Introduction to Data Science [B]

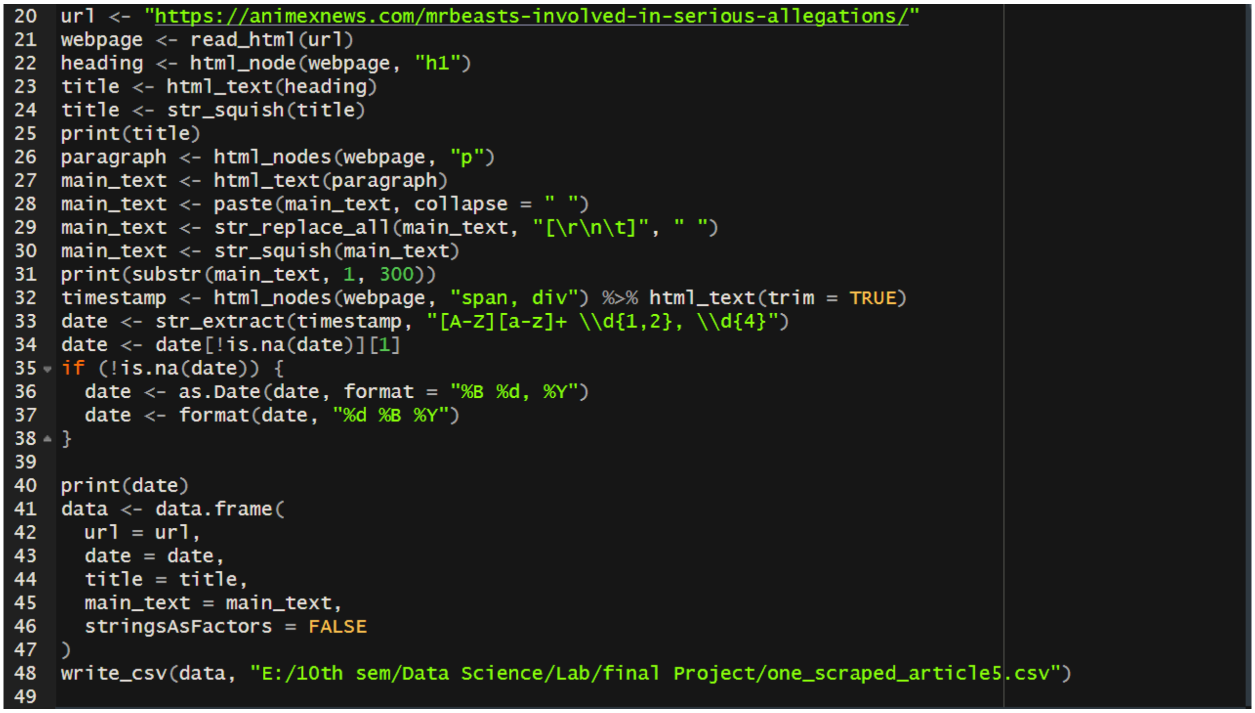
Faculty: TOHEDUL ISLAM
Title: Final Term Project Report
Group: 9

Group Member

Name	ID
MD TAFHIMUL HAQUE SADI	22-47071-1
RIFAH SANZIDA	22-47154-1
SADIA AFEOSE	21-45820-3

Final Project Part-1

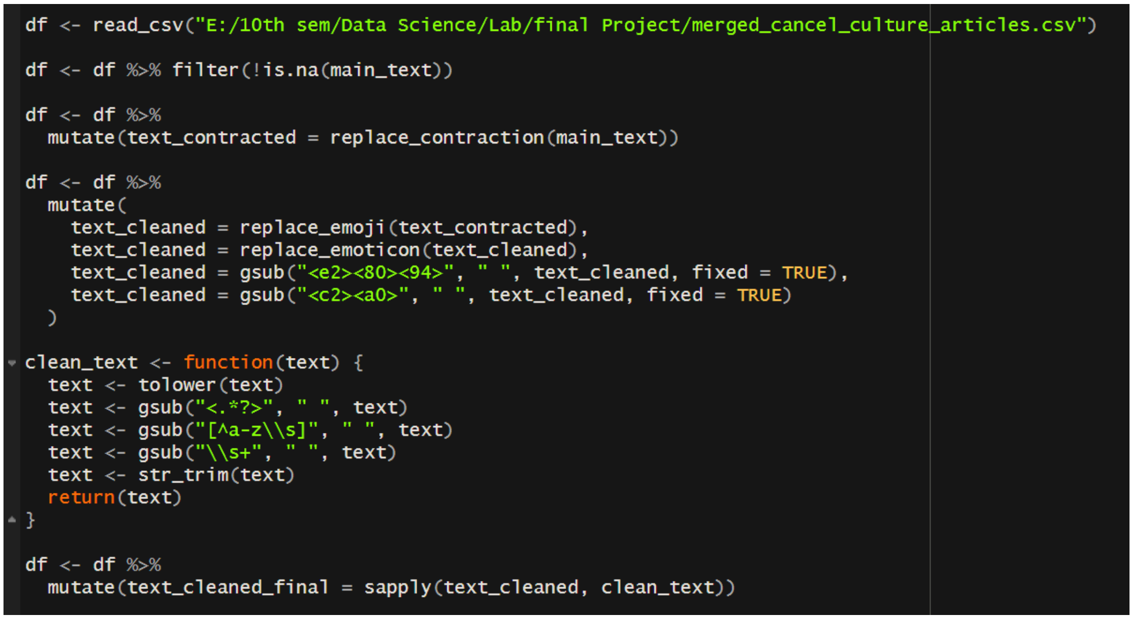
Web Scrapping data from URLs:



