Movie Plot LDA Topic Modeling Exploration

Leif Berg

2023-04-29

About

This Rmarkdown file demonstrates the use of LDA Topic Modelling in R with a set of movie plot descriptions.

Data downloaded from: https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots (https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots)

Setup

Read in Data

Read in the data from a CSV file

```
wiki_movies_plots <- read.csv("wiki_movie_plots_deduped.csv")
# Limit  number of plot descriptions (to ensure model finishes)
wiki_movies_plots <- wiki_movies_plots[1:2000,]
# wiki_movies_plots <- wiki_movies_plots %>% filter(Genre == "western")
```

View structure of data

```
str(wiki_movies_plots)
```

```
## 'data.frame':    2000 obs. of  8 variables:
##  $ Release.Year    : int  1901 1901 1901 1901 1902 1903 1903 1904 1905 1905 ...
##  $ Title           : chr  "Kansas Saloon Smashers" "Love by the Light of the Moon" "The Martyred Presidents" "Terrible Teddy, the Grizzly King" ...
##  $ Origin.Ethnicity: chr  "American" "American" "American" "American" ...
##  $ Director        : chr  "Unknown" "Unknown" "Unknown" "Unknown" ...
##  $ Cast            : chr  "" "" "" "" ...
##  $ Genre           : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ Wiki.Page       : chr  "https://en.wikipedia.org/wiki/Kansas_Saloon_Smashers" "https://en.wikipedia.org/wiki/Love_by_the_Light_of_the_Moon" "https://en.wikipedia.org/wiki/The_Martyred_Presidents" "https://en.wikipedia.org/wiki/Terrible_Teddy,_the_Grizzly_King" ...
##  $ Plot            : chr  "A bartender is working at a saloon, serving drinks to customers. After he fills a stereotypically Irish man's b"| __truncated__ "The moon, painted with a smiling face hangs over a park at night. A young couple walking past a fence learn on "| __truncated__ "The film, just over a minute long, is composed of two shots. In the first, a girl sits at the base of an altar "| __truncated__ "Lasting just 61 seconds and consisting of two shots, the first shot is set in a wood during winter. The actor r"| __truncated__ ...
```

View first few rows of data

```
head(wiki_movies_plots)
```

```
##   Release.Year                           Title Origin.Ethnicity
## 1         1901           Kansas Saloon Smashers         American
## 2         1901    Love by the Light of the Moon         American
## 3         1901          The Martyred Presidents         American
## 4         1901 Terrible Teddy, the Grizzly King         American
## 5         1902           Jack and the Beanstalk         American
## 6         1903              Alice in Wonderland         American
##                              Director      Cast   Genre
## 1                             Unknown           unknown
## 2                             Unknown           unknown
## 3                             Unknown           unknown
## 4                             Unknown           unknown
## 5 George S. Fleming, Edwin S. Porter           unknown
## 6                      Cecil Hepworth May Clark unknown
##                                                        Wiki.Page
## 1             https://en.wikipedia.org/wiki/Kansas_Saloon_Smashers
## 2      https://en.wikipedia.org/wiki/Love_by_the_Light_of_the_Moon
## 3            https://en.wikipedia.org/wiki/The_Martyred_Presidents
## 4   https://en.wikipedia.org/wiki/Terrible_Teddy,_the_Grizzly_King
## 5 https://en.wikipedia.org/wiki/Jack_and_the_Beanstalk_(1902_film)
## 6    https://en.wikipedia.org/wiki/Alice_in_Wonderland_(1903_film)
##
Plot
## 1
A bartender is working at a saloon, serving drinks to customers. After he fills a stereotypicall
y Irish man's bucket with beer, Carrie Nation and her followers burst inside. They assault the I
rish man, pulling his hat over his eyes and then dumping the beer over his head. The group then
begin wrecking the bar, smashing the fixtures, mirrors, and breaking the cash register. The bart
ender then sprays seltzer water in Nation's face before a group of policemen appear and order ev
erybody to leave.[1]
## 2
The moon, painted with a smiling face hangs over a park at night. A young couple walking past a
fence learn on a railing and look up. The moon smiles. They embrace, and the moon's smile gets b
igger. They then sit down on a bench by a tree. The moon's view is blocked, causing him to frow
n. In the last scene, the man fans the woman with his hat because the moon has left the sky and
is perched over her shoulder to see everything better.
## 3
The film, just over a minute long, is composed of two shots. In the first, a girl sits at the ba
se of an altar or tomb, her face hidden from the camera. At the center of the altar, a viewing p
ortal displays the portraits of three U.S. Presidents—Abraham Lincoln, James A. Garfield, and Wi
lliam McKinley—each victims of assassination.\nIn the second shot, which runs just over eight se
conds long, an assassin kneels feet of Lady Justice.
## 4
Lasting just 61 seconds and consisting of two shots, the first shot is set in a wood during wint
er. The actor representing then vice-president Theodore Roosevelt enthusiastically hurries down
a hillside towards a tree in the foreground. He falls once, but rights himself and cocks his rif
le. Two other men, bearing signs reading "His Photographer" and "His Press Agent" respectively,
follow him into the shot; the photographer sets up his camera. "Teddy" aims his rifle upward at
the tree and fells what appears to be a common house cat, which he then proceeds to stab. "Tedd
y" holds his prize aloft, and the press agent takes notes. The second shot is taken in a slightl
y different part of the wood, on a path. "Teddy" rides the path on his horse towards the camera
and out to the left of the shot, followed closely by the press agent and photographer, still dut
```

```
ifully holding their signs.
## 5
The earliest known adaptation of the classic fairytale, this films shows Jack trading his cow fo
r the beans, his mother forcing him to drop them in the front yard, and beig forced upstairs. As
he sleeps, Jack is visited by a fairy who shows him glimpses of what will await him when he asce
nds the bean stalk. In this version, Jack is the son of a deposed king. When Jack wakes up, he f
inds the beanstalk has grown and he climbs to the top where he enters the giant's home. The gian
t finds Jack, who narrowly escapes. The giant chases Jack down the bean stalk, but Jack is able
to cut it down before the giant can get to safety. He falls and is killed as Jack celebrates. Th
e fairy then reveals that Jack may return home as a prince.
## 6 Alice follows a large white rabbit down a "Rabbit-hole". She finds a tiny door. When she fi
nds a bottle labeled "Drink me", she does, and shrinks, but not enough to pass through the door.
She then eats something labeled "Eat me" and grows larger. She finds a fan when enables her to s
hrink enough to get into the "Garden" and try to get a "Dog" to play with her. She enters the "W
hite Rabbit's tiny House," but suddenly resumes her normal size. In order to get out, she has to
use the "magic fan."\nShe enters a kitchen, in which there is a cook and a woman holding a baby.
She persuades the woman to give her the child and takes the infant outside after the cook starts
throwing things around. The baby then turns into a pig and squirms out of her grip. "The Duches
s's Cheshire Cat" appears and disappears a couple of times to Alice and directs her to the Mad H
atter's "Mad Tea-Party." After a while, she leaves.\nThe Queen invites Alice to join the "ROYAL
PROCESSION": a parade of marching playing cards and others headed by the White Rabbit. When Alic
e "unintentionally offends the Queen", the latter summons the "Executioner". Alice "boxes the ea
rs", then flees when all the playing cards come for her. Then she wakes up and realizes it was a
ll a dream.
```

