

topic

Zilu Sun

2024-11-07

```
movie_plot <- read.csv("movie_plots.csv")
custom_stop_words <- tibble(word = c("jack", "jim", "steve", "tom", "jake", "john"))
all_stop_words <- bind_rows(stop_words, custom_stop_words)
tidy_plots <- movie_plot %>%
  unnest_tokens(word, Plot) %>%
  anti_join(all_stop_words)
```

```
## Joining with `by = join_by(word)`
```

```
tidy_plots <- tidy_plots %>%
  filter(!str_detect(word, "[:punct:]")) %>%
  filter(!str_detect(word, "[:digit:]"))
```

```
dtm <- tidy_plots %>%
  count(Movie.Name, word) %>%
  cast_dtm(Movie.Name, word, n)
```

```
word_counts <- tidy_plots %>%
  count(word, sort = TRUE)
wordcloud(words = word_counts$word, freq = word_counts$n, max.words = 100, random.order = FALSE, colors
```

```
## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : returns could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : agent could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : bring could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : outlaws could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : plans could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : indian could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : stop could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : control could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : tells could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : rescue could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : sam could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : brothers could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : events could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : government could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : makes could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : meet could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : white could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : escape could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : human could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : leader could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = word_counts$word, freq = word_counts$n, max.words
## = 100, : lives could not be fit on page. It will not be plotted.
```



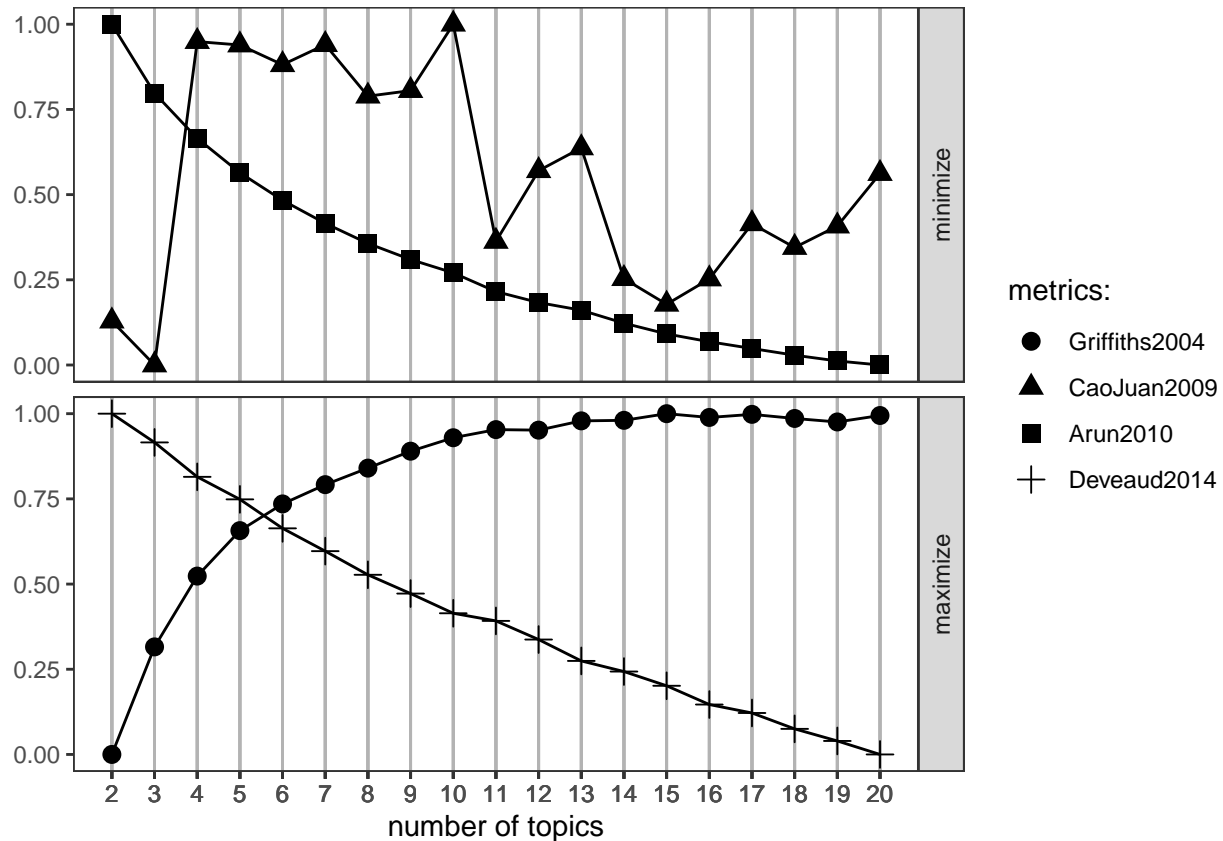
```
result <- FindTopicsNumber(
  dtm,
  topics = seq(2, 20, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234),
  mc.cores = 2L,
  verbose = TRUE
)
```

```
## fit models... done.
## calculate metrics:
## Griffiths2004... done.
## CaoJuan2009... done.
## Arun2010... done.
## Deveaud2014... done.
```

```
FindTopicsNumber_plot(result)
```

```
## Warning: The `scale` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the ldatuning package.
##   Please report the issue at <https://github.com/nikita-moor/ldatuning/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
```

