Topic Modeling

Gary Wang

2024-11-01

```r
library(lexicon)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.2     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
```

```r
library(topicmodels)
library(tidytext)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(dplyr)
library(ggplot2)
library(wordcloud)
```

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

```r
library(RColorBrewer)
```

```r
movies <- read.csv("movie_plots.csv")
```

Group by Genre, summarize common words, and find genre frequency

```r
# Extract genres based on common keywords found in the plot descriptions
movies <- movies %>%
  mutate(Genre = case_when(
    str_detect(Plot, "(?i)science|space|experiment|future|alien|
              robot|technology|planet|human|new world|earth") ~ "Sci-Fi",
    str_detect(Plot, "(?i)love|romance|relationship|affair|
```

```
               wedding|couple|meets|girl") ~ "Romance",
    str_detect(Plot, "(?i)war|battle|army|soldier|conflict|
               military|enemy|tank|battlefield|dead") ~ "War",
    str_detect(Plot, "(?i)ghost|haunt|horror|fear|terror|
               scary|supernatural|creepy") ~ "Horror",
    str_detect(Plot, "(?i)crime|detective|murder|investigate|
               thriller|mafia|heist|mystery") ~ "Crime",
    str_detect(Plot, "(?i)action|fight|fighting|adventure|hero|explosion|
               battle|rescue") ~ "Action",
    str_detect(Plot, "(?i)comedy|funny|humor|laugh|joke|
               satire|parody") ~ "Comedy",
    str_detect(Plot, "(?i)history|historical|biography|true story|
               period drama|century|ancient") ~ "History",
    str_detect(Plot, "(?i)fantasy|magic|myth|legend|superhero|
               kingdom|evil") ~ "Fantasy",
    str_detect(Plot, "(?i)western|cowboy|wild west|sheriff|ranch|
               town|outlaw") ~ "Western",
    str_detect(Plot, "(?i)documentary|docu|true events|reality|
               biopic") ~ "Documentary",
    str_detect(Plot, "(?i)sport|game|team|match|championship|
               wrestling") ~ "Sport",
    str_detect(Plot, "(?i)home|people|brother|daughter|brothers|
               friend|wife|son|father|mother") ~ "Family",
    TRUE ~ "Other"
  ))

# Calculate the frequency of each genre
genre_frequency <- movies %>%
  count(Genre, name = "Frequency")

# Tokenize the plots and remove stop words
plot_words <- movies %>%
  unnest_tokens(word, Plot) %>%
  anti_join(get_stopwords()) %>%
  count(Genre, word, sort = TRUE)
```

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

```
# Group by Genre, summarize common words, and find genre frequency
nested_data <- plot_words %>%
  group_by(Genre) %>%
  summarize(
    Words = paste(unique(word), collapse = ", ")
  ) %>%
  left_join(genre_frequency, by = "Genre")

view(nested_data)
```

Create a Document Term Matrix

```
dtm <- plot_words %>%
  cast_dtm(Genre, word, n)
```

```r
dtm
```

