Topic_Modeling_HW

Jie Fei

```
# install packages

# install.packages(c("tidyverse", "tm", "topicmodels", "ldatuning", "wordcloud", "quarto"))
# install.packages('tidytext')
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tm)
```

```
## Loading required package: NLP
##
## Attaching package: 'NLP'
##
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(topicmodels)
library(ldatuning)
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(quarto)
library(ggplot2)
library(tidytext)
```

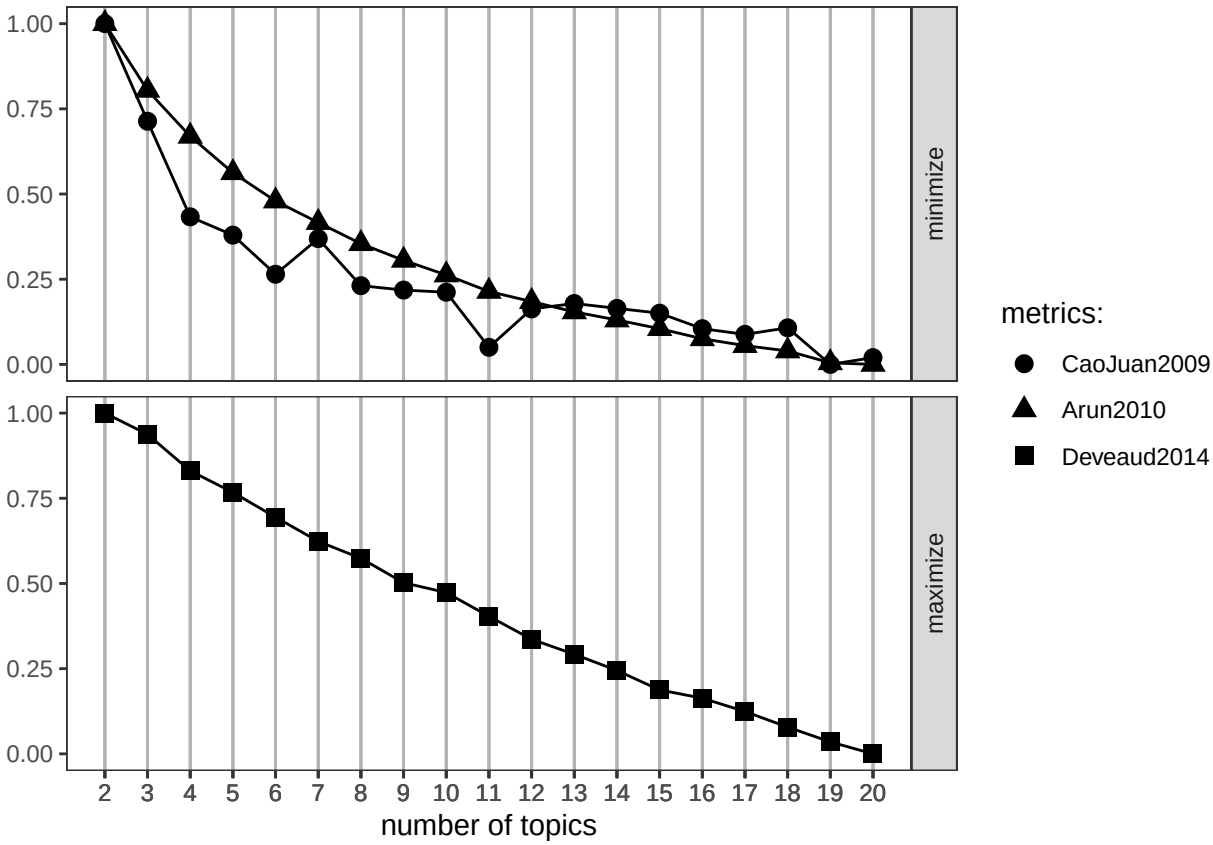
```
# load the dataset
movie_data <- read.csv("movie_plots.csv")
```

```r
# process text data
preprocess_text <- function(text) {
  text <- tolower(text)
  text <- removePunctuation(text)
  text <- removeNumbers(text)
  text <- removeWords(text, stopwords("english"))
  text <- stripWhitespace(text)
  return(text)
}

movie_data$Processed_Plot <- sapply(as.character(movie_data$Plot), preprocess_text)
```

```r
# create a corpus and Document-Term Matrix (DTM)
corpus <- Corpus(VectorSource(movie_data$Processed_Plot))
dtm <- DocumentTermMatrix(corpus, control = list(wordLengths = c(3, 15)))
```

```r
# determine optimal number of topics
result <- FindTopicsNumber(
  dtm,
  topics = seq(2, 20, by = 1),
  metrics = c("CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234)
)

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.
```



