  [8] "level"        "nearly"       "two"          "decades"      "fiscal"       "year"         "ended"
 [15] "march"        "sluggish"     "demand"       "united"       "states"       "china"        "leading"
 [22] "trade"        "body"         "said"         "monday"       "india"        "worlds"       "largest"
 [29] "cutting"      "polishing"    "hub"          "handling"     "nine"         "every"        "diamonds"
 [36] "processed"    "globally"     "sensitive"    "economic"     "uncertainty"  "particularly" "us"
 [43] "biggest"      "market"       "cut"          "polished"     "diamond"      "exports"      "usually"
 [50] "account"      "nearly"       "half"         "overall"      "gem"          "jewellery"    "shipments"
 [57] "fell"         "billion"      "yearonyear"   "gems"         "jewellery"    "export"       "promotion"
 [64] "council"      "gjepc"        "said"         "statement"    "slump"        "dragged"      "overall"
 [71] "gem"          "jewellery"    "exports"      "billion"      "fouryear"     "low"          "billion"
 [78] "previous"     "year"         "lower"        "demand"       "polished"     "diamonds"     "also"
 [85] "prompted"     "indian"       "processors"   "reduce"       "imports"      "rough"        "diamonds"
 [92] "billion"      "trade"        "body"         "said"         "gems"         "jewellery"    "exports"
 [99] "rose"         "yearonyear"   "march"        "however"      "billion"      "gjepc"        "said"
[106] "exporters"    "ramped"       "shipments"    "ahead"        "announced"    "us"           "tariffs"
[113] "us"           "president"    "donald"       "trump"        "initially"    "planned"      "place"
[120] "tariff"       "imported"     "indian"       "goods"        "april"        "part"         "duties"
[127] "targeting"    "dozens"       "countries"    "declared"     "day"          "pause"        "measure"
[134] "us"           "buyers"       "loading"      "march"        "tariffs"      "kicked"       "indian"
[141] "exporters"    "also"         "rushing"      "ship"         "us"           "orders"       "first"
[148] "wouldnt"      "get"          "hit"          "extra"        "costs"        "said"         "shaunak"
[155] "parikh"       "vicechairman" "gjepc"        "indias"       "gems"         "jewellery"    "exports"
[162] "unlikely"     "recover"      "year"         "one"          "major"        "mumbaibased"  "exporter"
[169] "told"         "reuters"      "us"           "tariffs"      "roiled"       "global"       "markets"
[176] "shaken"       "buyer"        "confidence"

> |
```

```
508

"indias exports cut polished diamonds plummeted lowest level nearly two decades fiscal year ended march sluggish demand
united states china leading trade body said monday india worlds largest cutting polishing hub handling nine every diamo
nds processed globally sensitive economic uncertainty particularly us biggest market cut polished diamond exports usual
ly account nearly half overall gem jewellery shipments fell billion yearonyear gems jewellery export promotion council
gjepc said statement slump dragged overall gem jewellery exports billion fouryear low billion previous year lower deman
d polished diamonds also prompted indian processors reduce imports rough diamonds billion trade body said gems jeweller
y exports rose yearonyear march however billion gjepc said exporters ramped shipments ahead announced us tariffs us pre
sident donald trump initially planned place tariff imported indian goods april part duties targeting dozens countries d
eclared day pause measure us buyers loading march tariffs kicked indian exporters also rushing ship us orders first wou
ldnt get hit extra costs said shaunak parikh vicechairman gjepc indias gems jewellery exports unlikely recover year one
major mumbaibased exporter told reuters us tariffs roiled global markets shaken buyer confidence"
> |
```

The cleaned text is tokenized into individual words using the tokenize_words() function. Each list of tokens is then collapsed back into a single string using paste() to maintain sentence-like structure for further processing.

Stemming The Tokenized Text

CODE

```
corpus_nostop <- VCorpus(VectorSource(Tokens_Text))
corpus_stemmed <- tm_map(corpus_nostop, stemDocument)
Stemmed_Text <- sapply(corpus_stemmed, as.character)
```

OUTPUT