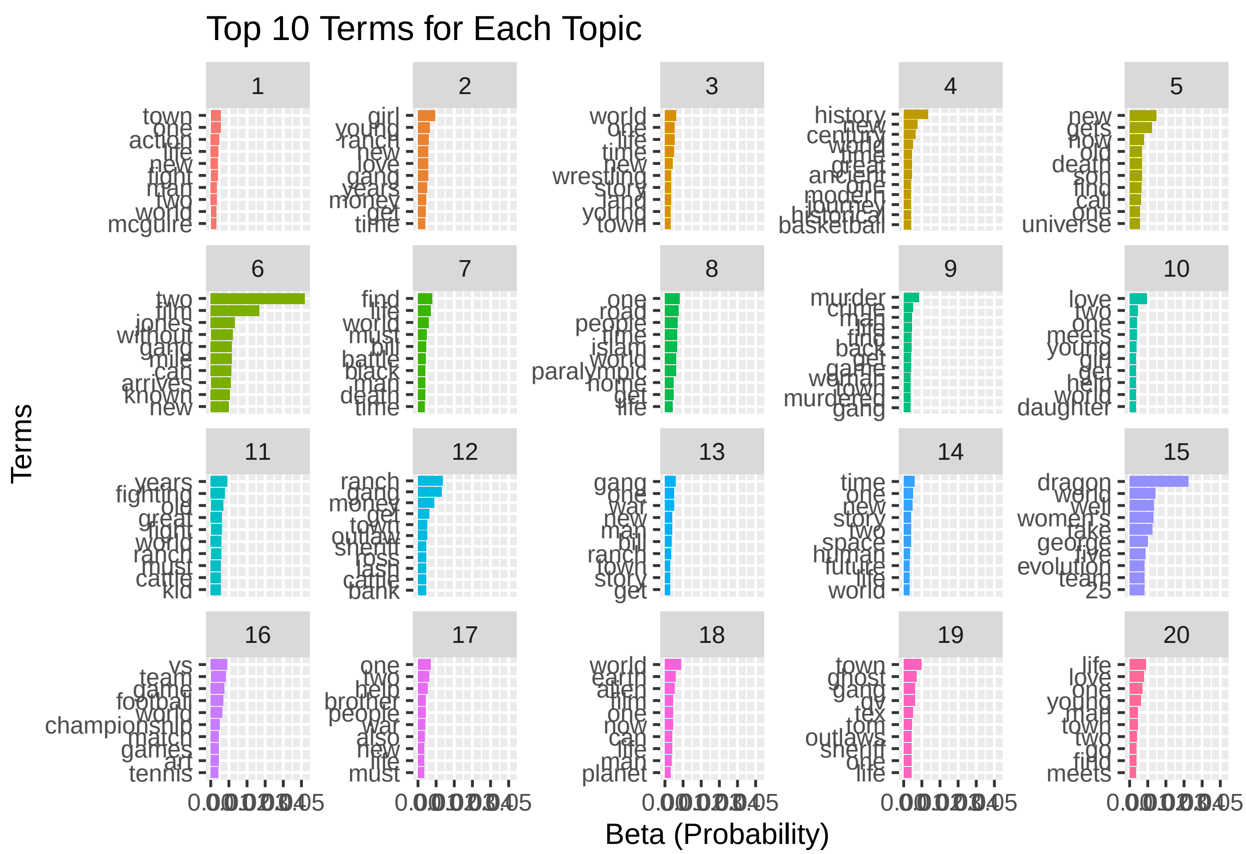
```
## <<DocumentTermMatrix (documents: 14, terms: 14842)>>
## Non-/sparse entries: 30812/176976
## Sparsity           : 85%
## Maximal term length: 17
## Weighting          : term frequency (tf)
```

Fir the LDA Model

```r
# Since we have 14 genres, we try to set k close to the number of genres.
k <- 20

# Fit the LDA model with k = 14
lda_model <- LDA(dtm, k = k, control = list(seed = 999))

# Extract the topic-term matrix
topics <- tidy(lda_model, matrix = "beta")

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

print(top_terms)
```

