Topic modeling of documentary reviews on IMDB

This is a brief overview of methods describing the preliminary work to inform the design of the Impact Assessment scale (IAS).

Data querying and selection:
Given the size of the data available, we first queried the top 100 rated documentaries on IMDB. This resulted in 57,109 reviews from a wide range of films and series categorized under the “Documentary” genre, and still with substantial noise-to-signal ratio. To keep our analyses in line with our a priori goal of focusing on themes of social impact (and/or impact), we used a subset of these documentaries to conduct our analyses. This resulted in a sub-sample of 2,704 reviews from 1999-2022 and from the following titles:

[1] "all the beauty and the bloodshed"
[2] "an inconvenient truth"
[3] "how to survive a plague"
[4] "food, inc."
[5] "harlan county u.s.a."
[6] "born into brothels: calcutta's red light kids"
[7] "hearts and minds"
[8] "lake of fire"
[9] "icarus"
[10] "bowling for columbine"
[11] "blackfish"
[12] "fed up"
[13] "the times of harvey milk"
[14] "the cove"
[15] "man on wire"
[16] "fahrenheit 9/11"

Data modeling and analysis:
We first processed and prepared the data for structural topic modeling (stm). After removing the infrequent terms (lower threshold <.05) and non-text or empty documents, the corpus for analysis included 2,704 documents, 4,259 terms (dictionary) and 237,277 tokens. The number (k) of topics in the corpus was informed through the diagnostic measures of held-out likelihood, semantic coherence, and residual dispersion (see Figure below). The first bend in the line indicates suggested optimal k number of topics, which in our case is k=5 topics. Accordingly, we ran a stm with k=5 on the corpus described above. We further inspected the relationships between the model identified features representative of the topics using co-occurrence network graphs.[1] For all analyses and plots, R libraries including stm, lda, quanteda.textplots were used.

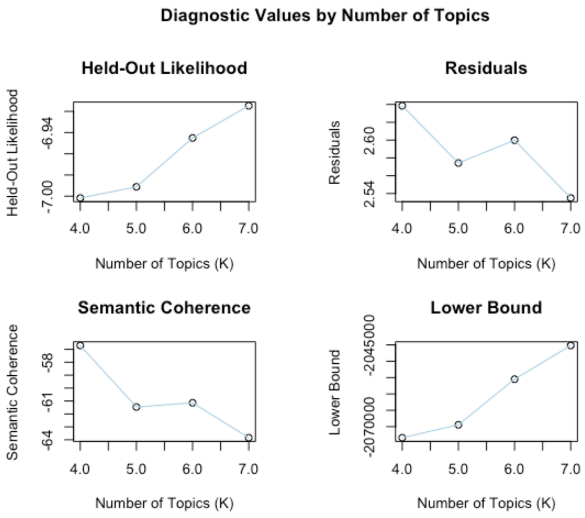


Figure 1. Diagnostic plots indicating optimal k number of topics for modeling.

Results:

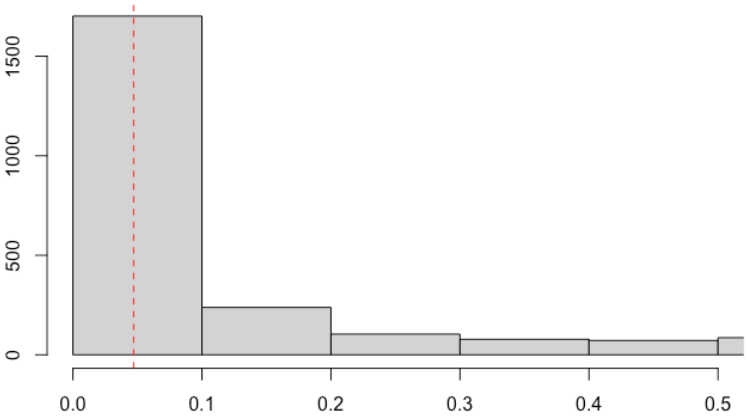
We provide the stm results summarized by top words for each topic (See Table 1) calculated based on 4 metrics: 1) highest probability (words within each topic with the highest probability)[2], 2) FREX (words that are both frequent and exclusive, identifying words that distinguish topics)[3], 3) lift (ratio of topic-word distribution by the empirical word count probability distribution)[4], and 4) score (log probability of seeing word v conditional on topic k)[5]. FREX is calculated by taking the harmonic mean of rank by probability within the topic (frequency) and rank by distribution of topic given word (exclusivity).

Table 1. Top words for each of the 5 topics identified from the stm.