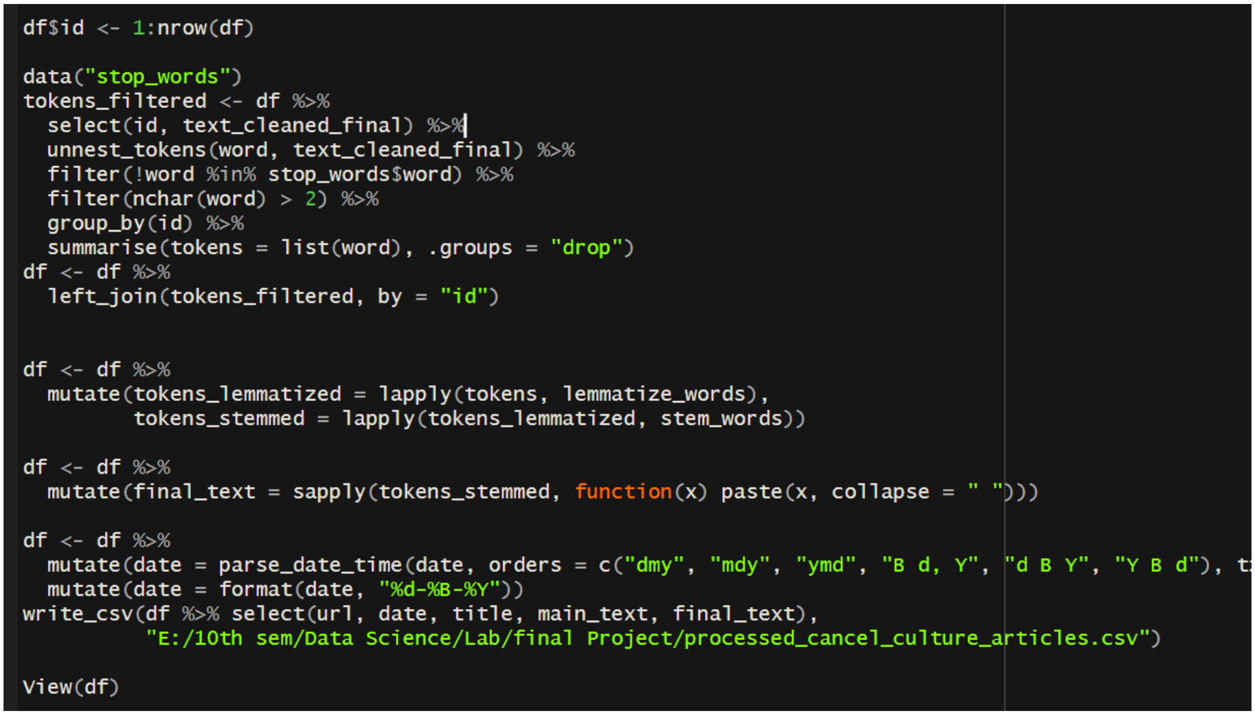
```r
url <- "https://animexnews.com/mrbeasts-involved-in-serious-allegations/"
webpage <- read_html(url)
heading <- html_node(webpage, "h1")
title <- html_text(heading)
title <- str_squish(title)
print(title)
paragraph <- html_nodes(webpage, "p")
main_text <- html_text(paragraph)
main_text <- paste(main_text, collapse = " ")
main_text <- str_replace_all(main_text, "[\r\n\t]", " ")
main_text <- str_squish(main_text)
print(substr(main_text, 1, 300))
timestamp <- html_nodes(webpage, "span, div") %>% html_text(trim = TRUE)
date <- str_extract(timestamp, "[A-Z][a-z]+ \\d{1,2}, \\d{4}")
date <- date[!is.na(date)][1]
if (!is.na(date)) {
  date <- as.Date(date, format = "%B %d, %Y")
  date <- format(date, "%d %B %Y")
}

print(date)
data <- data.frame(
  url = url,
  date = date,
  title = title,
  main_text = main_text,
  stringsAsFactors = FALSE
)
write_csv(data, "E:/10th sem/Data Science/Lab/final Project/one_scraped_article5.csv")
```

We initially select 63 url from different website about Toxicity and Cancel culture. For every URL we run the above picture code of web scrapping and save this in a csv file in my local machine. We extract heading by selecting "h1" tag, all the content for all the paragraph as <p> tag and extract date. After web scrapping for 63 url we ger 63 csv file. And finally we marge those csv file.

Data Preprocessing:



```r
df <- read_csv("E:/10th sem/Data Science/Lab/final Project/merged_cancel_culture_articles.csv")

df <- df %>% filter(!is.na(main_text))

df <- df %>%
  mutate(text_contracted = replace_contraction(main_text))

df <- df %>%
  mutate(
    text_cleaned = replace_emoji(text_contracted),
    text_cleaned = replace_emoticon(text_cleaned),
    text_cleaned = gsub("<e2><80><94>", " ", text_cleaned, fixed = TRUE),
    text_cleaned = gsub("<c2><a0>", " ", text_cleaned, fixed = TRUE)
  )

clean_text <- function(text) {
  text <- tolower(text)
  text <- gsub("<.*?>", " ", text)
  text <- gsub("[^a-z\\s]", " ", text)
  text <- gsub("\\s+", " ", text)
  text <- str_trim(text)
  return(text)
}

df <- df %>%
  mutate(text_cleaned_final = sapply(text_cleaned, clean_text))
```

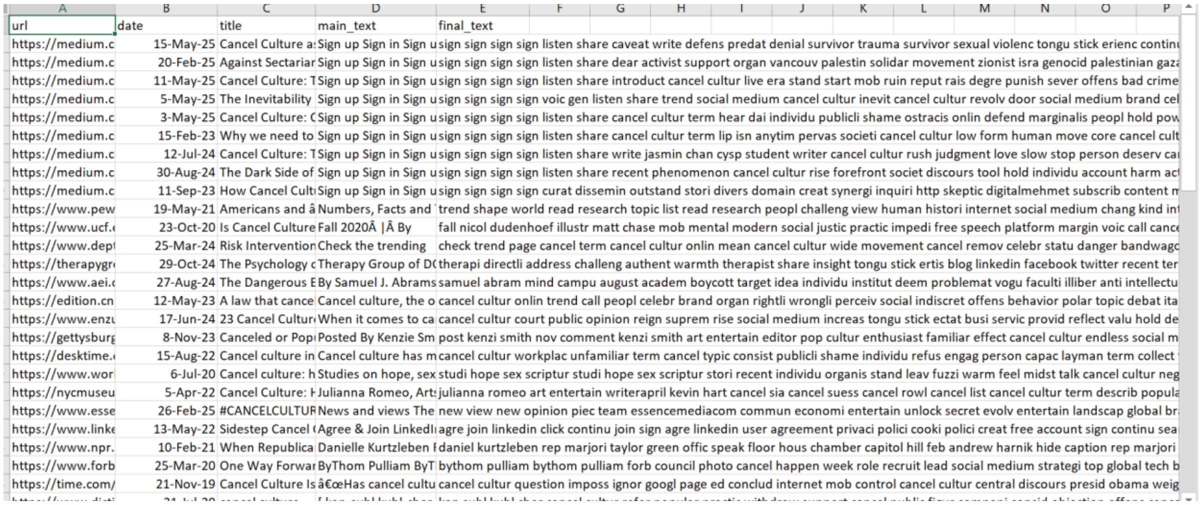


```r
df$id <- 1:nrow(df)

data("stop_words")
tokens_filtered <- df %>%
  select(id, text_cleaned_final) %>%
  unnest_tokens(word, text_cleaned_final) %>%
  filter(!word %in% stop_words$word) %>%
  filter(nchar(word) > 2) %>%
  group_by(id) %>%
  summarise(tokens = list(word), .groups = "drop")
df <- df %>%
  left_join(tokens_filtered, by = "id")


df <- df %>%
  mutate(tokens_lemmatized = lapply(tokens, lemmatize_words),
         tokens_stemmed = lapply(tokens_lemmatized, stem_words))

df <- df %>%
  mutate(final_text = sapply(tokens_stemmed, function(x) paste(x, collapse = " ")))

df <- df %>%
  mutate(date = parse_date_time(date, orders = c("dmy", "mdy", "ymd", "B d, Y", "d B Y", "Y B d"), t
  mutate(date = format(date, "%d-%B-%Y"))
write_csv(df %>% select(url, date, title, main_text, final_text),
          "E:/10th sem/Data Science/Lab/final Project/processed_cancel_culture_articles.csv")

View(df)
```

This R script that does extensive text processing to a dataset of cancel culture articles. It starts with loading the data-frame and filtering out rows with NA in main_text. Following that, you get a few cleaning transforms like contraction expansion, emoji, and emoticon removal (don't want no emojis on my dirty data!), and some cleanup for HTML code symbols. An additional custom function standardizes the text by lowercasing text, and cleaning with respect to special characters and additional space.

