Topic Modeling 2

Chang Lu

2024-12-09

```
# Load necessary libraries
library(tidyverse)
library(topicmodels)
library(tidyr)
library(dplyr)
library(ggplot2)
library(wordcloud)
library(tm)
library(textrank)
library(tidytext)
library(ggforce)
library(factoextra)

# Load dataset and clean the data
movies <- read.csv("~/Desktop/615topicmodeling/movie_plots.csv", stringsAsFactors = FALSE)
movies <- movies %>% filter(!is.na(Plot))  # Remove rows with missing Plot
```

Creating the Document-type Matrix

```
# Tokenize words and create document-term matrix
plot_word_counts <- movies %>%
  unnest_tokens(word, Plot) %>%
  count(Movie.Name, word, sort = TRUE) %>%
  ungroup()

plots_dtm <- plot_word_counts %>%
  cast_dtm(Movie.Name, word, n)

# Check DTM dimensions
dim(plots_dtm)
```

```
## [1]  1063 15001
```

Creating LDA model

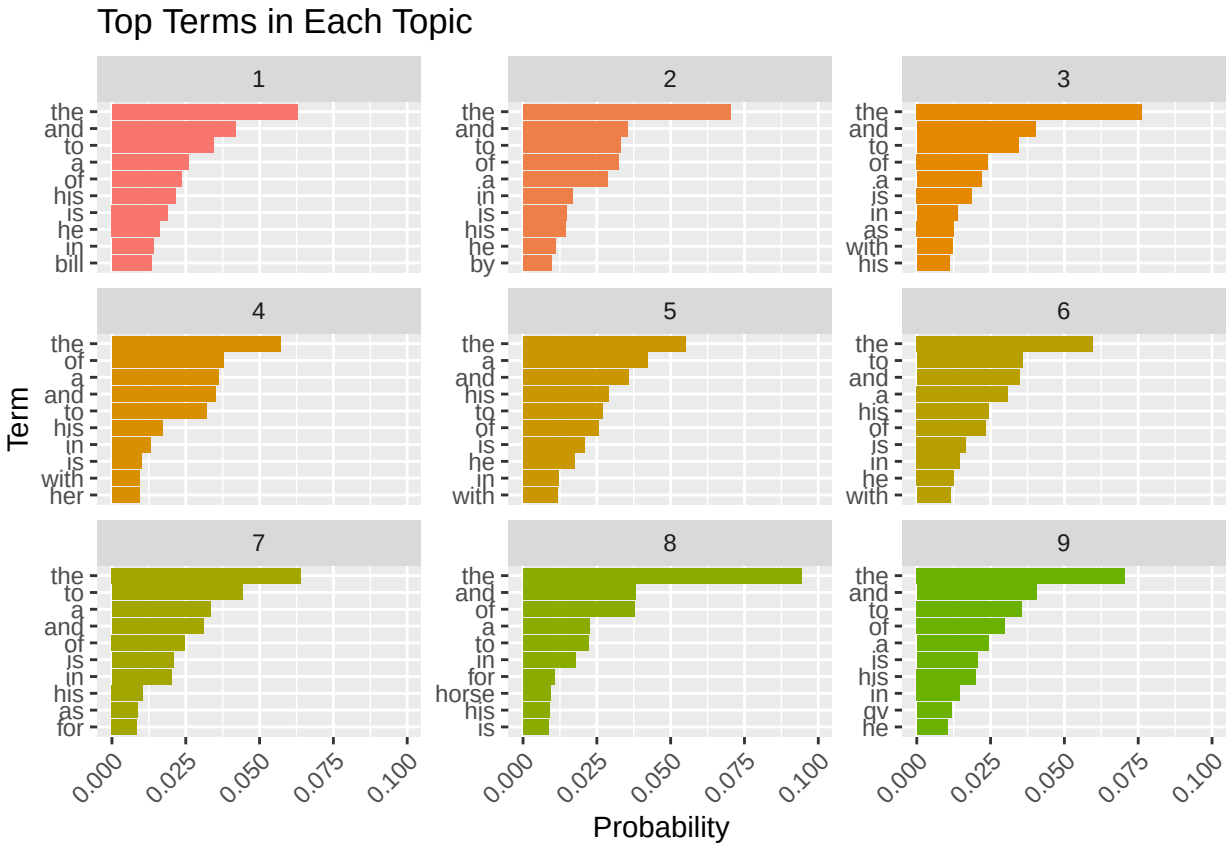
```
# Set up and run LDA model with 30 topics
set.seed(1234)
plots_lda <- LDA(plots_dtm, k = 30, control = list(seed = 1234))
```

```r
# Extract topic-term matrix (beta values)
topics <- tidy(plots_lda, matrix = "beta")

# Get top terms for each topic
top_terms <- topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(topic, -beta)

# View top terms
head(top_terms)
```