Topic number (k)	Top words computed based on 4 metrics
Topic 1	Highest Probility: moor, gun, film, columbin, documentari, movi, one, make, america, michael FREX: gun, heston, violenc, columbin, nra, charlton, bowl, canada, manson, fear Lift: barri, handgun, mose, prochoic, rocker, shooter, sixyearold, manson, nichol, ownership Score: gun, moor, columbin, heston, bowl, barri, nra, violenc, charlton, manson
Topic 2	Highest Probility: gore, global, warm, film, movi, will, chang, present, one, can FREX: warm, global, gore, inconveni, earth, climat, temperatur, data, planet, scientif Lift: antarct, antarctica, scifi, cyclic, futurama, guggenheim, pole, glacier, warmer, kyoto Score: gore, warm, global, scifi, inconveni, temperatur, scientif, carbon, earth, climat
Topic 3	Highest Probility: moor, film, bush, movi, see, peopl, war, michael, one, make FREX: iraq, bush, war, fahrenheit, saudi, soldier, administr, georg, bin, laden Lift: gotta, havemor, goat, photoop, wolfowitz, comb, wmd, britney, cheney, arabian Score: moor, bush, iraq, war, saudi, fahrenheit, gotta, bin, soldier, iraqi
Topic 4	Highest Probility: film, documentari, one, world, anim, watch, see, dolphin, stori, like FREXility: petit, wire, trainer, whale, philipp, seaworld, dolphin, orca, blackfish, cove Lift: marsh, zana, abund, anni, brancheau, caper, cathedr, composit, cowperthwait, dame Score: whale, dolphin, petit, trainer, wire, seaworld, orca, philipp, abund, blackfish
Topic 5	Highest Probility: food, industri, peopl, documentari, film, movi, eat, like, product, can FREX: food, eat, inc, industri, corn, farmer, product, chicken, sugar, meat Lift: ecoli, pollan, syrup, contamin, corn, couric, genet, hormon, inc, kenner Score: food, uneth, eat, corn, inc, farmer, meat, chicken, kenner, industri

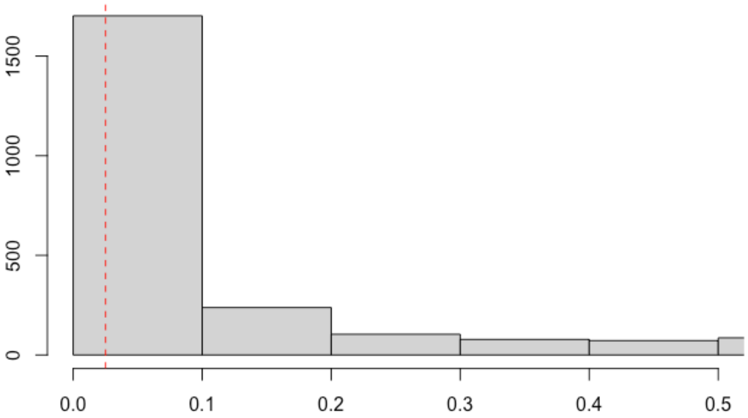
Distribution of MAP Estimates of Document-Topic Proportions

Topic 1: moor, gun, film, columbin, documentari, movi, one



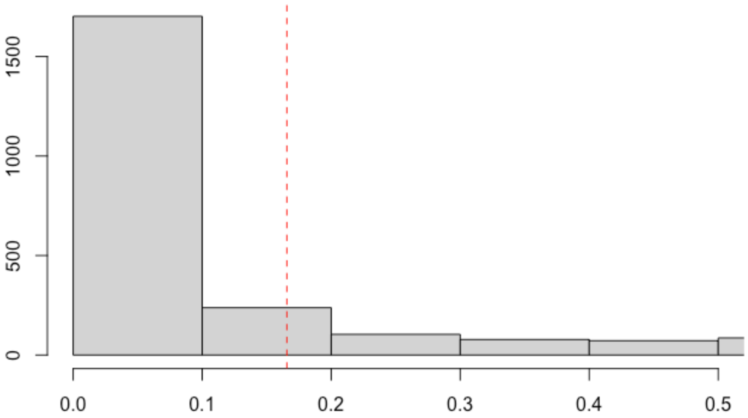
Distribution of MAP Estimates of Document-Topic Proportions