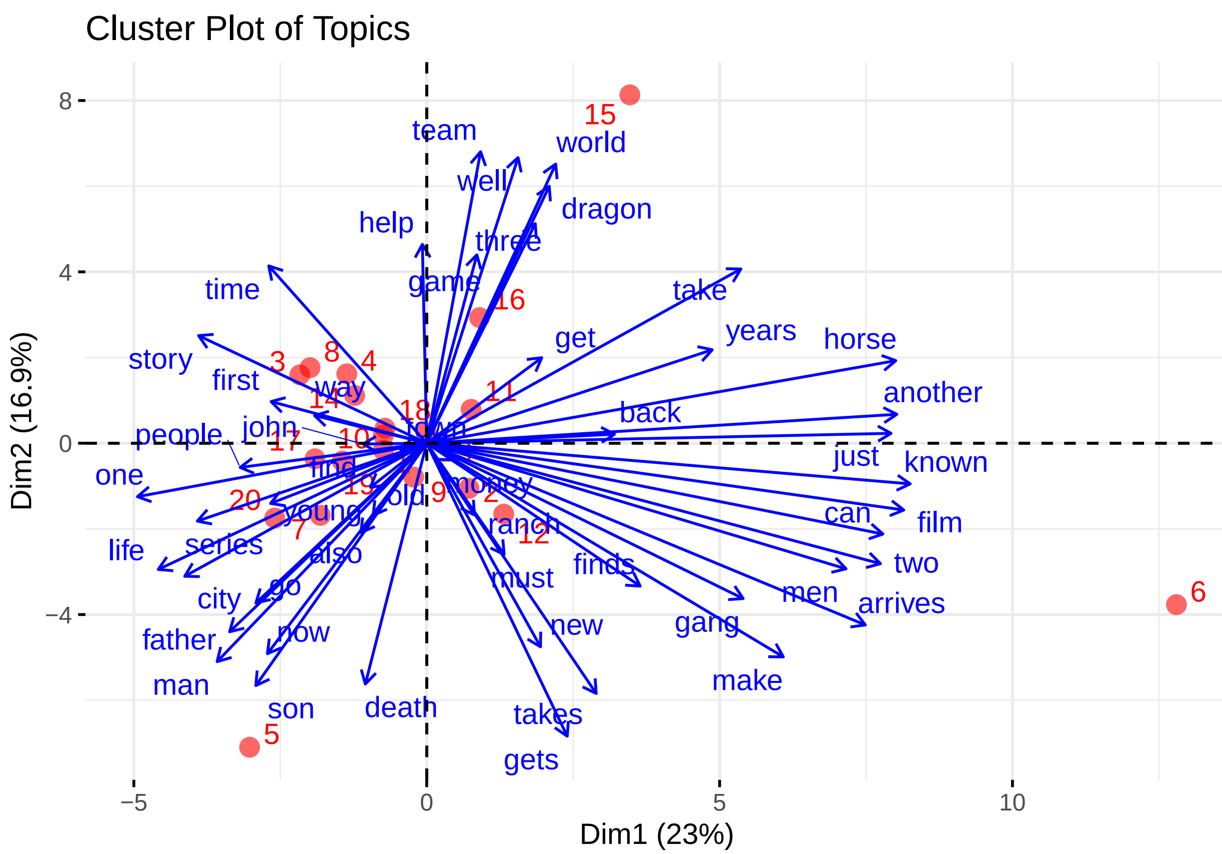
```
## # A tibble: 205 x 3
##    topic term       beta
##    <int> <chr>     <dbl>
##  1     1 town    0.00552
##  2     1 one     0.00545
##  3     1 action  0.00474
##  4     1 life    0.00427
##  5     1 new     0.00387
##  6     1 fight   0.00378
##  7     1 man     0.00347
##  8     1 two     0.00339
##  9     1 world   0.00300
## 10     1 mcguire 0.00296
## # i 195 more rows
```

```r
# Plot the top terms for visualization
ggplot(top_terms, aes(x = reorder_within(term, beta, topic), y = beta,
                      fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free_y") +
  coord_flip() +
  scale_x_reordered() +
  labs(title = "Top 10 Terms for Each Topic", x = "Terms",
       y = "Beta (Probability)")
```



Cluster Plot and Gamma Plot

```
# Filter the topic-term matrix to include only the top 50 most frequent terms
filtered_topics <- topics %>%
  group_by(term) %>%
  summarize(total_beta = sum(beta)) %>%
  top_n(50, total_beta)

topic_term_matrix <- topics %>%
  filter(term %in% filtered_topics$term) %>%
  spread(term, beta) %>%
  column_to_rownames(var = "topic")

# Perform PCA on the filtered topic-term matrix
pca_result <- prcomp(topic_term_matrix, scale. = TRUE)

# Plot the PCA result
fviz_pca_biplot(pca_result,
                repel = TRUE,
                col.var = "blue",
                col.ind = "red",
                pointsize = 3,
                alpha.ind = 0.6,
                title = "Cluster Plot of Topics")
```