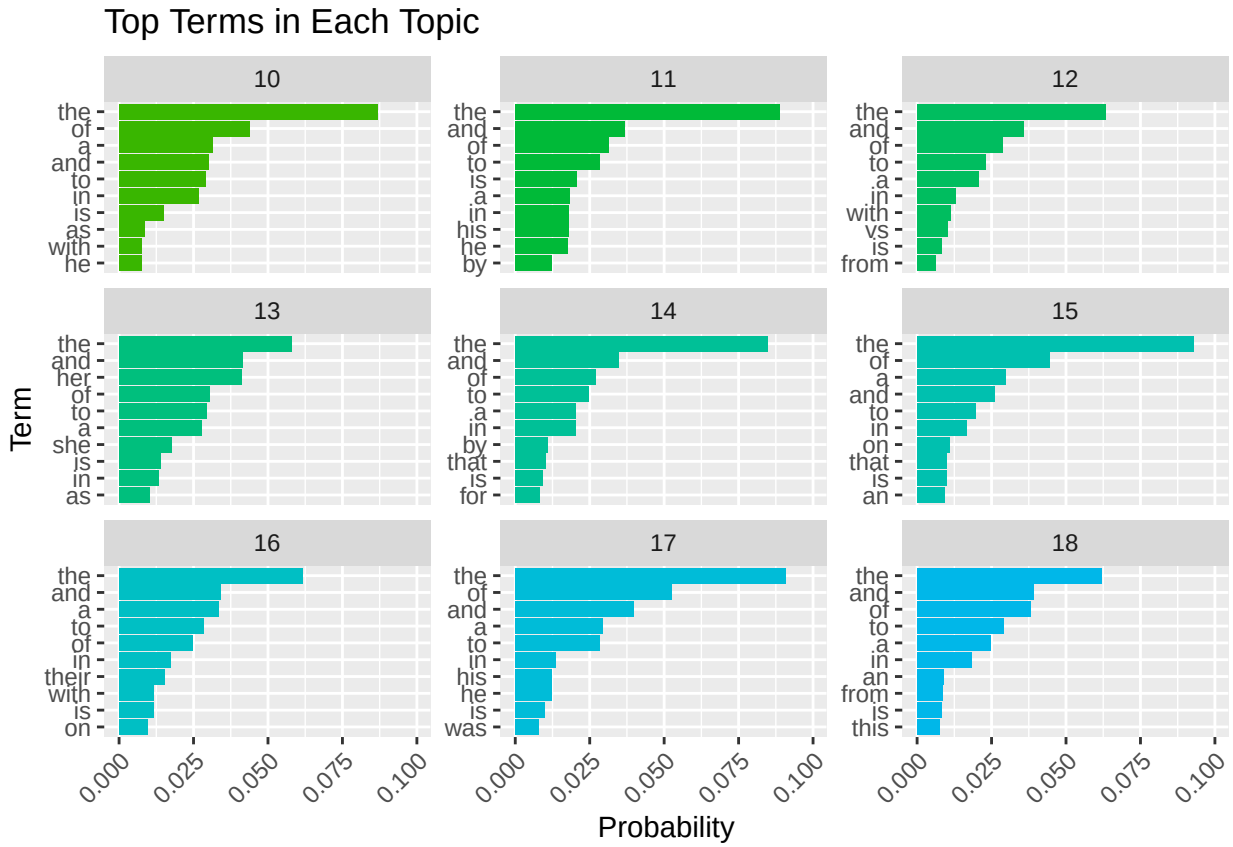
```
## # A tibble: 6 x 3
##   topic term    beta
##   <int> <chr>  <dbl>
## 1     1 the   0.0627
## 2     1 and   0.0419
## 3     1 to    0.0344
## 4     1 a     0.0259
## 5     1 of    0.0235
## 6     1 his   0.0216
```