The script then tokenizes the cleaned text, removes short words and stop words, followed by lemmatization and stemming on each list of tokens. The processed tokens are then revected back into cleaned text (final_text) for each article at last. The date field is also getting parsed and formatted in the script to a particular format. The resulting processed dataset is saved to a new CSV file so that it can be analysed in more detail.

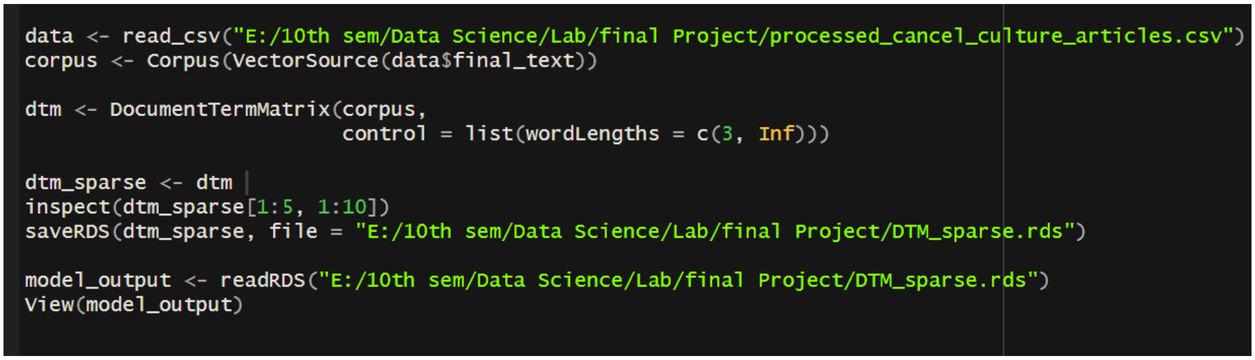
After preprocessing, The dataset result was:



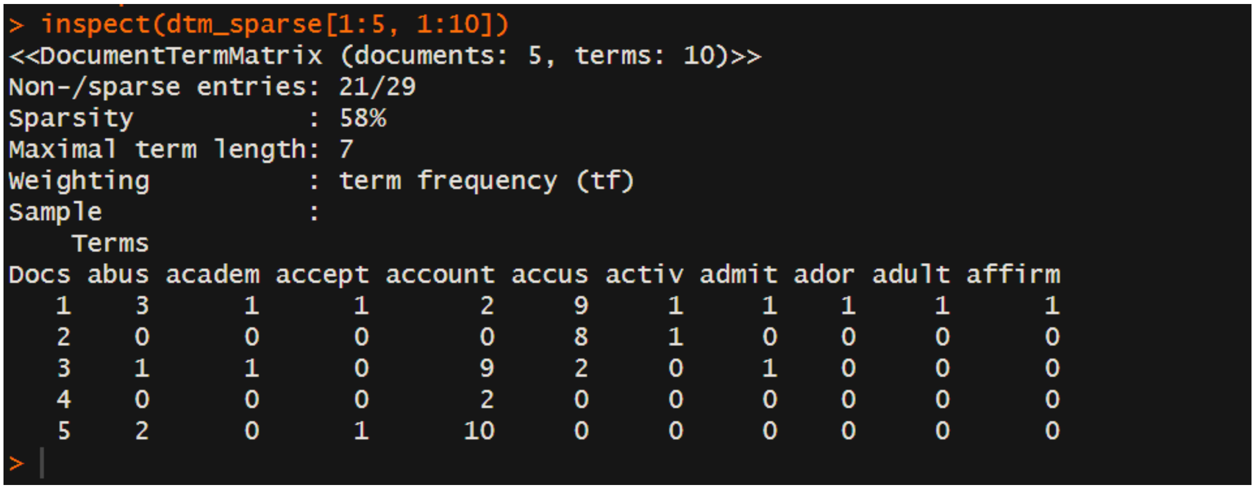
url	date	title	main_text	final_text
https://medium.c	15-May-25	Cancel Culture a:	Sign up Sign in Sign u	sign sign sign sign listen share caveat write defens predat denial survivor trauma survivor sexual violenc tongu stick erienc continu
https://medium.c	20-Feb-25	Against Sectariar	Sign up Sign in Sign u	sign sign sign sign listen share dear activist support organ vancouv palestin solidar movement zionist isra genocid palestinian gaza
https://medium.c	11-May-25	Cancel Culture: T	Sign up Sign in Sign u	sign sign sign sign listen share introduct cancel cultur live era stand start mob ruin reput rais degre punish sever offens bad crime
https://medium.c	5-May-25	The Inevitability	Sign up Sign in Sign u	sign sign sign sign voic gen listen share trend social medium cancel cultur inevit cancel cultur revolv door social medium brand cel
https://medium.c	3-May-25	Cancel Culture: (	Sign up Sign in Sign u	sign sign sign sign listen share cancel cultur term hear dai individu publicli shame ostracis onlin defend marginalis peopl hold pow
https://medium.c	15-Feb-23	Why we need to	Sign up Sign in Sign u	sign sign sign sign listen share cancel cultur term lip isn anytim pervas societi cancel cultur low form human move core cancel cult
https://medium.c	12-Jul-24	Cancel Culture: T	Sign up Sign in Sign u	sign sign sign sign listen share write jasmin chan cysp student writer cancel cultur rush judgment love slow stop person deserv ca
https://medium.c	30-Aug-24	The Dark Side of	Sign up Sign in Sign u	sign sign sign sign listen share recent phenomenon cancel cultur rise forefront societ discours tool hold individu account harm act
https://medium.c	11-Sep-23	How Cancel Cult	Sign up Sign in Sign u	sign sign sign sign curat dissemin outstand stori divers domain creat synergi inquiri http skeptic digitalmehmet subscrib content m
https://www.pew	19-May-21	Americans and â	Numbers, Facts and '	trend shape world read research topic list read research peopl challeng view human histori internet social medium chang kind int
https://www.ucf.e	23-Oct-20	Is Cancel Culture	Fall 2020Â |Â By	fall nicol dudenhoef illustr matt chase mob mental modern social justic practic impedi free speech platform margin voic call cance
https://www.dept	25-Mar-24	Risk Intervention	Check the trending	check trend page cancel term cancel cultur onlin mean cancel cultur wide movement cancel remov celebr statu danger bandwago
https://therapygr	29-Oct-24	The Psychology c	Therapy Group of DC	therapi directli address challeng authent warmth therapist share insight tongu stick ertis blog linkedin facebook twitter recent ter
https://www.aei.c	27-Aug-24	The Dangerous E	By Samuel J. Abrams	samuel abram mind campu august academ boycott target idea individu institut deem problemat vogu faculti illiber anti intellectu
https://edition.cn	12-May-23	A law that cance	Cancel culture, the o	cancel cultur onlin trend call peopl celebr brand organ rightli wrongli perceiv social indiscret offens behavior polar topic debat ita
https://www.enzu	17-Jun-24	23 Cancel Cultur	When it comes to ca	cancel cultur court public opinion reign suprem rise social medium increas tongu stick ectat busi servic provid reflect valu hold de
https://gettysburg	8-Nov-23	Canceled or Pop	Posted By Kenzie Sm	post kenzi smith nov comment kenzi smith art entertain editor pop cultur enthusiast familiar effect cancel cultur endless social m
https://desktime.c	15-Aug-22	Cancel culture in	Cancel culture has m	cancel cultur workplac unfamiliar term cancel typic consist publicli shame individu refus engag person capac layman term collect
https://www.wor	6-Jul-20	Cancel culture: h	Studies on hope, sex	studi hope sex scriptur studi hope sex scriptur stori recent individu organis stand leav fuzzi warm feel midst talk cancel cultur neg
https://nycmuseu	5-Apr-22	Cancel Culture: H	Julianna Romeo, Arts	julianna romeo art entertain writerapril kevin hart cancel sia cancel suess cancel rowl cancel list cancel cultur term describ popula
https://www.esse	26-Feb-25	#CANCELCULTUR	News and views The	new view new opinion piec team essencemediacom commun economi entertain unlock secret evolv entertain landscap global br
https://www.linke	13-May-22	Sidestep Cancel (	Agree & Join LinkedI	agre join linkedin click continu join sign agre linkedin user agreement privaci polici cooki polici creat free account sign continu sea
https://www.npr.	10-Feb-21	When Republica	Danielle Kurtzleben I	daniel kurtzleben rep marjori taylor green offic speak floor hous chamber capitol hill feb andrew harnik hide caption rep marjori
https://www.forb	25-Mar-20	One Way Forwar	ByThom Pulliam ByT	bythom pulliam bythom pulliam forb council photo cancel happen week role recruit lead social medium strategi top global tech b
https://time.com/	21-Nov-19	Cancel Culture Is	â€œHas cancel cultu	cancel cultur question imposs ignor googl page ed conclud internet mob control cancel cultur central discours presid obama weig