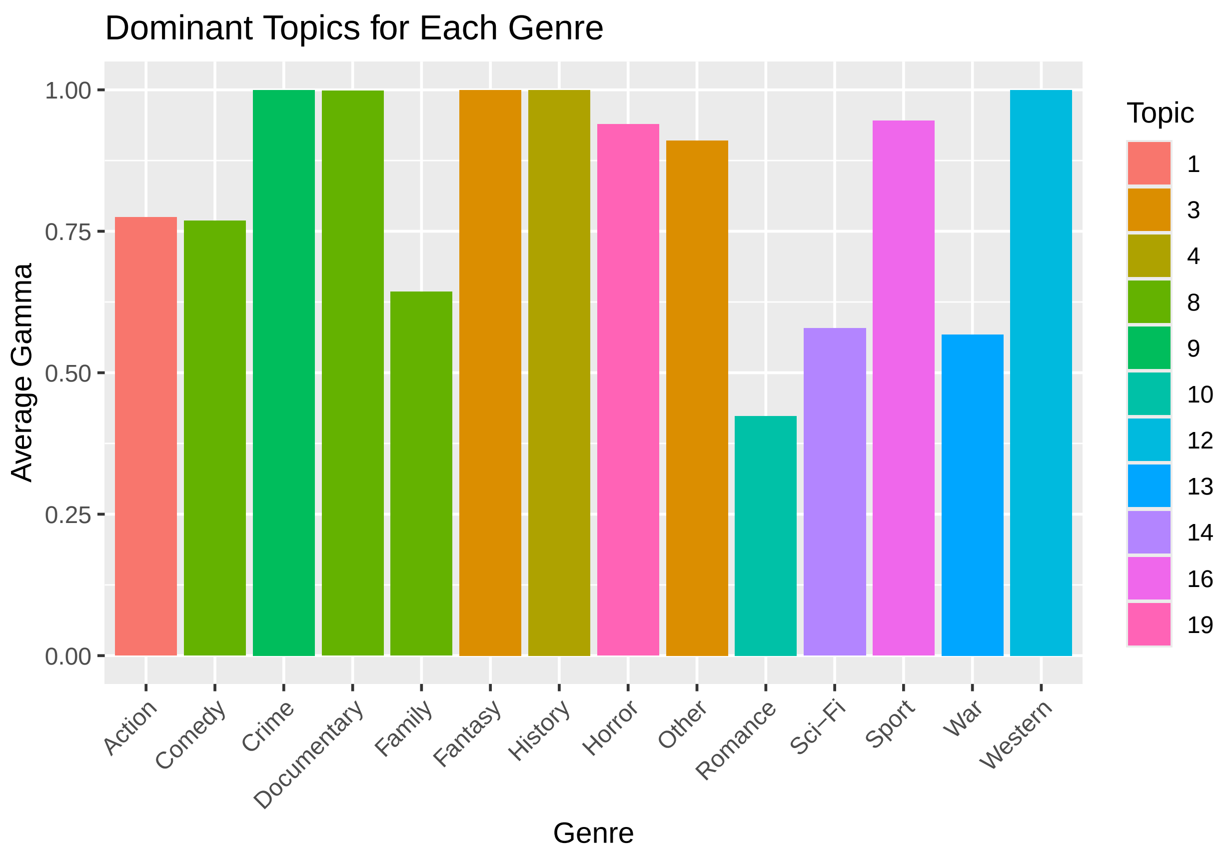



```
# Extract the document-topic distribution (gamma values)
doc_topic_matrix <- tidy(lda_model, matrix = "gamma")

# Aggregate by genre and calculate the average gamma per topic
genre_topic_distribution <- doc_topic_matrix %>%
  rename(Genre = document) %>%
  group_by(Genre, topic) %>%
  summarize(avg_gamma = mean(gamma, na.rm = TRUE), .groups = "drop")

# Identify the dominant topic for each genre
dominant_topics <- genre_topic_distribution %>%
  group_by(Genre) %>%
  slice_max(avg_gamma, n = 1) %>%
  ungroup()

print(dominant_topics)
```

```
## # A tibble: 14 x 3
##    Genre       topic avg_gamma
##    <chr>       <int>     <dbl>
##  1 Action          1     0.775
##  2 Comedy          8     0.769
##  3 Crime           9     1.00
##  4 Documentary     8     0.998
##  5 Family          8     0.643
```