Read in lists of common first names for males and females (ensure they are lowercase)

```
male_names <- tolower(read.delim("male.txt",header = FALSE)$V1)
# Select first 3k female names (full list is too much for regex function)
female_names <- tolower(read.delim("female.txt",header = FALSE)$V1)[1:3000]
```

Clean the data - remove common words/names

```
#Ensure plot descriptions are in lowercase
wiki_movies_plots$Plot <- tolower(wiki_movies_plots$Plot)
#Collect list of common words from tidytext package
stop_words <- tidytext::stop_words
#remove common words and names from plot descriptions
wiki_movies_plots$Plot <- removeWords(wiki_movies_plots$Plot,c(stop_words$word))
wiki_movies_plots$Plot <- removeWords(wiki_movies_plots$Plot,c(male_names))
wiki_movies_plots$Plot <- removeWords(wiki_movies_plots$Plot,c(female_names))
```

LDA Model

Create a Corpus object from the vector of movie plots

```
corpus <- Corpus(VectorSource(wiki_movies_plots$Plot))
```

Create a Document Term Matrix

```
tdm <- DocumentTermMatrix(corpus)
```

Create an LDA model with 12 topics and seed (to ensure consistent randomness)

```
ap_lda <- topicmodels::LDA(tdm, k = 12, control = list(seed = 1234))
ap_lda
```

```
## A LDA_VEM topic model with 12 topics.
```