Visualize Top Terms by Topic

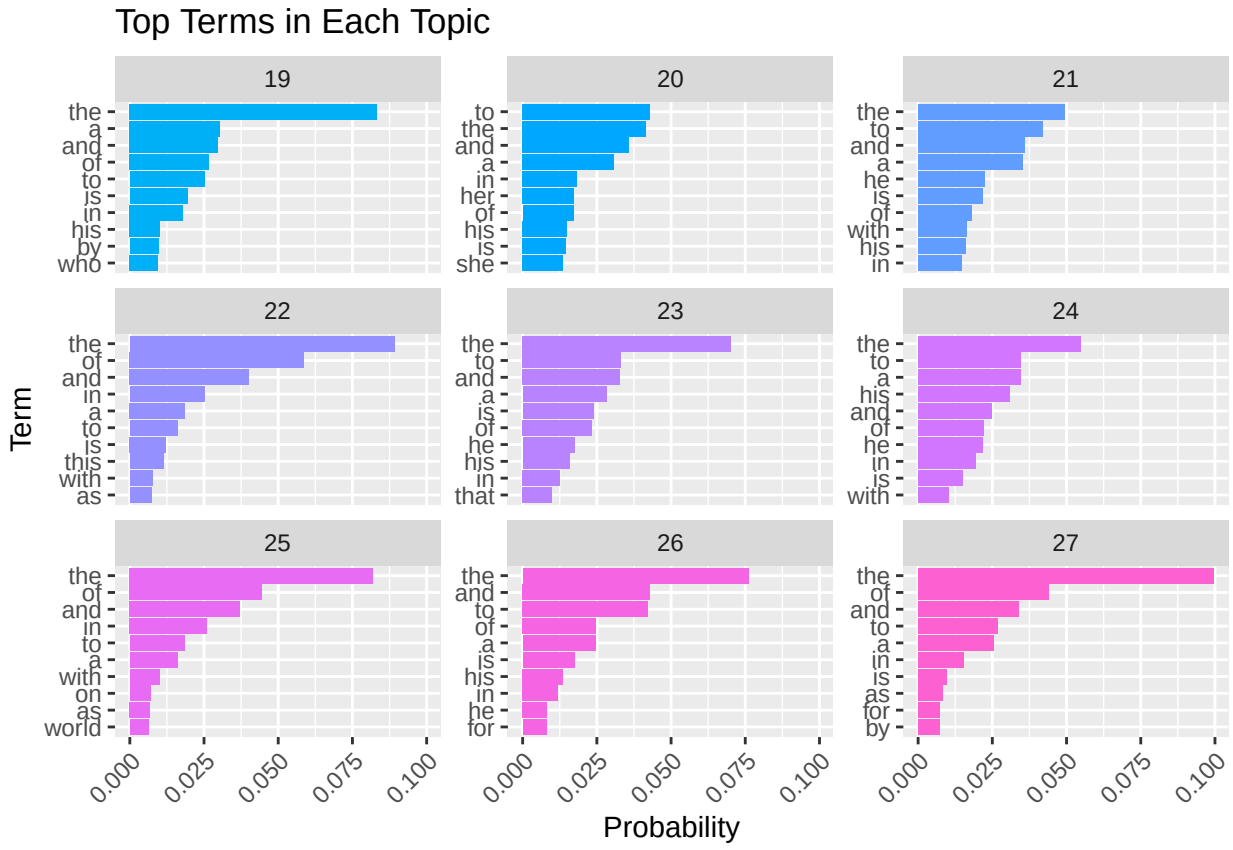
```r
# Visualization: Top terms for topics with improved layout
top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap_paginate(~ topic, scales = "free_y", ncol = 3, nrow = 3, page = 1) +
  scale_y_reordered() +
  scale_x_continuous(labels = scales::number_format(accuracy = 0.001)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(
    title = "Top Terms in Each Topic",
    x = "Probability",
    y = "Term"
  )
```