Final Term project Report Part 2

Document Term Matrix (DTM):



```
data <- read_csv("E:/10th sem/Data Science/Lab/final Project/processed_cancel_culture_articles.csv")
corpus <- Corpus(VectorSource(data$final_text))

dtm <- DocumentTermMatrix(corpus,
                          control = list(wordLengths = c(3, Inf)))

dtm_sparse <- dtm
inspect(dtm_sparse[1:5, 1:10])
saveRDS(dtm_sparse, file = "E:/10th sem/Data Science/Lab/final Project/DTM_sparse.rds")

model_output <- readRDS("E:/10th sem/Data Science/Lab/final Project/DTM_sparse.rds")
View(model_output)
```

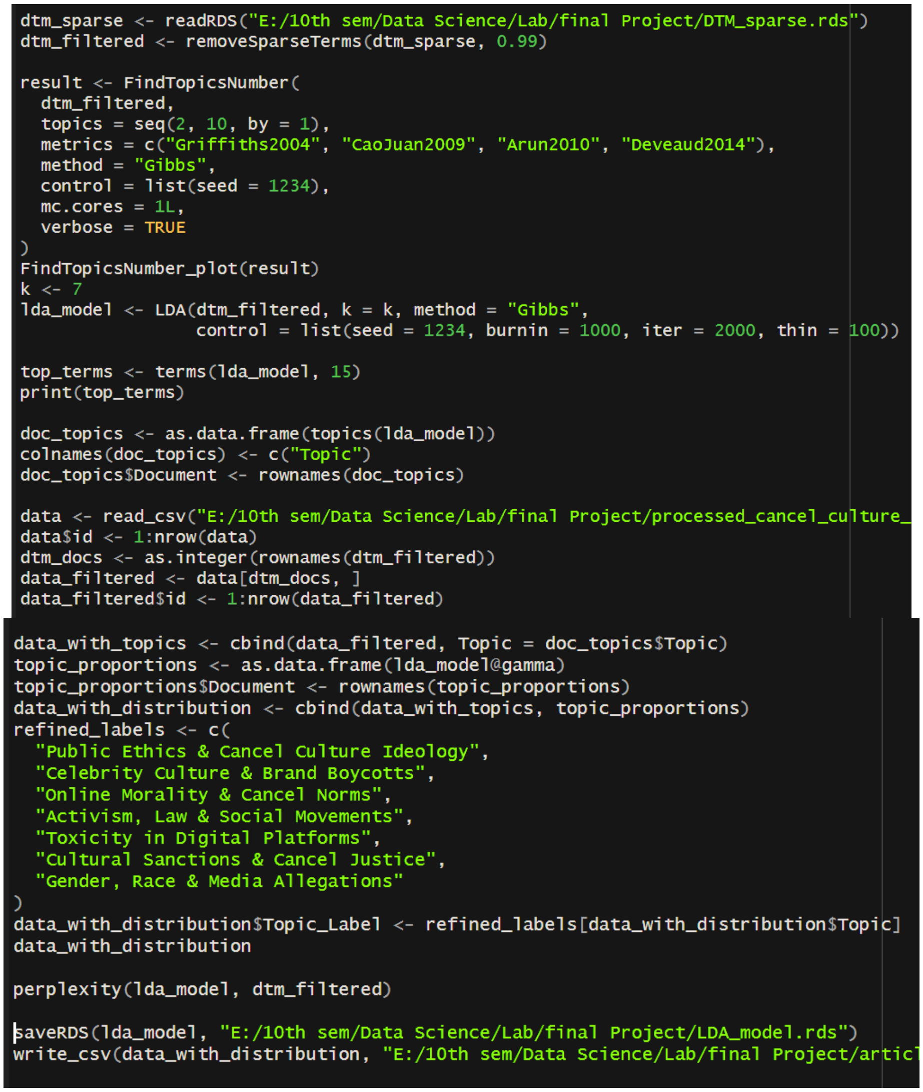


```
> inspect(dtm_sparse[1:5, 1:10])
<<DocumentTermMatrix (documents: 5, terms: 10)>>
Non-/sparse entries: 21/29
Sparsity           : 58%
Maximal term length: 7
Weighting          : term frequency (tf)
Sample             :
    Terms
Docs abus academ accept account accus activ admit ador adult affirm
   1    3      1      1       2     9     1     1    1     1      1
   2    0      0      0       0     8     1     0    0     0      0
   3    1      1      0       9     2     0     1    0     0      0
   4    0      0      0       2     0     0     0    0     0      0
   5    2      0      1      10     0     0     0    0     0      0
> 
```

A Document-Term Matrix (DTM) is created from the comment column. This matrix represents the frequency of terms (words) in each document (comment). Sparse terms (words that appear in less than 1% of documents) are removed to reduce noise. Rows with zero total term frequencies are removed to ensure only meaningful documents remain.

Load the pre-processed article data using the following R code, and create a text corpus with the column final_text. It then creates a kind of frequency matrix with texts, called the Document-Term Matrix (DTM) with frequency of terms (words with atleast 3 are characters) in a document. The matrix is succinctly represented and is examined for 5 documents and 10 terms as sample. Finally, the DTM is output as a. rds file and reload it for checking or further use.