```
##   6 Fantasy         3     1.00
##   7 History         4     1.00
##   8 Horror         19     0.939
##   9 Other           3     0.910
##  10 Romance        10     0.423
##  11 Sci-Fi         14     0.579
##  12 Sport          16     0.945
##  13 War            13     0.567
##  14 Western        12     1.00
```

```r
ggplot(dominant_topics, aes(x = Genre, y = avg_gamma, fill = factor(topic))) +
  geom_col() +
  labs(title = "Dominant Topics for Each Genre",
       x = "Genre",
       y = "Average Gamma",
       fill = "Topic") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

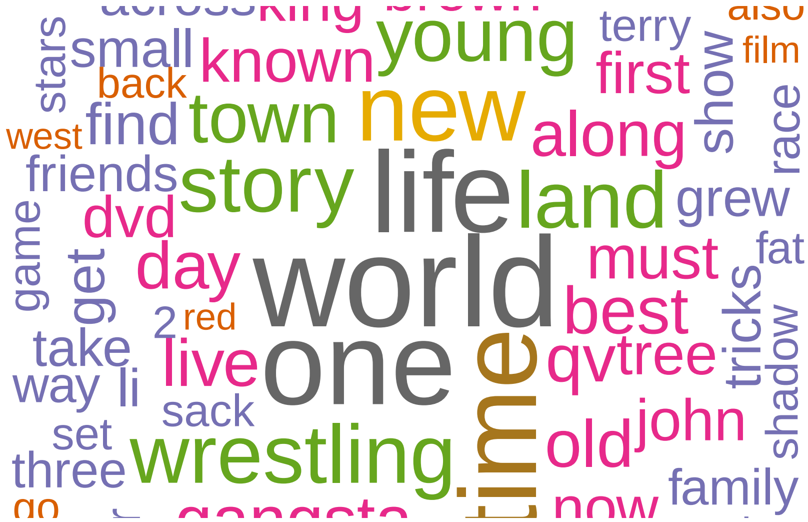


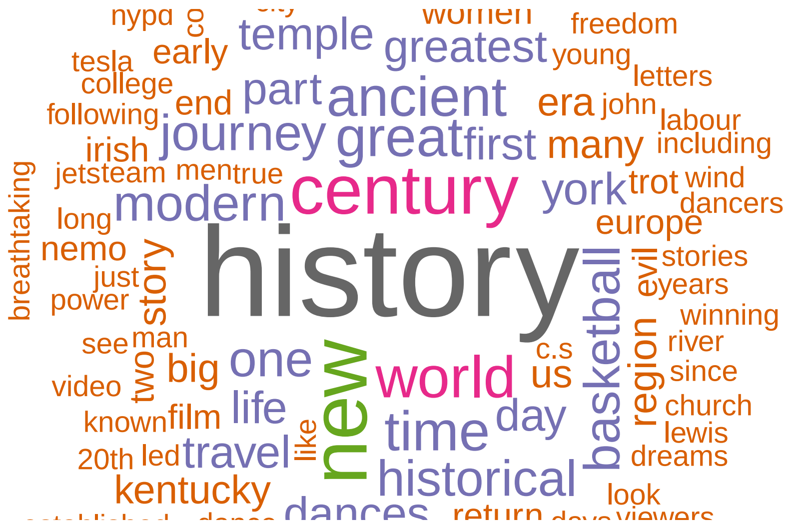
Create a Word Clound for each genre