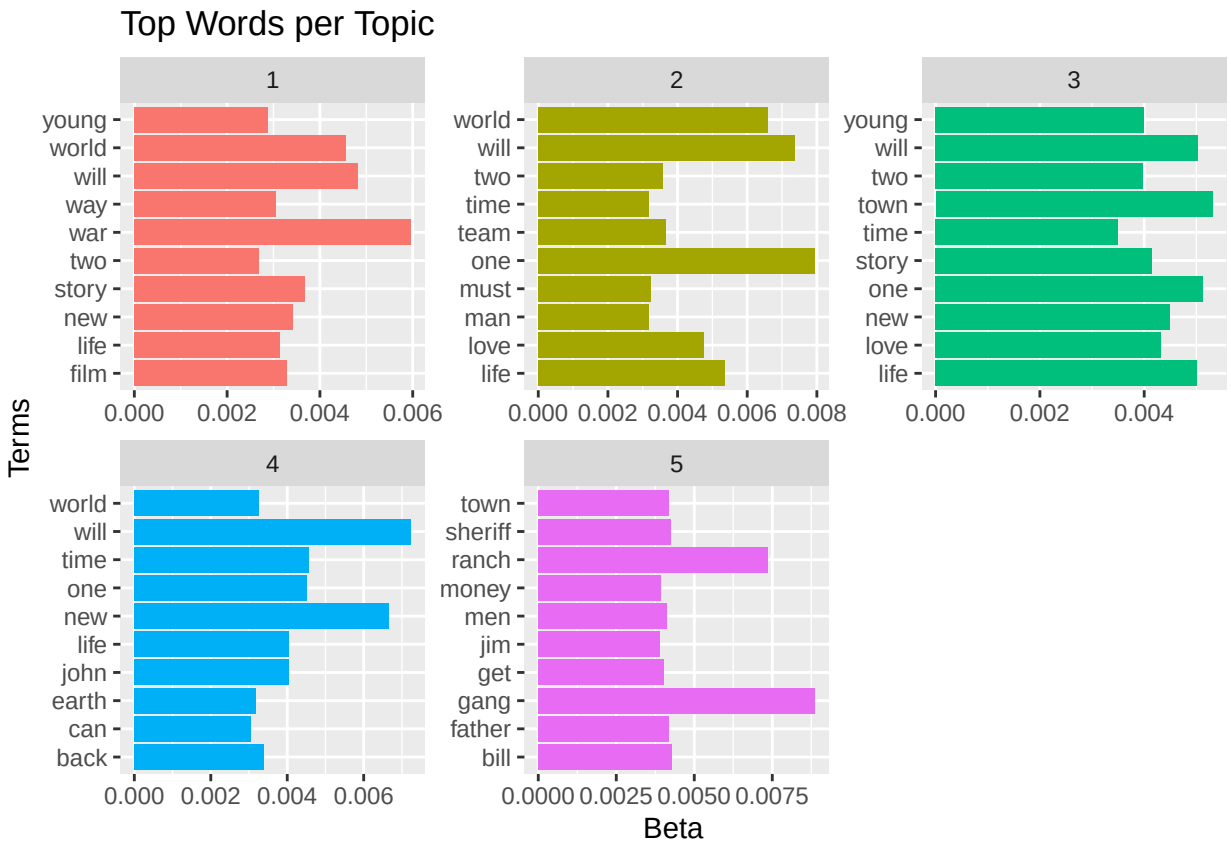

```
# fit the LDA model
optimal_k <- 5 # use the number determined from the scree plot
lda_model <- LDA(dtm, k = optimal_k, control = list(seed = 1234))
```

```
# extract topics and top words
terms <- terms(lda_model, 10) # top 10 words per topic
terms
```

```
##       Topic 1 Topic 2 Topic 3 Topic 4 Topic 5
##  [1,] "war"   "one"   "town"  "will"  "gang"
##  [2,] "will"  "will"  "one"   "new"   "ranch"
##  [3,] "world" "world" "will"  "time"  "bill"
##  [4,] "story" "life"  "life"  "one"   "sheriff"
##  [5,] "new"   "love"  "new"   "john"  "town"
##  [6,] "film"  "team"  "love"  "life"  "father"
##  [7,] "life"  "two"   "story" "back"  "men"
##  [8,] "way"   "must"  "young" "world" "get"
##  [9,] "young" "man"   "two"   "earth" "money"
## [10,] "two"   "time"  "time"  "can"   "jim"
```