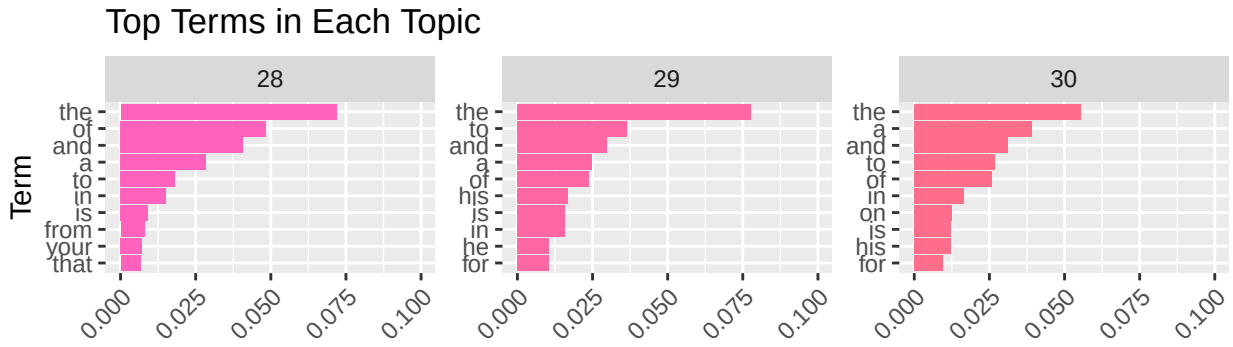

```
top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap_paginate(~ topic, scales = "free_y", ncol = 3, nrow = 3, page = 2) +
  scale_y_reordered() +
  scale_x_continuous(labels = scales::number_format(accuracy = 0.001)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(
    title = "Top Terms in Each Topic",
    x = "Probability",
    y = "Term"
  )
```



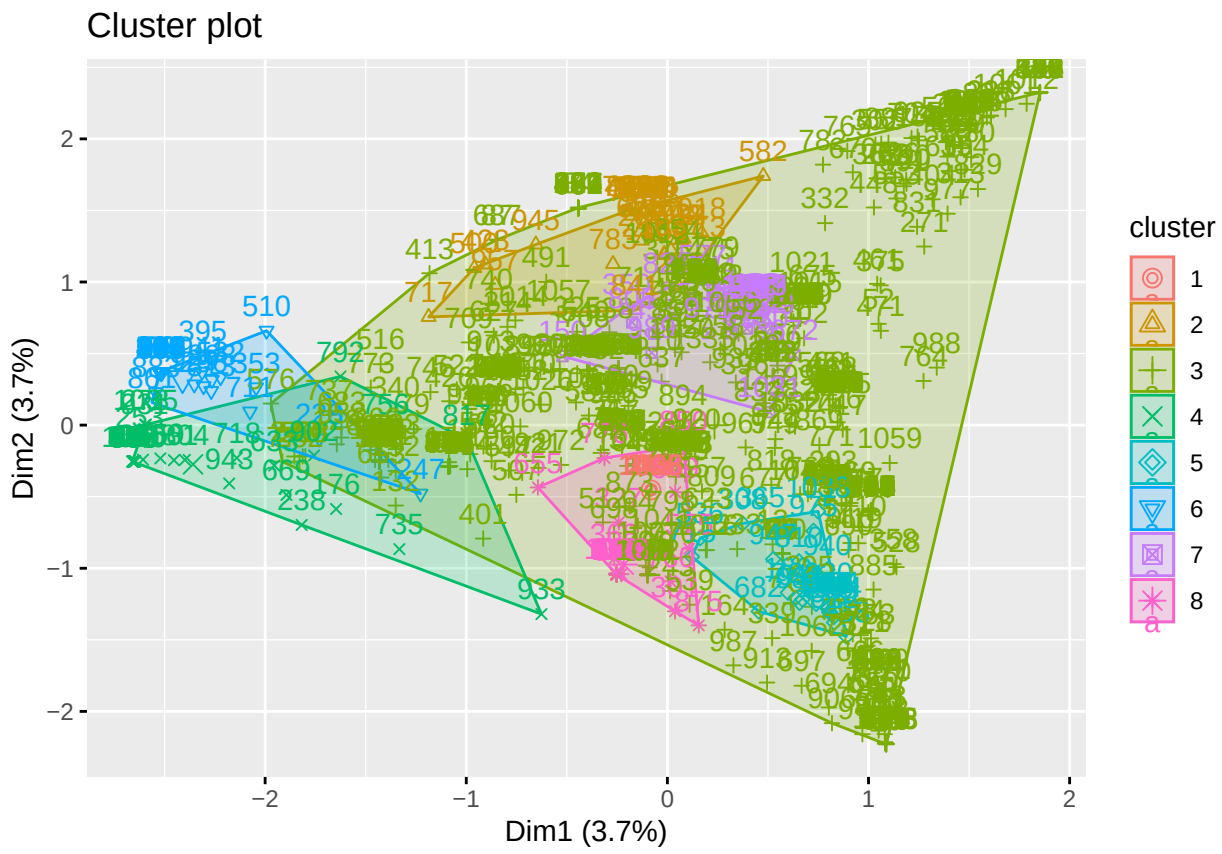

```
top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap_paginate(~ topic, scales = "free_y", ncol = 3, nrow = 3, page = 3) +
  scale_y_reordered() +
  scale_x_continuous(labels = scales::number_format(accuracy = 0.001)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(
    title = "Top Terms in Each Topic",
    x = "Probability",
    y = "Term"
  )
```




```
top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap_paginate(~ topic, scales = "free_y", ncol = 3, nrow = 3, page = 4) +
  scale_y_reordered() +
  scale_x_continuous(labels = scales::number_format(accuracy = 0.001)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(
    title = "Top Terms in Each Topic",
    x = "Probability",
    y = "Term"
  )
```




```r
# Extract the document-topic distribution matrix (gamma values)
plots_gamma <- tidy(plots_lda, matrix = "gamma")

# Transform the document-topic matrix into a wide format for clustering
plots_gamma_wider <- plots_gamma %>%
  pivot_wider(names_from = topic, values_from = gamma)

# Perform K-means clustering on the document-topic distributions
set.seed(1234)
plots_gamma_wider_no_na <- plots_gamma_wider %>% drop_na()
cluster <- kmeans(plots_gamma_wider_no_na %>% select(-document), centers = 8)

# Visualize the clusters using a scatter plot
library(factoextra)
fviz_cluster(cluster, data = plots_gamma_wider_no_na %>% select(-document))
```




```
# Add the cluster assignments to the original dataset
plots_gamma_wider$cluster <- cluster$cluster
```