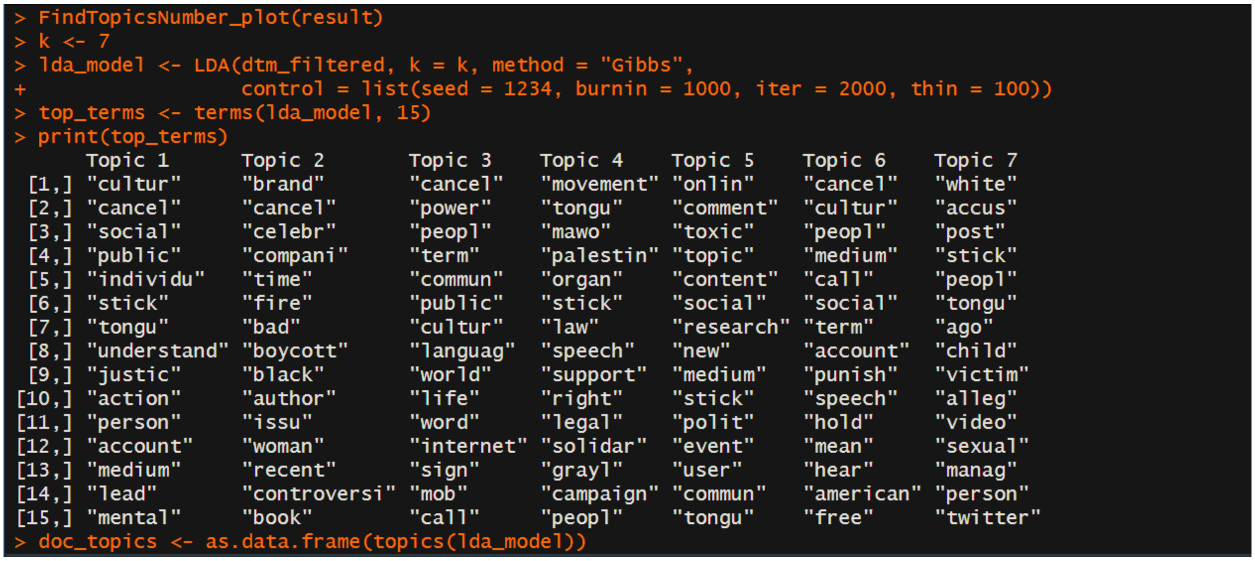
Build LDA model:



```r
dtm_sparse <- readRDS("E:/10th sem/Data Science/Lab/final Project/DTM_sparse.rds")
dtm_filtered <- removeSparseTerms(dtm_sparse, 0.99)

result <- FindTopicsNumber(
  dtm_filtered,
  topics = seq(2, 10, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234),
  mc.cores = 1L,
  verbose = TRUE
)
FindTopicsNumber_plot(result)
k <- 7
lda_model <- LDA(dtm_filtered, k = k, method = "Gibbs",
                 control = list(seed = 1234, burnin = 1000, iter = 2000, thin = 100))

top_terms <- terms(lda_model, 15)
print(top_terms)

doc_topics <- as.data.frame(topics(lda_model))
colnames(doc_topics) <- c("Topic")
doc_topics$Document <- rownames(doc_topics)

data <- read_csv("E:/10th sem/Data Science/Lab/final Project/processed_cancel_culture_
data$id <- 1:nrow(data)
dtm_docs <- as.integer(rownames(dtm_filtered))
data_filtered <- data[dtm_docs, ]
data_filtered$id <- 1:nrow(data_filtered)
data_with_topics <- cbind(data_filtered, Topic = doc_topics$Topic)
topic_proportions <- as.data.frame(lda_model@gamma)
topic_proportions$Document <- rownames(topic_proportions)
data_with_distribution <- cbind(data_with_topics, topic_proportions)
refined_labels <- c(
  "Public Ethics & Cancel Culture Ideology",
  "Celebrity Culture & Brand Boycotts",
  "Online Morality & Cancel Norms",
  "Activism, Law & Social Movements",
  "Toxicity in Digital Platforms",
  "Cultural Sanctions & Cancel Justice",
  "Gender, Race & Media Allegations"
)
data_with_distribution$Topic_Label <- refined_labels[data_with_distribution$Topic]
data_with_distribution

perplexity(lda_model, dtm_filtered)

saveRDS(lda_model, "E:/10th sem/Data Science/Lab/final Project/LDA_model.rds")
write_csv(data_with_distribution, "E:/10th sem/Data Science/Lab/final Project/artic
```

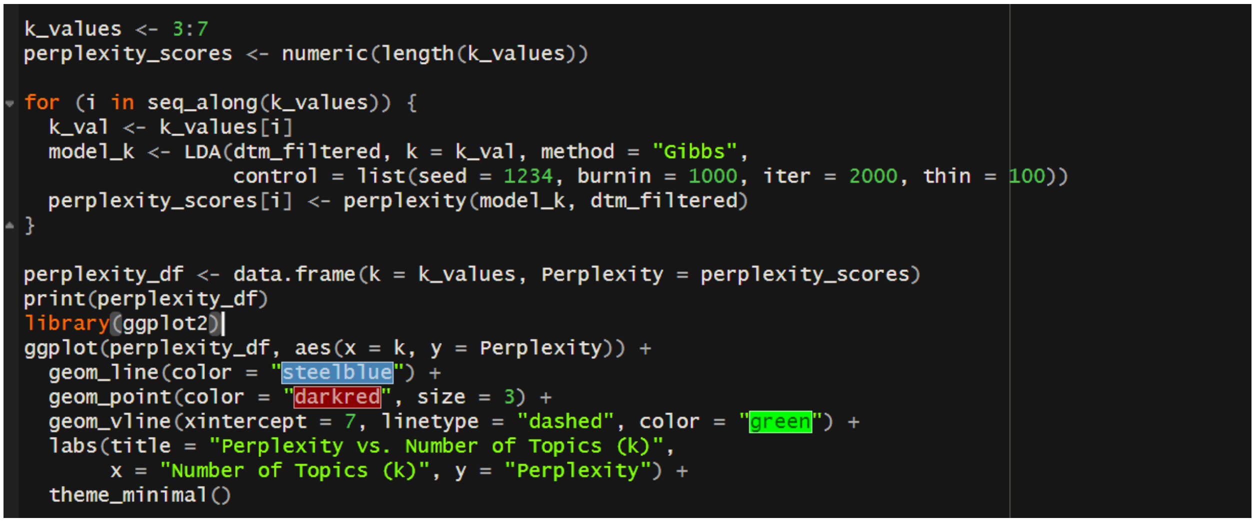
This code does the topic modelling with Latent Dirichlet Allocation (LDA) on the preprocessed text dataset. It begins with loading the saved Document-Term Matrix (DTM) and clip sparse terms to achieve concentration on the more frequent words. The best number of topics that must be considered is attested with four of the evaluation methods established above (Griffiths2004, CaoJuan2009, Arun2010, Deveaud2014) by means of the FindTopicsNumber function. With this, topics are used (k = 7) and LDA model is trained using Gibbs sampling.

We extract the top 15 terms for each topic & classify each document by its most probable topic. - (ii) Topic distribution per document is inferred and added to the original data. Custom labels are added for each topic to enhance readability, such as “Public Ethics & Cancel Culture Ideology” or “Toxicity in Digital Platforms”. Finally, model perplexity is calculated for evaluation, and the results are saved for further analysis and visualization.