generated.



Griffiths2004: Evaluates the log-likelihood of the LDA model fit; a higher value is better. CaoJuan2009: Measures the similarity between topics; a lower value is better, indicating greater distinction between topics. Arun2010: Assesses the divergence between topic distribution and word frequency distribution; a lower value is preferred. Deveaud2014: Evaluates model quality based on topic coherence; a higher value is better.

```
#according to FindTopicsNumber_plot(result), best k = 11 or 15
lda_model <- LDA(dtm, k = 11, control = list(seed = 123))
lda_gamma <- tidy(lda_model, matrix = "gamma")
lda_topics <- tidy(lda_model, matrix = "beta")

top_terms <- lda_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(topic, -beta)

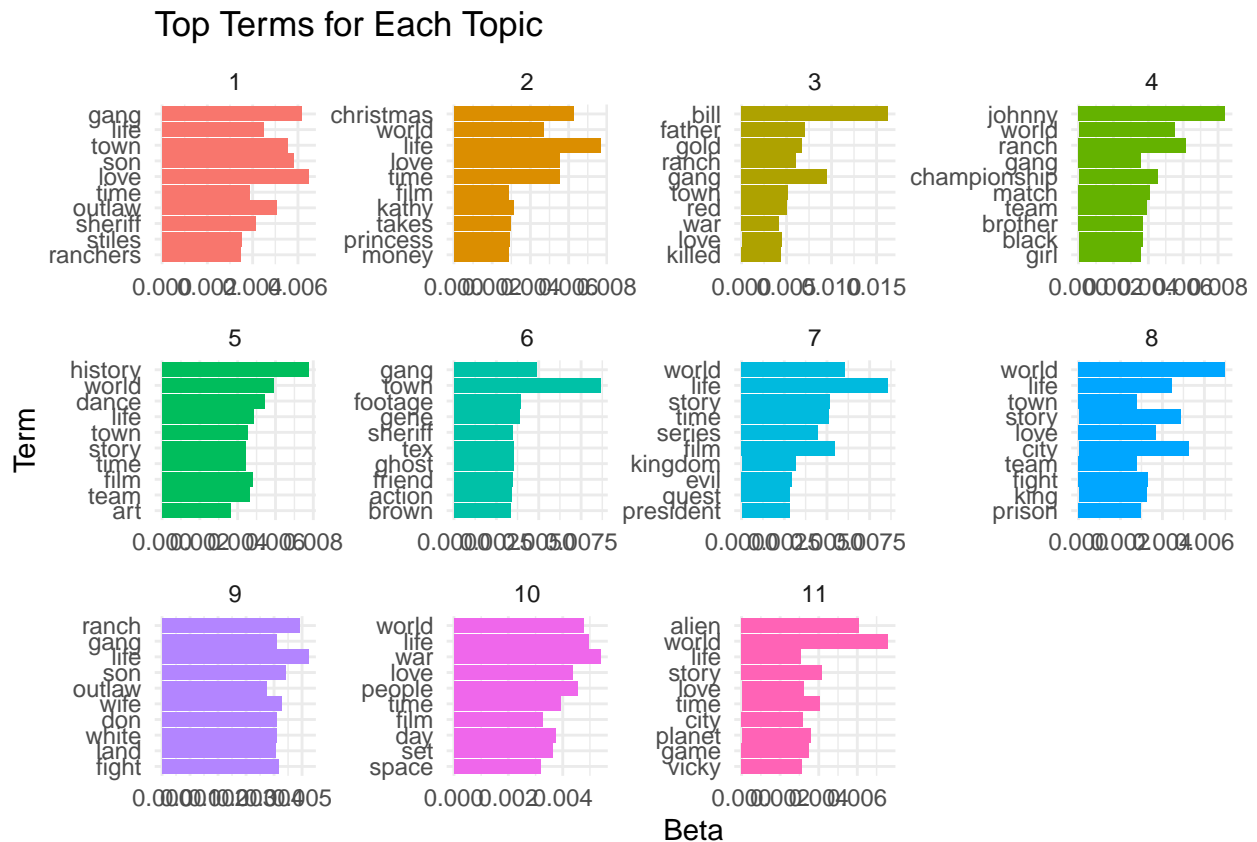
print(top_terms)
```

```
## # A tibble: 110 x 3
##   topic term      beta
##   <int> <chr>    <dbl>
## 1     1 love    0.00647
## 2     1 gang    0.00614
## 3     1 son     0.00580
```

```
## 4      1 town      0.00555
## 5      1 outlaw    0.00507
## 6      1 life      0.00446
## 7      1 sheriff   0.00413
## 8      1 time      0.00387
## 9      1 styles    0.00350
## 10     1 ranchers  0.00349
## # i 100 more rows
```