```r
# Set up color palette
palette <- brewer.pal(8, "Dark2")

# Create word clouds for each topic in a loop
k <- 14
for (i in 1:k) {
```

```
  # Filter words for the current topic
  topic_words <- topics %>%
    filter(topic == i) %>%
    arrange(desc(beta))

  # Create the word cloud for the current topic
  wordcloud(words = topic_words$term,
            freq = topic_words$beta,
            max.words = 100,
            random.order = FALSE,
            colors = palette)
}
```


become
redlife
money
dead
work
father
york
knows
never
like
outlaw
set
son
old
love
three
take
must
even
city
sister
also
new
ranch
back
family
men
make
gun
time
soon
man
girl
help
bill
use
years
latertry
friend
billy
goes
young
arrives
secret
land
get
falls
sheriff
takes
gang
know
buck
town
gold
finds
first
wild
chief
war
boy

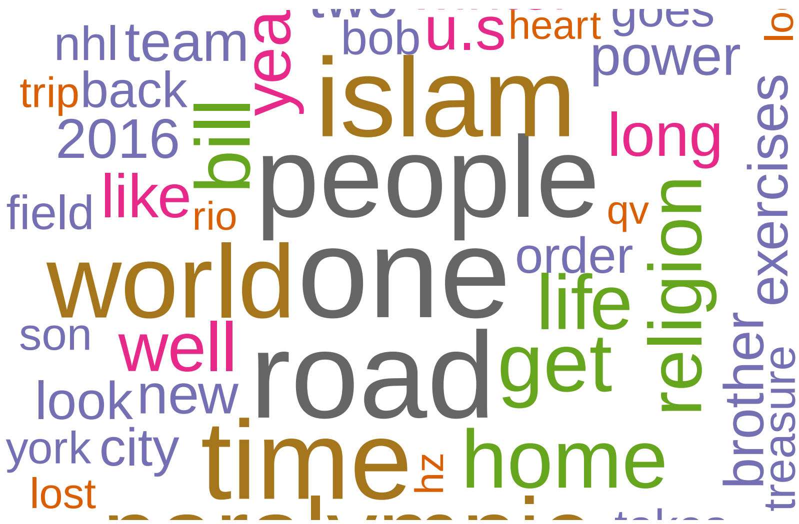

world
life
one
time
new
wrestling
story
land
young
town
along
known
day
live
best
must
first
old
john
qvtree
dvd
small
find
friends
grew
family
take
way
three
sack
set
get
game
stars
show
race
tricks
shadow
fat
terry
film
back
west
red
2
li


history
century
new
world
basketball
ancient
great
journey
modern
historical
time
temple
greatest
part
first
york
one
life
travel
day
dances
many
era
kentucky
region
story
big
us
trot
europe
early
end
irish
nemo
film
evil
two
freedom
young
letters
john
labour
including
wind
dancers
stories
years
winning
river
since
church
lewis
dreams
look
viewers
women
nypd
tesla
college
following
jetsteam
men
true
long
just
power
breathtaking
see
man
video
known
20th
led
like
c.s
return

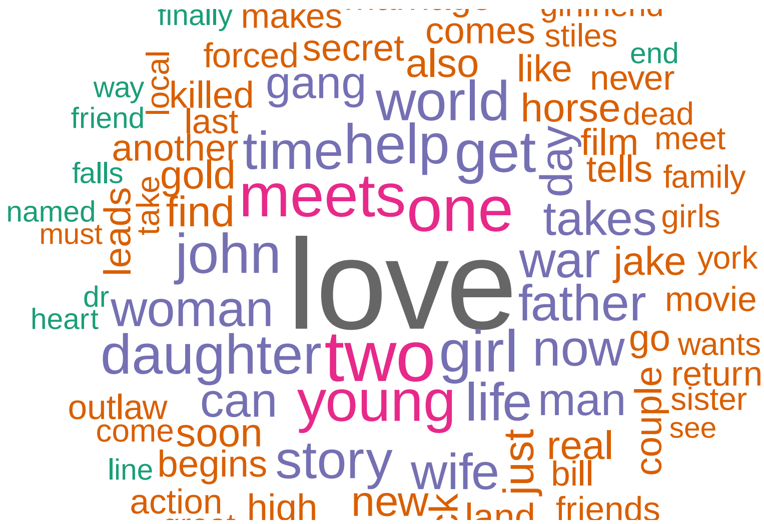

return
episode
christmas
violent
energy
group
people
machine
brothers
series
takes
two
kill
go
mission
sides
every
old
death
treasure
like
storm
year
city
johnny
son
now
re
last
religion
maverick
telling
bring
get
new
call
slaves
time
father
men
life
way
young
wife
jack
map
gets
find
half
black
many
can
prison
one
man
lies
united
players
self
use
night
tim
universe
brings


two
film
jones
without
known
another
gang
can
arrives
mile
working
order
water
jackson
end
unknown
heart
horse
hunt
seven
tells
tom
new
learns
gets
mind
dawn
take
roberts
bring.
single
kills
wonderful
carrying
losing
meanwhile
together
stop
takes
men
establish
biker
wild
back
place
riders
away
unlikely
trail
years
just
office