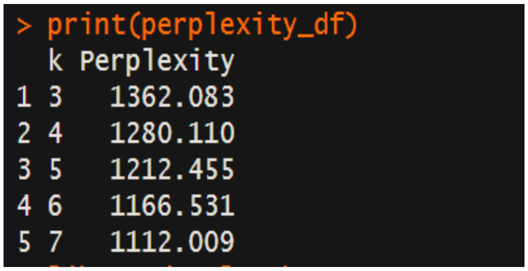


```
> FindTopicsNumber_plot(result)
> k <- 7
> lda_model <- LDA(dtm_filtered, k = k, method = "Gibbs",
+                  control = list(seed = 1234, burnin = 1000, iter = 2000, thin = 100))
> top_terms <- terms(lda_model, 15)
> print(top_terms)
      Topic 1      Topic 2        Topic 3     Topic 4     Topic 5     Topic 6     Topic 7
 [1,] "cultur"     "brand"        "cancel"    "movement"  "onlin"     "cancel"    "white"
 [2,] "cancel"     "cancel"       "power"     "tongu"     "comment"   "cultur"    "accus"
 [3,] "social"     "celebr"       "peopl"     "mawo"      "toxic"     "peopl"     "post"
 [4,] "public"     "compani"      "term"      "palestin"  "topic"     "medium"    "stick"
 [5,] "individu"   "time"         "commun"    "organ"     "content"   "call"      "peopl"
 [6,] "stick"      "fire"         "public"    "stick"     "social"    "social"    "tongu"
 [7,] "tongu"      "bad"          "cultur"    "law"       "research"  "term"      "ago"
 [8,] "understand" "boycott"      "languag"   "speech"    "new"       "account"   "child"
 [9,] "justic"     "black"        "world"     "support"   "medium"    "punish"    "victim"
[10,] "action"     "author"       "life"      "right"     "stick"     "speech"    "alleg"
[11,] "person"     "issu"         "word"      "legal"     "polit"     "hold"      "video"
[12,] "account"    "woman"        "internet"  "solidar"   "event"     "mean"      "sexual"
[13,] "medium"     "recent"       "sign"      "grayl"     "user"      "hear"      "manag"
[14,] "lead"       "controversi"  "mob"       "campaign"  "commun"    "american"  "person"
[15,] "mental"     "book"         "call"      "peopl"     "tongu"     "free"      "twitter"
> doc_topics <- as.data.frame(topics(lda_model))
```

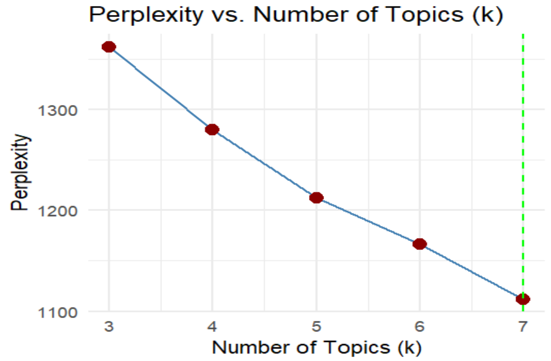
Model Evaluation using Perplexity:



```
k_values <- 3:7
perplexity_scores <- numeric(length(k_values))

for (i in seq_along(k_values)) {
  k_val <- k_values[i]
  model_k <- LDA(dtm_filtered, k = k_val, method = "Gibbs",
                 control = list(seed = 1234, burnin = 1000, iter = 2000, thin = 100))
  perplexity_scores[i] <- perplexity(model_k, dtm_filtered)
}

perplexity_df <- data.frame(k = k_values, Perplexity = perplexity_scores)
print(perplexity_df)
library(ggplot2)
ggplot(perplexity_df, aes(x = k, y = Perplexity)) +
  geom_line(color = "steelblue") +
  geom_point(color = "darkred", size = 3) +
  geom_vline(xintercept = 7, linetype = "dashed", color = "green") +
  labs(title = "Perplexity vs. Number of Topics (k)",
       x = "Number of Topics (k)", y = "Perplexity") +
  theme_minimal()
```



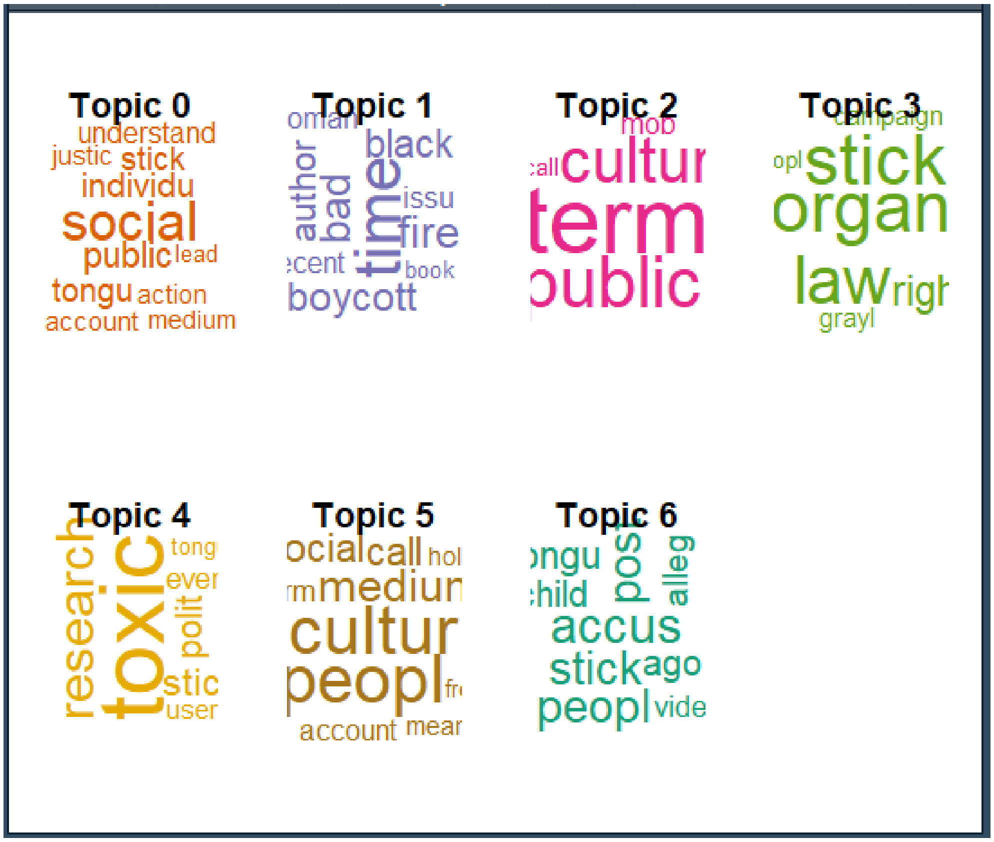
```
> print(perplexity_df)
  k Perplexity
1 3   1362.083
2 4   1280.110
3 5   1212.455
4 6   1166.531
5 7   1112.009
```



This section evaluates LDA model performance for different topic numbers (`k = 3` to `7`) using **perplexity**, a common metric for model quality (lower is better). For each `k`, an LDA model is trained, and its perplexity is recorded. The results are stored in a data frame and visualized using `ggplot2`. The plot displays how perplexity changes with the number of topics, helping identify the optimal `k`. A dashed green line marks `k = 7`, highlighting it as the chosen model based on earlier evaluation and topic coherence.