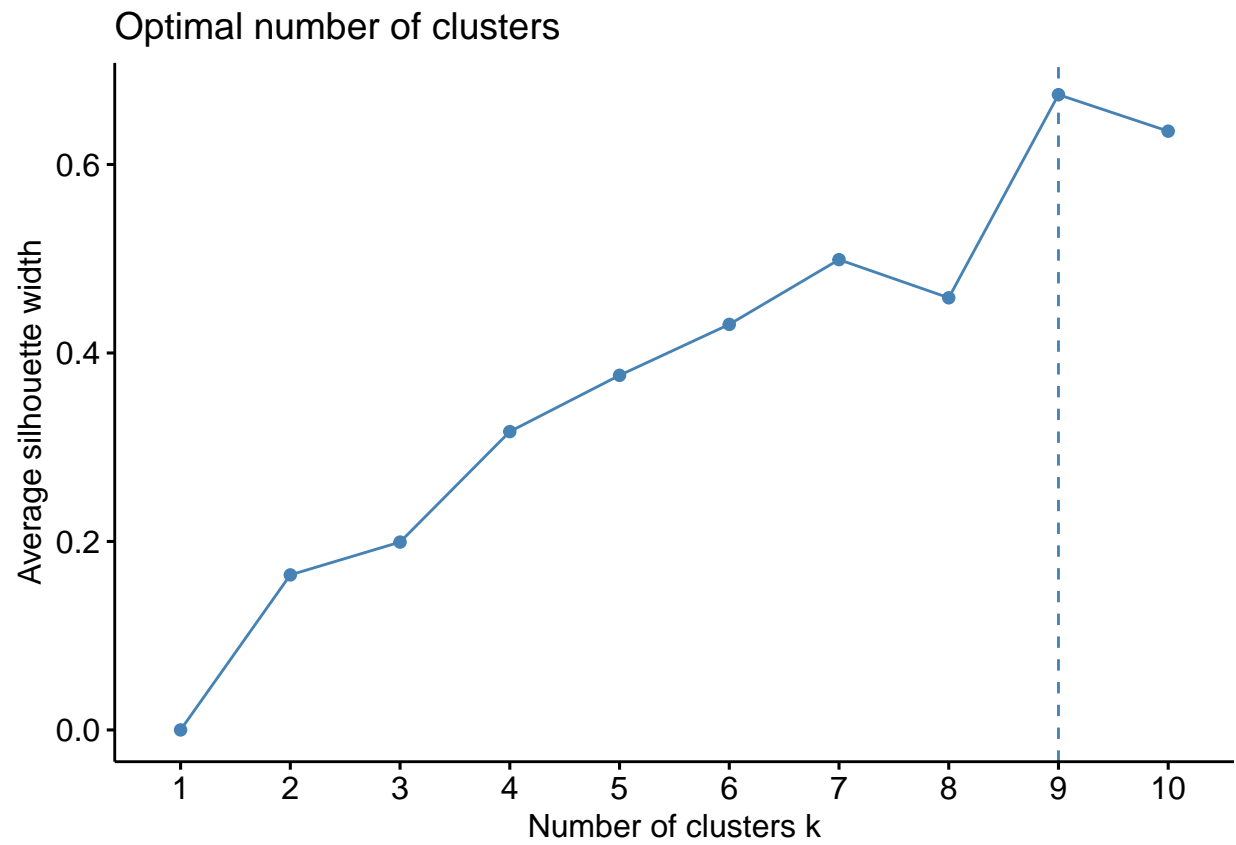
#beta plot

```
ggplot(top_terms, aes(x = reorder(term, beta), y = beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(x = "Term", y = "Beta", title = "Top Terms for Each Topic") +
  theme_minimal()
```



```
gamma_matrix <- lda_gamma %>% pivot_wider(names_from = topic,
                                             values_from = gamma)%>%
  drop_na()
gamma_matrix_numeric <- gamma_matrix %>%
  select(where(is.numeric))

fviz_nbclust(gamma_matrix_numeric, kmeans, method = "silhouette")
```



```
cluster <- kmeans(gamma_matrix_numeric, centers = 9)
fviz_cluster(cluster, data = gamma_matrix_numeric)
```