Achieve each cluster information (e.g. cluster 6, cluster 2).

```
# Extract the list of movie titles in a specific cluster
cluster_6_names <- plots_gamma_wider %>%
  filter(cluster == 6) %>%
  pull(document)

cluster_2_names <- plots_gamma_wider %>%
  filter(cluster == 2) %>%
  pull(document)

# Display the movies
cluster_6_movies <- movies %>%
  filter(Movie.Name %in% cluster_6_names)

cluster_2_movies <- movies %>%
  filter(Movie.Name %in% cluster_2_names)

print(cluster_6_movies)
```

```
##                                    Movie.Name
```

```
## 1                          Rough and Ready
## 2                     Ride in the Whirlwind
## 3                           Ecstasy of Gold
## 4                       The Fighting Ranger
## 5                        Pals of the Saddle
## 6                           South of Sonora
## 7                  Overland with Kit Carson
## 8                           The Grey Vulture
## 9                         The Gospel of Lou
## 10                         Mojave Firebrand
## 11                  The Far Side of Jericho
## 12                 North from the Lone Star
## 13                    Riders of Black River
## 14                 The Return of Wild Bill
## 15                       Call of the Desert
## 16                       Code of the Cactus
## 17                          The Loaded Door
## 18                     Terror of the Plains
## 19                        South of Santa Fe
## 20                         Law of the North
## 21       "Lonesome Dove: The Outlaw Years"
## 22                     Days of Buffalo Bill
## 23                          Overland Riders
## 24                     Trail of Kit Carson
## 25                  Bandits of the Badlands
## 26 American Bandits: Frank and Jesse James
## 27                     Bad Men of Tombstone
## 28                       The Road to Denver
## 29         The Lone Rider in Texas Justice
## 30                             Law and Order
## 31                The Canyon of Missing Men
## 32                   Law of the Golden West
## 33                    Renegades of the West
## 34                       Change in the Wind
## 35                     The Mounted Stranger
## 36                      Under the Tonto Rim
## 37                     Billy the Kid Wanted
## 38           The Legend of the Lone Ranger
## 39         Strike of the Thunderkick Tiger
## 40                     Knight of the Plains
## 41                       King of Dodge City
## 42                       West of the Brazos
## 43                     Courage of the West
## 44                     Brand of the Outlaws
## 45                    The Dream of Alvareen
## 46                          The Bushwhackers
## 47                             The Law Rides
##
## 1
## 2
## 3
## 4
## 5
## 6
```

```
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14 The Return of Wild Bill  :  To get possession of choice ranch lands, Matt and Jake Kilgore frame S
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
## 31
## 32
## 33
## 34
## 35
## 36
## 37
## 38
## 39
## 40
## 41
## 42
## 43
## 44
## 45
## 46
## 47
```

```
print(cluster_2_movies)
```

```
##                                                                Movie.Name
## 1                                                  The Vengeance Trail
## 2                                  Behind the Action in 'Biker Boyz'
## 3                                                  "The Re-Inventors"
## 4                                                  Thieves of Fortune
## 5                                Saga of a Crew 2008 Special Edition
## 6                                     Dale Evans: Queen of the West
## 7                                                  The Painted Desert
## 8                                  Battle Jitni: The Danger Element
## 9                                            Four Guns to the Border
```

```
## 10                                Race Across the Sky: The Leadville Trail 100
## 11                                                      Eve & the Bigger Apple
## 12                                                    A Christmas to Remember
## 13 An Evening with Walter Hill & Lawrence Gordon. A Tribute to Andrew Laszlo
## 14                                                             Circle of Fury
## 15                                    National Geographic Inside: Cat & Mouse
## 16                                                               The Artifice
## 17                                                    Lipstick and Bullets
## 18                                                         Hard to Give Up
## 19                                                    Sélection officielle
## 20                                Bullitt and the Mystery of the Devil's Root
## 21                                         Standpoint: The Diamond War
## 22                                            "Because of Meeting You"
## 23                                      Joshua Bailey and the Island of Death
## 24                                                          Made in the USA
## 25                                                       Tiger Zero Three
## 26                                          Compton: The Antwon Ross Story
## 27                                                      The Soul of Buddha
## 28                                           In the Land of Fire & Ice
## 29                                              Hit the Ground Running
## 30                                                  The Danger Element
## 31                                "Beijing 2008: Games of the XXIX Olympiad"
## 32                                                        The Subject
## 33                                                     Double Crossed
## 34                                  The Complete History of the New York Jets
## 35                                The Turtle Stone: The Legacy of Abbott Farm
## 36                                                    Streets of Laredo
## 37                                                         Art of War 3
## 38                                                  Blood of the Hunter
## 39                                                Brothers in the Saddle
## 40                                          "Chronexia and the Eight Seals"
## 41                  The Adventures of Young Indiana Jones: Love's Sweet Song
## 42                                                  When the City Sleeps
## 43                                                         Reincarnation
## 44                                                "Pride and Prejudice"
## 45                                                      Tracing Skylines
## 46                                                The French Encounter
## 47                                                   Oath of Vengeance
## 48                                                They Were Not Divided
## 49                                                Death Run to Istanbul
## 50                          Television's Opening Night: How the Box Was Born
## 51                                                    The Indy Wrestler
## 52              TSF Channel Presents How to Really Pack a Backpack Parts 1-3
## 53                                      Freestyle Fighting Championship XV
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
```

```
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
## 31
## 32
## 33
## 34
## 35
## 36
## 37
## 38
## 39
## 40
## 41
## 42
## 43
## 44
## 45
## 46 The French Encounter  :  For the first time, a love story goes beyond national boundaries. When J
## 47
## 48
## 49
## 50
## 51
## 52
## 53
```