```
508

"india export cut polish diamond plummet lowest level near two decad fiscal year end march sluggish demand unit state c
hina lead trade bodi said monday india world largest cut polish hub handl nine everi diamond process global sensit econ
om uncertainti particular us biggest market cut polish diamond export usual account near half overal gem jewelleri ship
ment fell billion yearonyear gem jewelleri export promot council gjepc said statement slump drag overal gem jewelleri e
xport billion fouryear low billion previous year lower demand polish diamond also prompt indian processor reduc import
rough diamond billion trade bodi said gem jewelleri export rose yearonyear march howev billion gjepc said export ramp s
hipment ahead announc us tariff us presid donald trump initi plan place tariff import indian good april part duti targe
t dozen countri declar day paus measur us buyer load march tariff kick indian export also rush ship us order first woul
dnt get hit extra cost said shaunak parikh vicechairman gjepc india gem jewelleri export unlik recov year one major mum
baibas export told reuter us tariff roil global market shaken buyer confid"
> |
```

The tokenized text is converted into a new corpus for stemming. The stemDocument() function is applied to reduce words to their root forms ("polished" => "polish"). The stemmed text is then extracted as plain character strings.

Lemmatization The Tokenized Text

CODE

```
corpus_lemmatized <- tm_map(corpus_nostop,
content_transformer(lemmatize_strings))
mydata$Lemmatized_Text <- sapply(corpus_lemmatized, as.character)
mydata$Lemmatized_Text
```

OUTPUT

```
[508] "indias export cut polish diamond plummet low level nearly two decade fiscal year end march sluggish demand unite
state china lead trade body say monday india world large cut polish hub handle nine every diamond process globally sens
itive economic uncertainty particularly us big market cut polish diamond export usually account nearly half overall gem
jewellery shipment fall billion yearonyear gem jewellery export promotion council gjepc say statement slump drag overal
l gem jewellery export billion fouryear low billion previous year low demand polish diamond also prompt indian processo
r reduce import rough diamond billion trade body say gem jewellery export rise yearonyear march however billion gjepc s
ay exporter ramp shipment ahead announce us tariff us president donald trump initially plan place tariff import indian
good april part duty target dozen country declare day pause measure us buyer load march tariff kick indian exporter als
o rush ship us order first wouldnt get hit extra cost say shaunak parikh vicechairman gjepc indias gem jewellery export
unlikely recover year one major mumbaibased exporter tell reuters us tariff roil global market shake buyer confidence"
>
```

Lemmatization is applied to the corpus to convert each word to its dictionary base form ("plummeted" => "plummet") The processed text is extracted and stored in a new column named Lemmatized_Text for further analysis.

- # Part-2

Document-Term Matrix

CODE

```
corpus <- VCorpus(VectorSource(mydata$Lemmatized_Text))
dtm <- DocumentTermMatrix(corpus)
```

OUTPUT

```
> dtm <- DocumentTermMatrix(corpus)
> dtm
<<DocumentTermMatrix (documents: 508, terms: 22154)>>
Non-/sparse entries: 113326/11140906
Sparsity           : 99%
Maximal term length: 69
Weighting          : term frequency (tf)
>
```

A Document-Term Matrix (DTM) was created from the lemmatized text corpus using the DocumentTermMatrix() function. The output shows that the matrix contains 508 documents (news articles) and 22,154 unique terms (words). The matrix is highly sparse (99%), meaning most words appear in only a few documents.

The DTM uses term frequency (tf) as the weighting scheme, where each cell indicates how many times a specific word occurs in each document.

Applying LDA

CODE

```
lda_model <- LDA(dtm, k = 5, control = list(seed = 1234))
mydata$Assigned_Topic <- topics(lda_model)
```

OUTPUT