Visualization

Word Cloud:

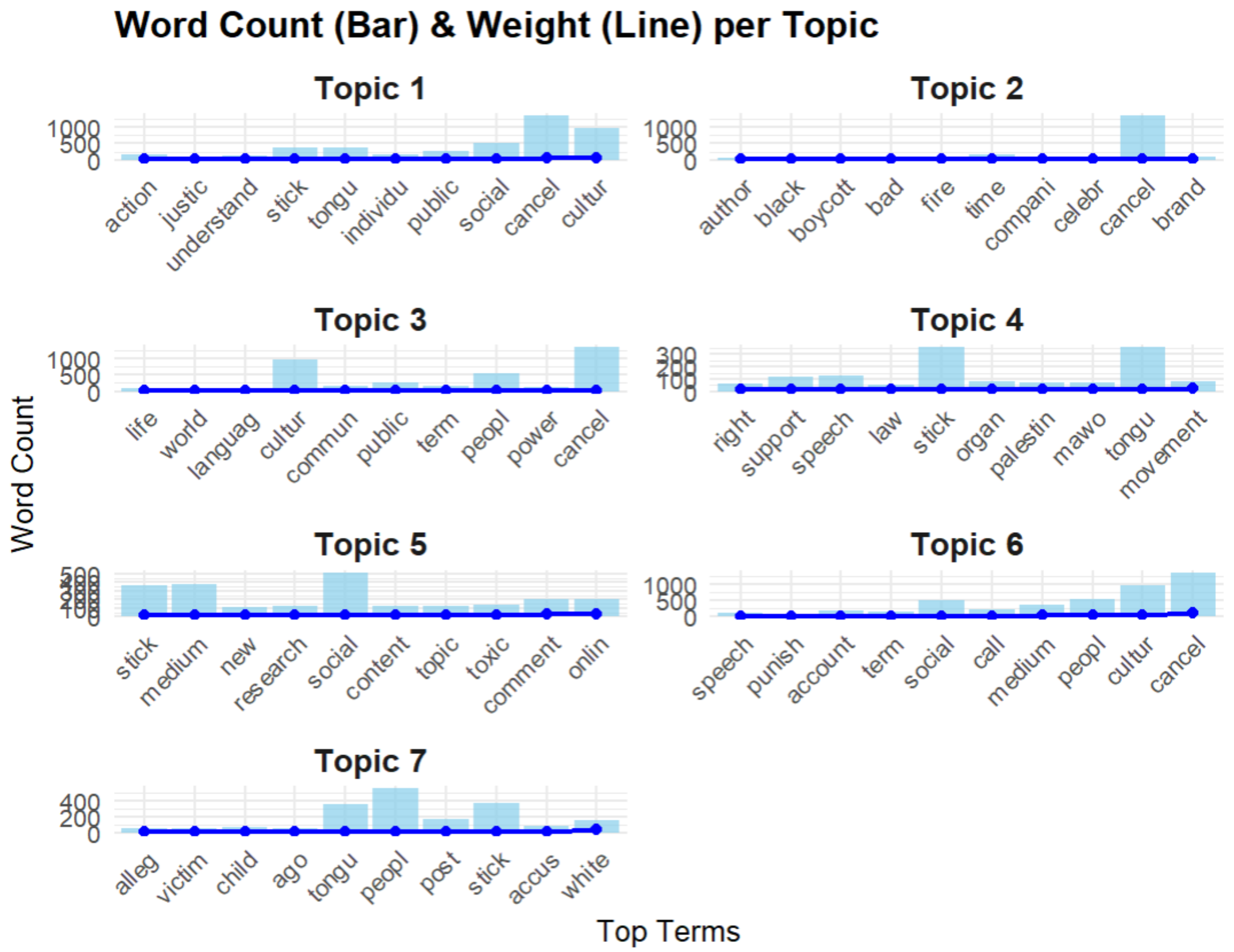


The image shows a word cloud visualization of the 7 LDA topics generated from the dataset. Each topic highlights its most representative terms based on frequency and importance. The topics have been manually interpreted and labeled for clarity:

1. Public Ethics & Cancel Culture Ideology
2. Celebrity Culture & Brand Boycotts
3. Online Morality & Cancel Norms
4. Activism, Law & Social Movements
5. Toxicity in Digital Platforms
6. Cultural Sanctions & Cancel Justice
7. Gender, Race & Media Allegations

These labels reflect the dominant themes in each topic, enhancing interpretability and linking textual patterns to real-world narratives in cancel culture discourse.

Word Count and Weight For each Topic:

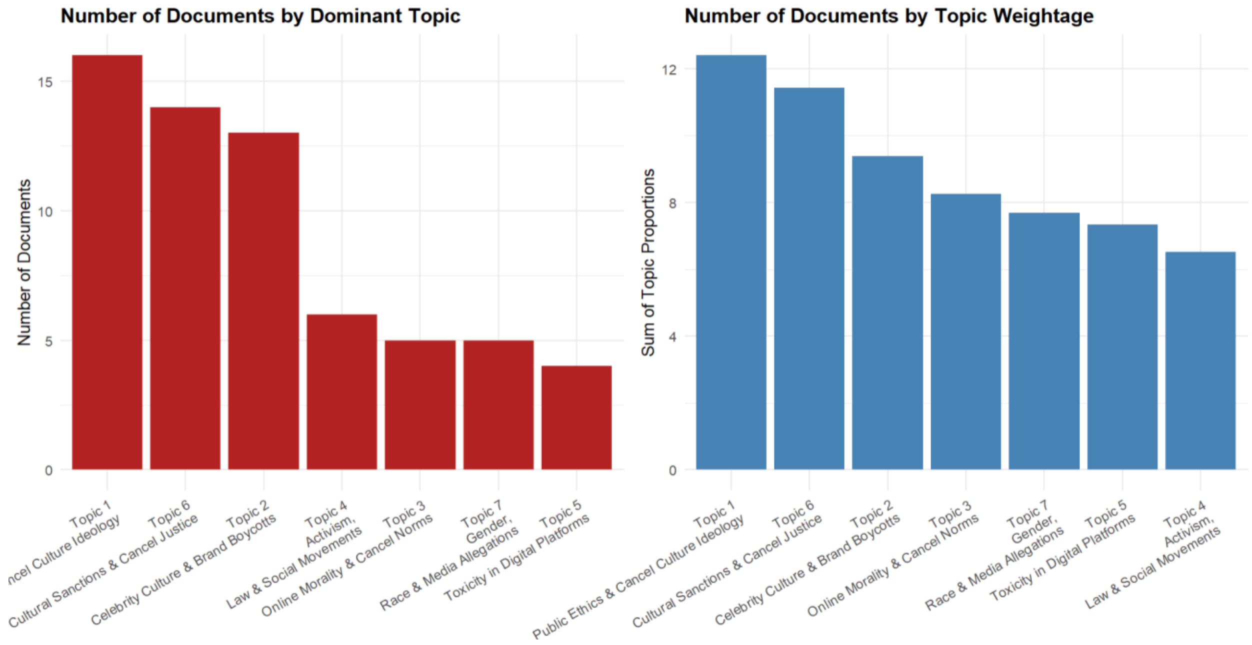


The figure displays the word count (bars) and term weight (lines) for the top keywords in each of the 7 LDA topics derived from the cancel culture dataset. Each subplot corresponds to one topic, showing how frequently key terms appear (bar height) and their relative importance in the topic distribution (blue line). This dual view helps identify which words are both common and semantically central to each topic, aiding in meaningful topic labeling and interpretation.