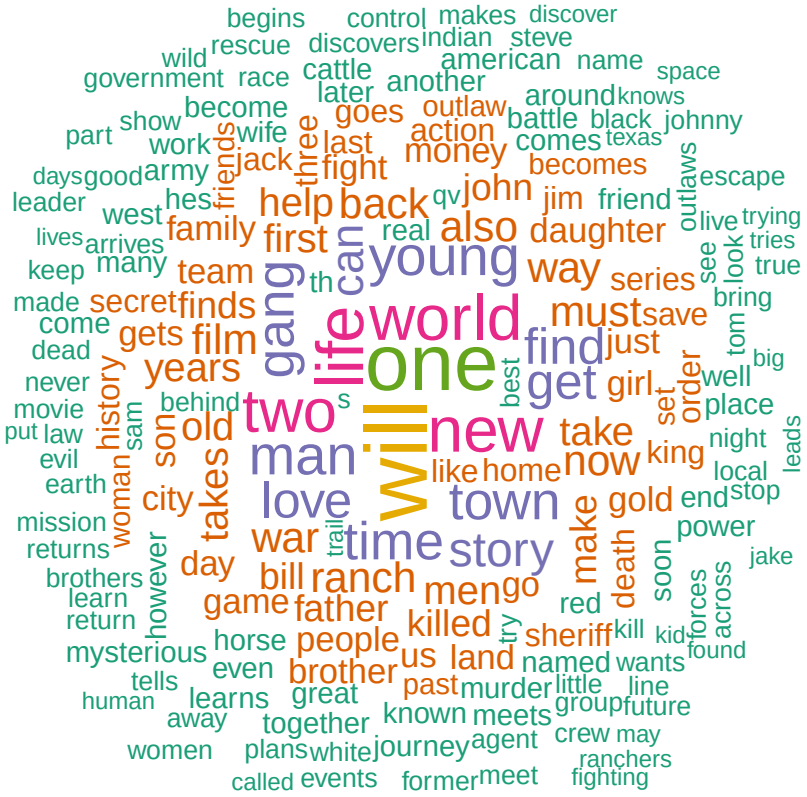
```
# visualize top words per topic
topics <- tidy(lda_model) %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
```

```
  arrange(topic, -beta)

ggplot(topics, aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title = "Top Words per Topic", y = "Beta", x = "Terms")
```




```
# The beta plots display the top terms and their probabilities for each topic generated by the LDA mode
```

```
# create word clouds

all_text <- paste(movie_data$Processed_Plot, collapse = " ")
wordcloud(words = names(table(unlist(strsplit(all_text, " ")))),
          freq = table(unlist(strsplit(all_text, " "))),
          min.freq = 2,
          max.words = 200,
          random.order = FALSE,
          colors = brewer.pal(8, "Dark2"))
```


begins
control
makes
discover
rescue
discovers
indian
steve
wild
american
name
government
race
cattle
another
space
later
around
knows
become
goes
outlaw
battle
black
johnny
part
show
wife
action
comes
texas
work
last
money
jack
three
fight
becomes
days
good
army
friends
escape
leader
hes
help
back
qv
john
jim
friend
outlaws
west
family
first
real
also
daughter
live
trying
lives
arrives
can
young
keep
many
team
th
way
series
see
look
tries
true
made
secret
finds
gang
life
world
must
save
bring
come
gets
film
one
find
just
tom
dead
years
best
get
girl
big
never
history
two
s
well
movie
behind
set
order
place
put
law
sam
old
will
new
take
night
evil
son
man
like
home
now
king
local
earth
woman
takes
city
love
town
gold
end
stop
mission
trail
war
time
story
make
power
returns
however
day
bill
ranch
men
go
death
soon
jake
brothers
learn
game
father
red
forces
across
return
killed
try
sheriff
kill
kid
mysterious
horse
people
us
land
named
wants
found
even
brother
tells
past
murder
little
line
human
learns
great
group
future
away
together
known
meets
women
plans
white
journey
agent
crew
may
ranchers
called
events
former
meet
fighting