```
> mydata$Assigned_Topic
  [1] 4 1 4 4 4 4 1 4 4 4 4 4 5 2 3 3 2 3 4 4 4 3 1 4 4 4 4 4 4 1 4 4 4 4 4 4 4 4 4 4 3 4 3 4 4 4 4 4 4 4 4 5 2 4 4 4 4 4 4 4 3
 [56] 4 4 3 4 4 3 3 1 4 4 3 4 4 3 4 3 4 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 3 4 4 4 4 4 4 1 3 4 4 5 4 4 4 3 4 3 3 5 2 2 5 2 5
[111] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 1 1 5 3 2 2 5 5 5 1 2 5 1 2 5 3 5 5 5 5 5 5 5 5 2 5 2 1 2 5 5 5
[166] 1 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 2 3 5 2 5 2 5 5 5 3 5 5 2 5 5 2 5 1 5 3 5 5 5 5 1 5 2 4 1 3 2 3 4 2 3 3 4 3 1 3
[221] 4 2 3 3 1 4 4 4 1 4 5 2 2 1 5 1 1 1 3 4 4 3 1 2 2 1 4 1 4 3 2 2 1 1 3 1 1 2 2 2 5 1 3 4 1 2 3 3 3 3 5 2 5 2 3
[276] 2 4 4 4 3 5 2 3 4 2 2 4 4 5 5 5 3 5 1 5 2 2 5 2 1 1 2 1 2 2 1 1 1 1 1 1 1 1 2 1 2 2 5 1 1 1 2 1 1 5 1 1 1 1 1
[331] 1 1 1 5 2 5 1 1 1 4 2 1 1 5 2 1 2 1 1 1 5 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 5 1 5 2 2 2 2 2 1 1 2 1 1 2 2 1
[386] 1 2 2 2 4 4 4 4 4 1 4 4 2 1 1 1 1 1 1 2 5 5 2 2 2 1 2 2 1 5 2 5 2 3 5 4 2 2 2 5 3 2 3 5 2 2 3 4 2 5 5 5 2 5 1 2
[441] 1 5 3 2 3 1 2 5 5 2 2 2 2 2 4 5 1 4 5 1 5 2 5 3 2 2 1 2 2 5 5 2 5 2 2 4 4 4 5 3 2 5 5 3 3 1 2 2 2 5 5 2 3 2 3
[496] 3 3 2 2 2 2 1 2 3 2 2 2 2
>
```

Latent Dirichlet Allocation (LDA) was applied with **k = 5**, instructing the model to identify five latent topics across the dataset. Using the topics() function, each of the 508 news articles was assigned to its most probable topic, stored in the Assigned_Topic column. The output confirms that all articles were successfully labeled, with topic numbers distributed throughout the dataset.

Extraction of Top 10 Words per Topic

CODE

```
top_terms <- terms(lda_model, 10)

for (i in 1:5) {
  cat("\nTopic ", i, "\n")
  print(top_terms[, i])
}
```

OUTPUT

```
Topic  1
 [1] "company"    "use"        "new"        "launch"     "service"    "plan"       "good"       "work"
 [9] "technology" "system"

Topic  2
 [1] "trump"     "good"      "tariff"    "high"      "use"       "come"      "time"      "president" "market"
[10] "trade"

Topic  3
 [1] "good"  "time"  "may"   "price" "day"   "see"   "come"  "give"  "world" "find"

Topic  4
 [1] "good" "film" "work" "show" "see"  "time" "know" "come" "like" "get"

Topic  5
 [1] "trump"     "official"  "state"     "president" "attack"    "force"     "day"       "war"       "tell"
[10] "country"
```

The top 10 most important words for each of the 5 topics identified by the LDA model are printed by this code. These words are considered the most relevant terms for interpreting the themes within each topic.

Extraction of Word Probabilities for Each Topic

CODE