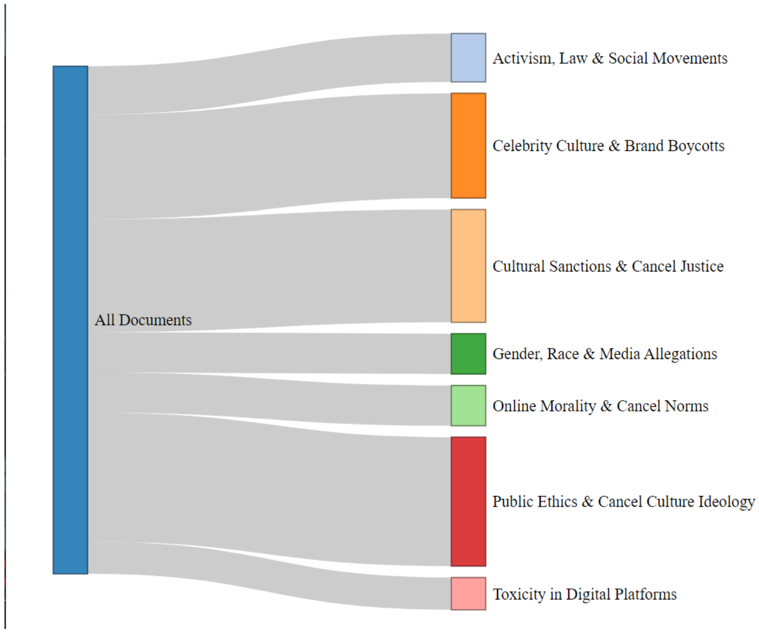
Number of document by Dominant topic and Weight:

The two bar charts show topic prominence in the cancel culture dataset. The first chart highlights how many articles are mainly about each topic, with Cancel Culture Ideology and Cultural Sanctions & Cancel Justice being the most dominant. Topics like Toxicity in Digital Platforms and Law & Social Movements appear less often as primary themes.

The second chart sums topic weight across all articles, reflecting overall influence. Even less dominant topics still contribute significantly. Cancel Culture Ideology and Cultural Sanctions again rank highest, confirming their central role in the discourse.

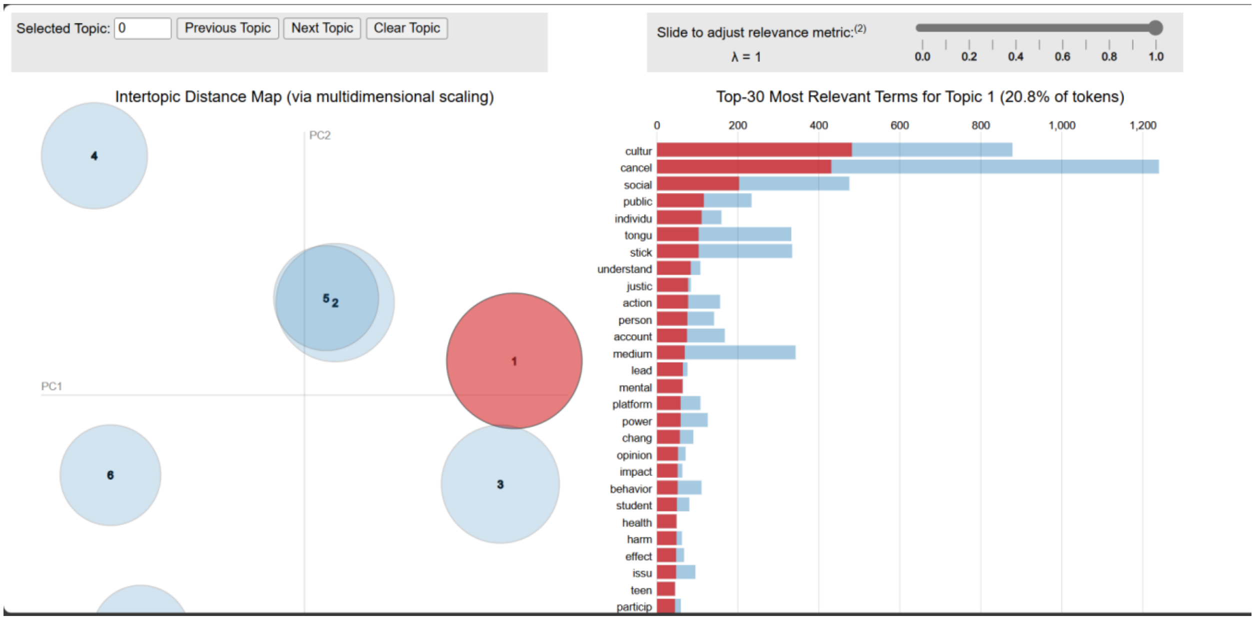


Sankey Diagram:



The Sankey diagram visualizes the flow of all documents into their assigned dominant topics. Each stream represents how many articles are linked to each of the seven identified themes. The widest flows correspond to Public Ethics & Cancel Culture Ideology, Cultural Sanctions & Cancel Justice, and Celebrity Culture & Brand Boycotts, indicating their prominence in the dataset. This visual effectively illustrates the thematic distribution and relative volume of documents across topics.

LDAvis Output:

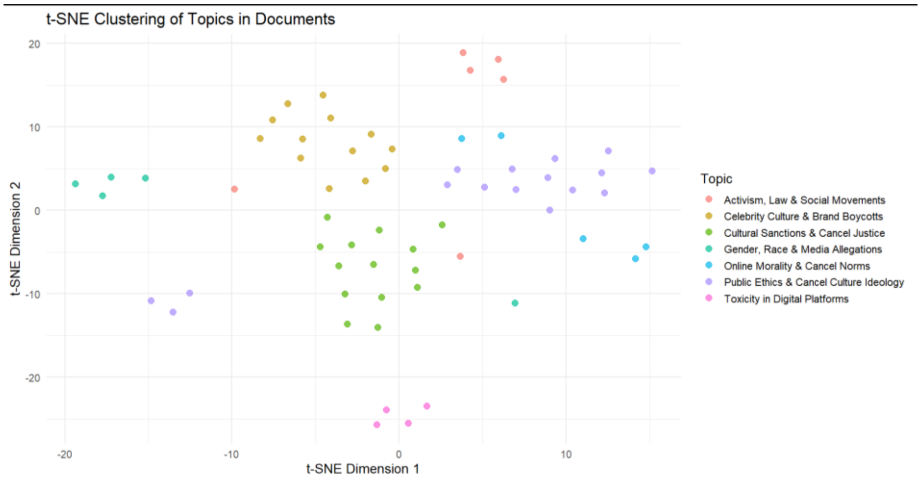

Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 1

The LDAvis visualization offers an interactive overview of topic structure. On the left, the Intertopic Distance Map (via multidimensional scaling) shows how distinct or overlapping the topics are—Topic 1 appears well-separated, indicating a unique theme. On the right, the Top-30 Most Relevant Terms for Topic 1 are shown, with bars representing frequency and relevance. Key terms like culture, cancel, social, and public suggest this topic aligns with "Public Ethics & Cancel Culture Ideology." The visualization helps interpret topic meanings and evaluate their separation in semantic space.

t-SNE Clustering:



The t-SNE plot shows how documents cluster by topic. Each color represents a topic, with clear groupings indicating distinct themes like *Celebrity Culture* and *Cultural Sanctions*. Some overlaps suggest mixed-topic content, supporting the effectiveness of the LDA model.