Topic Modeling Report

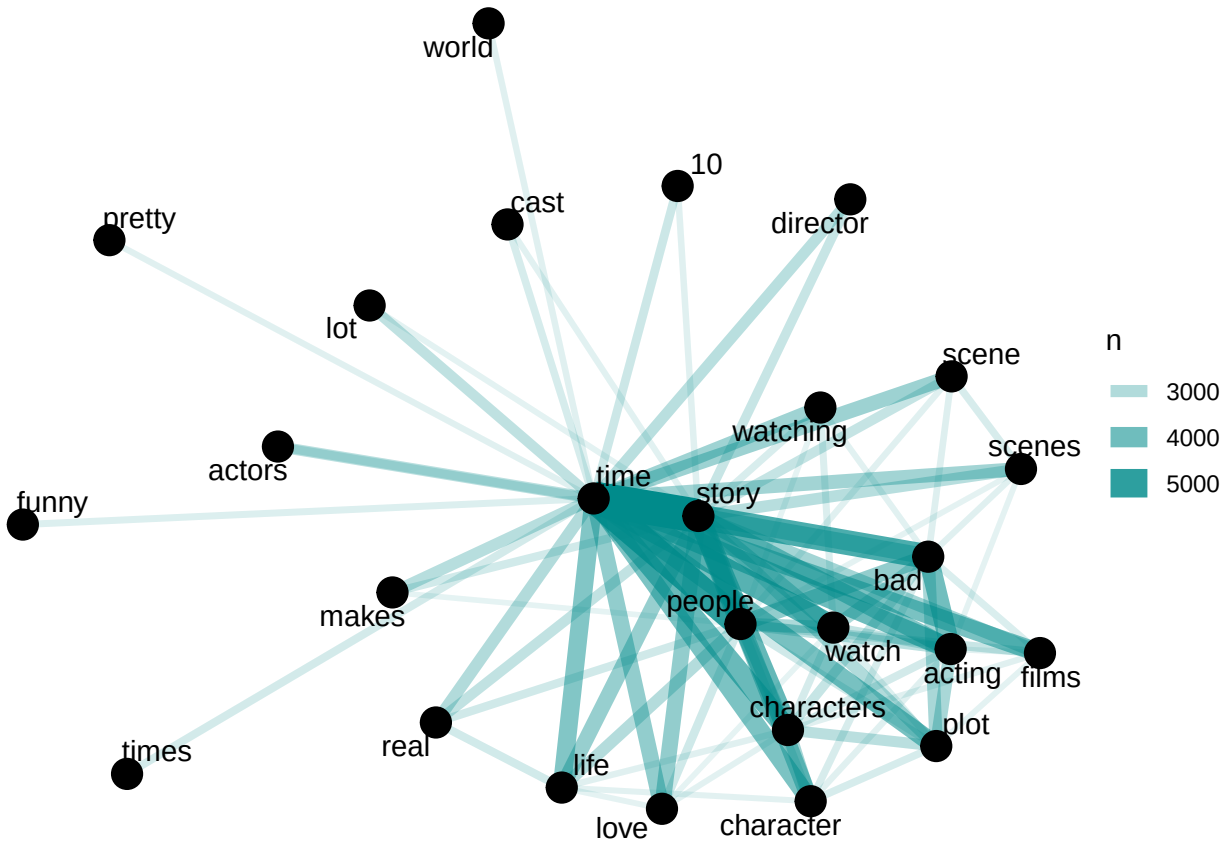
Group 7

2022-11-12

We use IMDB reviews as our data and do the following anylsis. First, we set up three stop words, movie, movies and film, since these three words appear most frequently and do not make much sense in movie reviews. Then, we use $unnest_tokens()$ function to both break the text into individual tokens and transform it to a tidy data structure.