```
topic_term_probs <- posterior(lda_model)$terms
top_n <- 10
for (i in 1:5) {
  cat(paste0("\nTopic ", i, " Word Probabilities \n"))
  sorted <- sort(topic_term_probs[i, ], decreasing = TRUE)
  top_words <- head(sorted, top_n)
  df <- data.frame(Word = names(top_words),
                   Probability = round(top_words, 4))
  print(df, row.names = FALSE)
}
```

OUTPUT

Topic 1 Word Probabilities

Word	Probability
company	0.0047
use	0.0039
new	0.0029
launch	0.0025
service	0.0025
plan	0.0025
good	0.0025
work	0.0025
technology	0.0024
system	0.0024

Topic 2 Word Probabilities

Word	Probability
trump	0.0037
good	0.0036
tariff	0.0032
high	0.0029
use	0.0028
come	0.0027
time	0.0027
president	0.0027
market	0.0026
trade	0.0025

Topic 3 Word Probabilities

Word	Probability
good	0.0035
time	0.0029
may	0.0025
price	0.0023
day	0.0022
see	0.0022
come	0.0021
give	0.0020
world	0.0020
find	0.0019

Topic 4 Word Probabilities

Word	Probability
good	0.0035
film	0.0033
work	0.0029
show	0.0028
see	0.0028
time	0.0027
know	0.0025
come	0.0025
like	0.0025
get	0.0025

Topic 5 Word Probabilities

Word	Probability
trump	0.0038
official	0.0036
state	0.0035
president	0.0033
attack	0.0033
force	0.0032
day	0.0029
war	0.0029
tell	0.0028
country	0.0028

posterior(lda_model)$terms is used to extract the word-topic probability matrix from the LDA model and stores it in topic_term_probs, which holds the probabilities of each word across all topics. For each topic, it sorts all words by their probability and prints the top 10 most representative terms along with their probability values, helping to interpret the key themes within each topic.

Topic Proportions for 10 Sample Documents

CODE

```
topic_proportions <- posterior(lda_model)$topics
head(topic_proportions, 10)
```

OUTPUT