Cluster plot



Based on the top 10 keywords for each category, the five classes are categorized as:

Family Drama, History, Adventure, Romance & Morality, and Crime.

```
dominant_topic <- tibble(
  Movie.Name = rownames(dtm),
  dominant_topic = apply(gamma_matrix, 1, which.max)
)
```

```
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
```

```
movie_plot_with_topic <- movie_plot %>%
  left_join(dominant_topic, by = "Movie.Name")
movie_plot_with_topic <- movie_plot_with_topic %>%
  mutate(dominant_topic = case_when(
    dominant_topic == 1 ~ "Western and Family Life",
    dominant_topic == 2 ~ "Holidays and Romance",
    dominant_topic == 3 ~ "Crime and Revenge",
    dominant_topic == 4 ~ "Sports and Competition",
    dominant_topic == 5 ~ "History and Art",
    dominant_topic == 6 ~ "Western and Adventure",
    dominant_topic == 7 ~ "Fantasy and Quest",
    dominant_topic == 8 ~ "Urban Life and Conflict",
    dominant_topic == 9 ~ "Family and Struggles",
    dominant_topic == 10 ~ "War and Humanity",
    dominant_topic == 11 ~ "Aliens and Science Fiction",
    TRUE ~ NA_character_
  ))

write.csv(movie_plot_with_topic, "movie_with_topic.csv", row.names = FALSE)
```