Making word cloud.

```
# Function to generate a word cloud for each topic
generate_wordcloud <- function(topic_number, topics_data) {
  # Filter terms for the specified topic
  topic_terms <- topics_data %>%
    filter(topic == topic_number) %>%
```

```
    arrange(desc(beta)) %>%
    slice_max(beta, n = 30) # Select top 30 words

  # Generate the word cloud
  wordcloud(words = topic_terms$term,
            freq = topic_terms$beta,
            max.words = 30,
            random.order = FALSE,
            colors = brewer.pal(8, "Dark2"),
            scale = c(4, 0.5))
}

# Generate word clouds for each topic (Example: first 5 topics)
par(mfrow = c(2, 3)) # Set up a layout to display multiple word clouds
for (i in 1:5) {
  generate_wordcloud(i, topics)
  title(paste("Topic", i))
}
```

What I Have Learned

1. Visualization by topic : Each topic is characterized by its top terms (words with the highest probabilities), which are visualized in bar plots. These help interpret the general themes or subjects of each topic in the data. Also, bar plots and cluster visualizations provide an overview of how topics are distributed and how documents relate to each cluster.

2. Code warning problem: Thanks to Prof.Haviland's in-class code, I can fix the warning problem.

3. Word clouds: Word clouds are generated to visually summarize the top terms in each topic, giving a quick sense of prominent themes.