```
> head(topic_proportions, 10)
              1            2            3            4            5
1  8.830420e-05 8.830420e-05 8.830420e-05 0.9996467832 8.830420e-05
2  9.992468e-01 1.882881e-04 1.882881e-04 0.0001882881 1.882881e-04
3  9.453835e-05 9.453835e-05 9.453835e-05 0.7709432931 2.287731e-01
4  9.247665e-05 2.164190e-01 9.247665e-05 0.7833035889 9.247665e-05
5  7.898384e-05 7.898384e-05 7.898384e-05 0.9996840647 7.898384e-05
6  6.087972e-01 9.247665e-05 9.247665e-05 0.3909253607 9.247665e-05
7  9.856707e-05 9.856707e-05 9.856707e-05 0.9996057317 9.856707e-05
8  2.229341e-04 2.229341e-04 2.229341e-04 0.9991082638 2.229341e-04
9  2.231055e-01 2.787201e-01 1.317414e-04 0.4979108450 1.317414e-04
10 1.742033e-01 1.331199e-04 1.331199e-04 0.8253973836 1.331199e-04
>
```

This matrix displays the topic proportions for the first 10 documents. Each row corresponds to a document, and each column shows the probability that the document belongs to a specific topic. The values in each row sum to 1, indicating how much each topic contributes to the document's content.

Topic Interpretation

- ### Topic 1 Interpretation

Top words of topic 1 are **[*company, use, new, launch, service, plan, good, work, technology, system*]**
Most probable word: **company**

Topic 1 is centered around business, technology, and service-related developments. Words like *"launch"*, *"technology"*, and *"system"* suggest news about product releases, innovation, or company operations. The presence of *"plan"* and *"work"* further indicates forward-looking or operational aspects of businesses. So, it can be said that Topic 1 likely represents corporate and technology-oriented news, focusing on product launches, services, and business plans.

- ### Topic 2 Interpretation

Top words of topic 2 are ***[trump, good, tariff, high, use, come, time, president, market, trade]***
Most probable word: **trump**

This topic focuses on politics and economics, particularly trade issues and government figures. Terms like *"tariff"*, *"market"*, and *"trade"* point to economic policy discussions, while *"trump"* and *"president"* indicate political leadership as a central theme. So, it can be said that Topic 2 likely

represents economic policy and political news, especially U.S.-centered trade and leadership discussions.

- ## Topic 3 Interpretation

Top words topic 3 are ***[good, time, may, price, day, see, come, give, world, find]***
Most probable word**: good**

This topic appears to be general or editorial in tone, possibly discussing broad trends, opinions, or speculative articles. Words like *"may"*, *"see"*, *"find"*, and *"give"* are common in commentary or feature stories. So, it can be said that Topic 3 likely reflects general or reflective news articles covering opinions, societal insights, or global commentary.

- ## Topic 4 Interpretation

Top words of topic 4 are ***[good, film, work, show, see, time, know, come, like, get]***
Most probable word: **good**

Topic 4 centers on entertainment and media-related content. Words such as *"film"*, *"show"*, *"see"*, and *"like"* suggest coverage of movies, performances, or celebrity news. The use of *"work"* and *"get"* may also relate to personal stories or reviews. So, it can be said that Topic 4 likely represents entertainment, film, and showbiz coverage.

- ## Topic 5 Interpretation

Top words of topic 4 are ***[trump, official, state, president, attack, force, day, war, tell, country]***
Most probable word: **trump**

Similar to Topic 2 ,Topic 5 strongly reflects political and international affairs, especially involving conflict, leadership, and official statements. Terms like *"war"*, *"attack"*, *"force"*, and *"state"* indicate serious geopolitical issues. . So, it can be said that Topic 5 likely represents hard news focused on international relations, government actions, and security issues

Topic Proportion Interpretation (Sample of 10 Documents)

The matrix shows how much each topic contributes to each document. Each row corresponds to one news article, and each column represents the proportion (probability) of a particular topic in that document.

- **Document 1:**
  - Dominated by **Topic 4** (≈ 99.96%)
    → This article likely relates to entertainment, events, or media, based on Topic 4's words like *film*, *show*, and *see*.

- **Document 2:**
  - Dominated by **Topic 1** (≈ 99.94%)
    → Strongly aligned with business and technology news — possibly about company launches or services.
- **Document 3:**
  - Mixed, but **Topic 3** is strongest (≈ 77%)
    → Likely a general update piece, perhaps economic or social news — based on Topic 3's frequent words like *may*, *price*, *world*.
- **Document 4:**
  - Highest in **Topic 4** (≈ 78%)
    → Suggests a media-related article again, possibly covering an event or show.
- **Document 5:**
  - Very high in **Topic 4** (≈ 99.96%)
    → Strongly aligned with entertainment/media content.
- **Document 6:**
  - Highest in **Topic 4** (≈ 39%), but more mixed
    → Could be a general article covering multiple themes, slightly leaning toward entertainment.
- **Document 7:**
  - Dominated by **Topic 4** (≈ 99.96%)
    → Again suggests coverage of an event, film, or similar content.
- **Document 8:**
  - Mostly **Topic 2** (≈ 99.91%)
    → Likely political/economic in nature, given Topic 2's association with *trump*, *market*, and *trade*.
- **Document 9:**
  - Highest in **Topic 2** (≈ 28%) and **Topic 1** (≈ 23%)
    → May be a mixed-genre article involving both business and political context.
- **Document 10:**
  - Strong in **Topic 4** (≈ 82%)
    → Most likely an entertainment article.

Overall Topic Modeling Interpretation:

The LDA model revealed five distinct topics across over 500 news articles. Topic 1 focuses on business, technology, and service-related news. Topic 2 captures political and economic discussions, particularly around U.S. leadership and trade. Topic 3 reflects general opinion pieces or societal commentary. Topic 4 covers entertainment and media-related content such as films and shows. Topic 5 centers on international affairs, government actions, and geopolitical issues. The document-topic proportion matrix revealed that a large portion of articles (especially in the first 10) were strongly influenced by Topic 4, showing a dominance of entertainment news in the

sample and some documents were dominated by Topic 1 and Topic 2, highlighting the presence of business and political reporting. The distribution of these topics suggests that the dataset offers a diverse mix of business, politics, media, and global news coverage.