events
times
win
mine
learn
arrested
bar
pop
big
goes
must
andy
asked
looking
works
discover
make
given
system
group
pay
vowed
trying
trip
leader
wants
head
frame
career
holds
sense
traveling
explores
engineer
name
tales
aid
decided
world
challenge
brewing


indians
daughter
sell
goes
behind
way
dan
lead
forced
takes
death
must
film
lost
team
black
world
go
son
dance
man
story
learns
get
can
find
save
tom
dead
time
events
secret
john
life
real
army
serial
west
new
kill
bill
war
many
bob
like
battle
two
land
fight
north
city
place

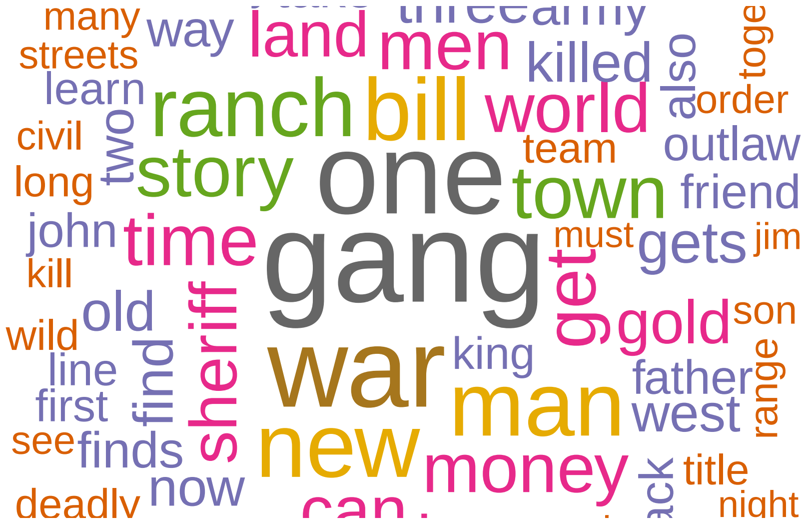

nhl
team
yea
bob
u.s
heart
power
trip
back
islam
2016
bill
long
exercises
people
field
like
rio
qv
world
one
order
life
religion
son
well
road
get
brother
look
new
treasure
york
city
time
hz
home
lost


jan
night
gold
young
town
come
sam
gang
arrive
death
killed
holt
texas
friend
ranch
find
now
men
gets
city
make
burke
becomes
man
life
game
card
way
one
qv
murder
frank
just
new
john
jim
baby
fight
crime
son
get
mary
garrett
can
story
back
time
killer
run
like
takes
two
woman
brother


love
two
meets
one
young
daughter
john
woman
girl
time
help
get
world
gang
takes
war
father
now
man
life
can
story
wife


films
like
keep
ex
justice
fight
cattle
great
get
brothers
way
back
kid
kill
old
must
later
team
ned
help
billy
years
name
action
young
taken
time
fighting
new
men
gang
joe
find
can
world
rue
wild
mine
father


ranch
gang
money
get
outlaw
lash
find
take
gene
bank
young
town
cattle
sheriff
ross
brady
jamison
king
fuzzy
man
identity
law
stolen
agent
back
arrives
sets
can
billy
years
parker
stage
wade
ranches
texas
john
finds
one
bull
also
sides
indians
dick
kid
jack
old
gets
buck
takes
judge
kills
age
along
game
bill
kill
jones
uncle
father
onto
rangers
comes
outlaws


gang
one
war
new
man
bill
ranch
town
story
time
world
men
land
money
sheriff
get
gold
killed
three
army
way
many
streets
learn
two
civil
long
john
kill
old
wild
line
first
see
finds
find
now
deadly
king
team
must
gets
jim
son
father
west
range
title
night
outlaw
friend
order
also
can


time
one
new
space
story
two
future
life
world
love
man
war
planet
earth
get
must
way
go
fight
humans
make
live
takes
team
like
town
movie
also
save
power
forces
find
take
race
battle
back
best
music
help
can
first
killed
crew
three
sky
fam
stop
secret
humanity
work
ghost