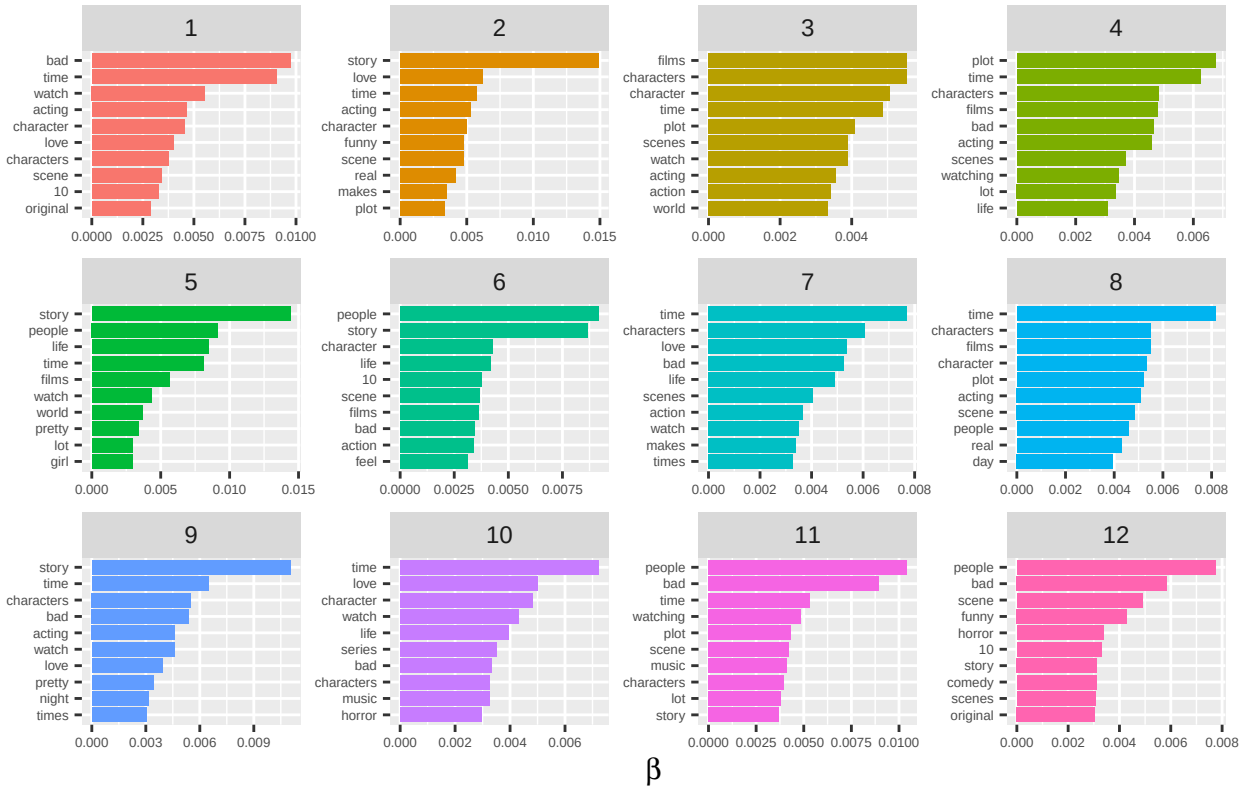
Afterwards, we look at the terms' inverse document frequency to measure if a word is important in our analysis. As a result, considering both the number of occurrences and term frequency(n/total), we have top 5 terms, 'trivialboring', 'mad', 'stop.oz', 'bad', and 'bought'. Three of them are negative, so that we may conclude that there are more negative review than the positive. In order to prove it, we further draw some plots.

Here, we draw a plot showing how words correlate to each other. We filter the words by saying that the total number of occurrence is larger than 2200 and get 26 words, which is centered around 'time', 'story', 'bad' and 'people'.



Another plot shows how often these words show up in the review. The larger the word, the more often it appears. Obviously, 'time' and 'story' are the two most frequent occurrences, followed by 'bad' and 'people', and then 'love', 'films', etc..

Finally, we use LDA model to find 12 topics that are extracted from the review. The most common words in topic 1 include "bad", "watch","10", and "original", which suggests it may represent literary film. Those most common in topic 2 include "makes", "acting", "funny", and "real", suggesting that this topic represents comedy or action movie. When analyze these topic, we tend to ignore the words that appears in every topic including "story", "character", and "scene".


Top 10 terms in each LDA topic
1
bad
time
watch
acting
character
love
characters
scene
10
original
0.0000 0.0025 0.0050 0.0075 0.0100
2
story
love
time
acting
character
funny
scene
real
makes
plot
0.000 0.005 0.010 0.015
3
films
characters
character
time
plot
scenes
watch
acting
action
world
0.000 0.002 0.004
4
plot
time
characters
films
bad
acting
scenes
watching
lot
life
0.000 0.002 0.004 0.006
5
story
people
life
time
films
watch
world
pretty
lot
girl
0.000 0.005 0.010 0.015
6
people
story
character
life
10
scene
films
bad
action
feel
0.0000 0.0025 0.0050 0.0075
7
time
characters
love
bad
life
scenes
action
watch
makes
times
0.000 0.002 0.004 0.006 0.008
8
time
characters
films
character
plot
acting
scene
people
real
day
0.000 0.002 0.004 0.006 0.008
9
story
time
characters
bad
acting
watch
love
pretty
night
times
0.000 0.003 0.006 0.009
10
time
love
character
watch
life
series
bad
characters
music
horror
0.000 0.002 0.004 0.006
11
people
bad
time
watching
plot
scene
music
characters
lot
story
0.0000 0.0025 0.0050 0.0075 0.0100
12
people
bad
scene
funny
horror
10
story
comedy
scenes
original
0.000 0.002 0.004 0.006 0.008
β