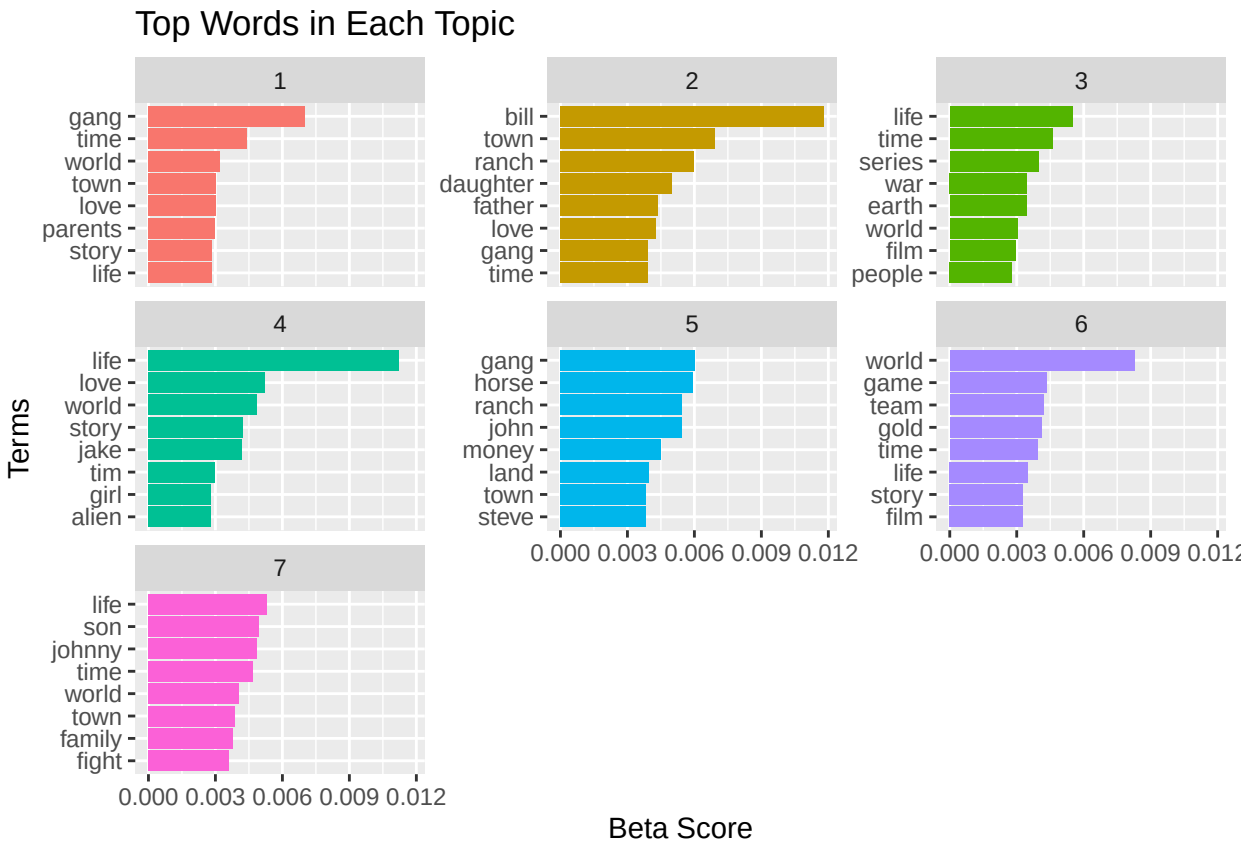
615 topic modeling

```r
# Load necessary libraries
library(dplyr)
library(stringr)
library(tidytext)
library(tm)
library(topicmodels)
library(ggplot2)
library(tidyr)
library(wordcloud)
library(RColorBrewer)

# 1. Load and preview the dataset
movies_data <- read.csv("movie_plots_with_genres.csv")
glimpse(movies_data)  # View the dataset structure for a quick overview
```

```
## Rows: 1,077
## Columns: 4
## $ row        <dbl> 31, 87, 146, 197, 314, 448, 489, 506, 521, 571, 680, 715, 7~
## $ Movie.Name <chr> "Pioneers of the West ", "The Infiltrators ", "\"Graviton: ~
## $ Genre      <chr> "western", "action", "sci-fi", "action", "history", "histor~
## $ Plot       <chr> "Pioneers of the West  :  Caught by the Piutes, pony Expres~
```

```r
# 2. Extract the title and detailed plot text
movies_data <- movies_data %>%
  mutate(
    Title_Extracted = str_extract(Plot, "^[^:]+"),  # Extract text before the colon as the title
    Plot_Only = str_replace(Plot, "^[^:]+: ", "")  # Extract text after the colon as the detailed plot
  )

# Add a unique row identifier
movies_data <- movies_data %>%
  mutate(row = row_number())  # Add row number as a unique identifier

# 3. Remove redundant information from the plot text
movies_data <- movies_data %>%
  rowwise() %>%
  mutate(Cleaned_Plot = str_remove_all(Plot, fixed(Movie.Name)))

# 4. Tokenize plot text and filter stop words
data("stop_words")
tokenized_movies <- movies_data %>%
  unnest_tokens(word, Cleaned_Plot) %>%
  anti_join(stop_words, by = "word")  # Remove stop words

# 5. Create a Document-Term Matrix (DTM)
```

```r
dtm <- tokenized_movies %>%
  count(row, word, sort = TRUE) %>%
  cast_dtm(row, word, n)  # Convert to a Document-Term Matrix

# 6. Fit an LDA model for topic modeling
num_topics <- 7  # Set the number of topics
set.seed(2024)  # Ensure reproducibility
lda_model <- LDA(dtm, k = num_topics, control = list(seed = 2024))

# 7. Extract top terms for each topic and visualize
topic_terms <- tidy(lda_model, matrix = "beta") %>%
  group_by(topic) %>%
  slice_max(beta, n = 8) %>%  # Extract top 8 terms for each topic
  ungroup()

ggplot(topic_terms, aes(x = reorder_within(term, beta, topic), y = beta, fill = as.factor(topic))) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  facet_wrap(~ topic, scales = "free_y") +
  scale_x_reordered() +
  labs(title = "Top Words in Each Topic", x = "Terms", y = "Beta Score")
```




```r
# 8. Assign topics to each movie and save results
assigned_topics <- tidy(lda_model, matrix = "gamma") %>%
  group_by(document) %>%
  slice_max(gamma, n = 1)  # Assign topic with the highest gamma
```

```r
movies_data <- movies_data %>%
  mutate(row = as.character(row))  # Ensure row is character to match document type

movies_with_topics <- movies_data %>%