Create Beta Table

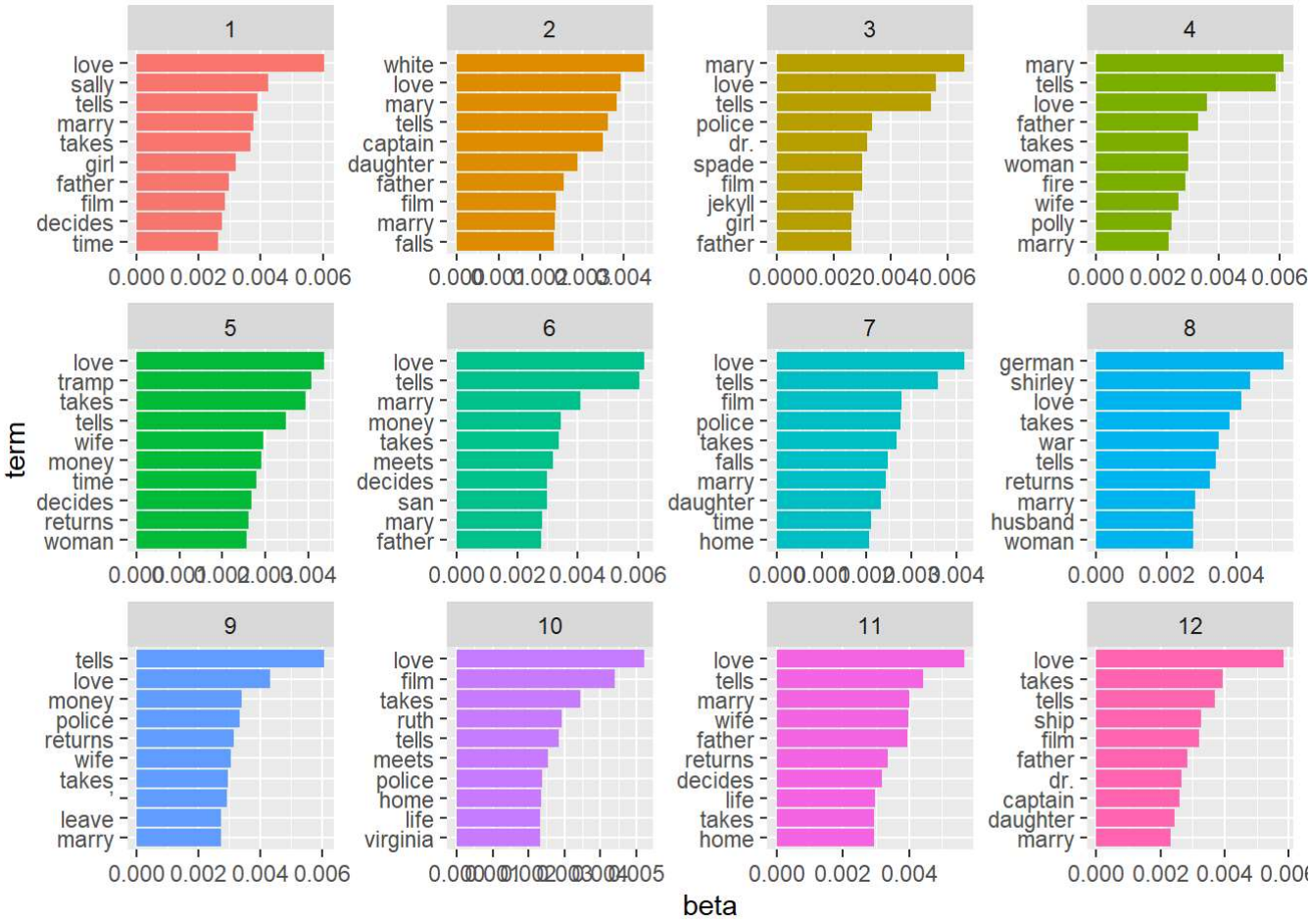
```
ap_topics <- tidy(ap_lda, matrix = "beta")
ap_topics
```

```
## # A tibble: 463,416 × 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1     1 assault 5.44e-  5
##  2     2 assault 1.09e-240
##  3     3 assault 4.83e- 32
##  4     4 assault 7.28e-  5
##  5     5 assault 6.08e-  5
##  6     6 assault 8.70e-243
##  7     7 assault 5.77e-  5
##  8     8 assault 2.69e-242
##  9     9 assault 5.80e-  5
## 10    10 assault 2.46e-  4
## # ℹ 463,406 more rows
```

Visualize Results

```
#Prepare the data for visualization
ap_top_terms <- ap_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(topic, -beta)

#Visualize the topics with ggplot
ap_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered()
```


1
love
sally
tells
marry
takes
girl
father
film
decides
time
0.000 0.002 0.004 0.006
2
white
love
mary
tells
captain
daughter
father
film
marry
falls
0.000 0.001 0.002 0.003 0.004
3
mary
love
tells
police
dr.
spade
film
jekyll
girl
father
0.000 0.002 0.004 0.006
4
mary
tells
love
father
takes
woman
fire
wife
polly
marry
0.000 0.002 0.004 0.006
5
love
tramp
takes
tells
wife
money
time
decides
returns
woman
0.000 0.001 0.002 0.003 0.004
6
love
tells
marry
money
takes
meets
decides
san
mary
father
0.000 0.002 0.004 0.006
7
love
tells
film
police
takes
falls
marry
daughter
time
home
0.000 0.001 0.002 0.003 0.004
8
german
shirley
love
takes
war
tells
returns
marry
husband
woman
0.000 0.002 0.004
9
tells
love
money
police
returns
wife
takes
leave
marry
0.000 0.002 0.004 0.006
10
love
film
takes
ruth
tells
meets
police
home
life
virginia
0.000 0.001 0.002 0.003 0.004 0.005
11
love
tells
marry
wife
father
returns
decides
life
takes
home
0.000 0.002 0.004
12
love
takes
tells
ship
film
father
dr.
captain
daughter
marry
0.000 0.002 0.004 0.006
term
beta