  left_join(assigned_topics, by = c("row" = "document"))

# Save the results to a CSV file
write.csv(movies_with_topics, "movies_topics_final.csv", row.names = FALSE)

# 9. Perform clustering on topic distributions
topic_dist <- posterior(lda_model)$topics
num_clusters <- 5  # Set number of clusters
set.seed(999)  # Ensure reproducibility for clustering
kmeans_model <- kmeans(topic_dist, centers = num_clusters)

movies_with_topics$Cluster <- kmeans_model$cluster
movies_with_topics$Distance_to_Centroid <- apply(topic_dist, 1, function(row) {
  cluster_center <- kmeans_model$centers[kmeans_model$cluster[which.max(row)], ]
  sqrt(sum((row - cluster_center)^2))
})

# 10. Generate word clouds for each cluster
palette_colors <- brewer.pal(9, "Pastel1")  # Define a color palette

for (cluster_id in 1:num_clusters) {
  cluster_words <- tokenized_movies %>%
    filter(row %in% movies_with_topics$row[movies_with_topics$Cluster == cluster_id]) %>%
    count(word, sort = TRUE) %>%
    filter(nchar(word) <= 12)  # Remove overly long words to avoid fitting issues

  wordcloud(
    words = cluster_words$word,
    freq = cluster_words$n,
    min.freq = 2,
    max.words = 50,  # Limit words to avoid overcrowding
    scale = c(2, 0.3),  # Adjust font size range
    random.order = FALSE,
    colors = palette_colors
  )
}
```

bring white ranchers
gold discovers
return city love daughter military
wife son
takes evil john dance fight
kathy gang life
friend
outlaw
war george
ranch tim town
home team
army girl world story
cattle
stiles time land murder
johnny family father
game christmas mother
brother buck race day judge
rancher
dead johnny
film sheriff cattle
day takes tells people
save tom set girl friends war
secret daughter jim world wife joe
discovers life bill jack party
brother
ranch town love king
death time story gold
stage killed kill
marshal fight gang father
queen
escape rescue mary
game
outlaw
game sheriff
cattle daughter
red
series gold people real earth
action money story city army
bill
killed world land
john
home film time war son learn
set team
history
horse fight
ranch life town save
power
jack gang
named love past
brother father day dead
family girl takes death
packed moments
live learn henry
series friend
brother nevada escapes
nelson steve footage bank
discovers family love money
tom action
killed star
life tex meets
bud
story evil forced
world real master
fans
films bill
mind baby movie
mtv game gang time
relationship town
violence wilkins
property rick starr
women
secret battle
mary land tom team
mission steve
woman story forces
history war town action stop
vicky family
takes world day series
home
love life gang indian
learns john jake son
film meet
people ranch
meets black time friends
peter brother alien brothers
lives
set planet father events
night sam