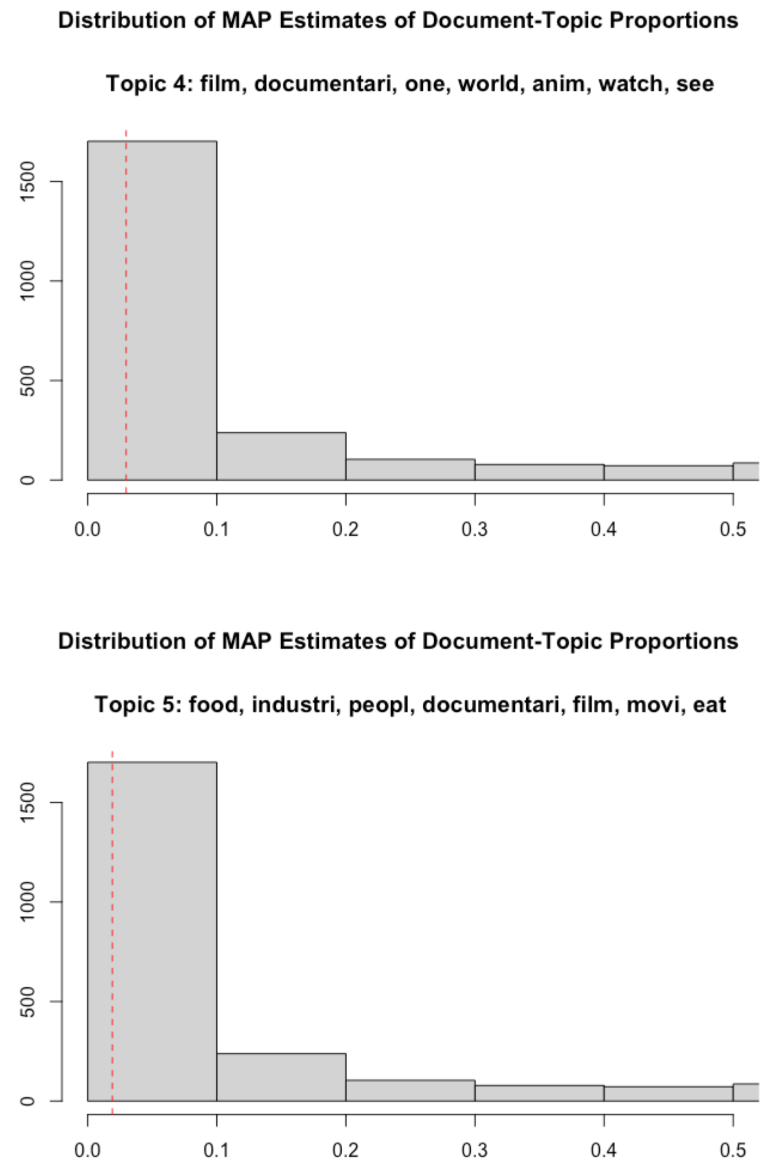
Topic 2: gore, global, warm, film, movi, will, chang



Distribution of MAP Estimates of Document-Topic Proportions

Topic 3: moor, film, bush, movi, see, peopl, war





Panel Figure 2. Histograms of the expected distribution of topic proportions across all the documents (N = 2,704) for all 5 topics. X-axis shows the maximum a posteriori probability (MAP) estimates of the document-topic loadings across all documents. The dashed red line indicates the median.

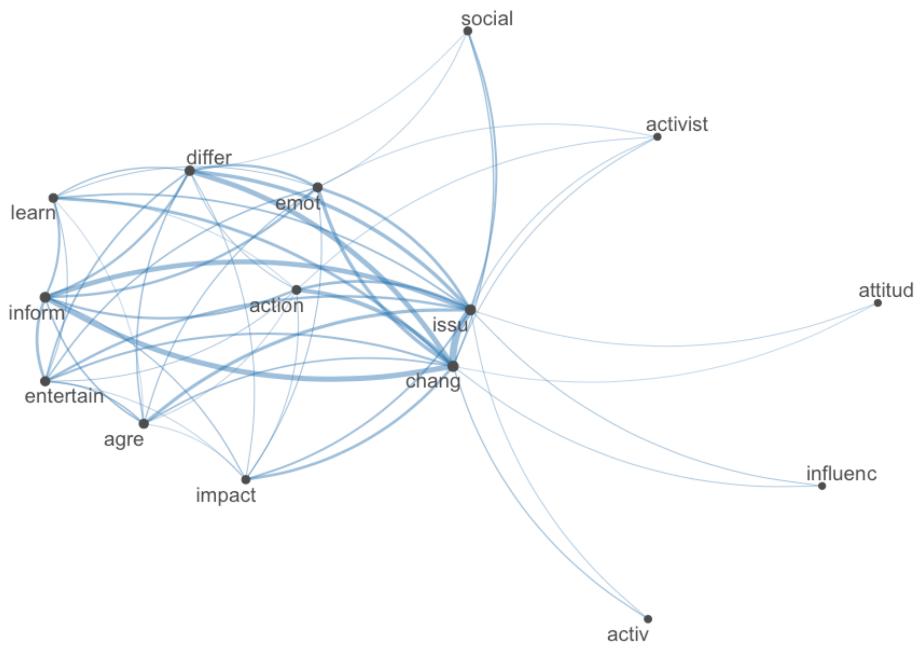


Figure 3. Network of feature relationships based on co-occurrence probability. Blue edges show co-occurrences of the features identified by the model. Distance between the features is directly proportional to their co-occurrence likelihood.[1] The lower frequency threshold for inclusion in the plot is set at 0.8 (i.e., minimum 8% occurrence in the corpus).

References

1. Benoit K, Watanabe K, Wang H, Obeng A, Müller S, Matsuo A, Fellows I. *quanteda.textplots: Plots for the Quantitative Analysis of Textual Data; 0.94.3*. 2023.
2. Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, Albertson B, Rand DG. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*. 2014;58:1064-1082. doi: https://doi.org/10.1111/ajps.12103
3. Bischof JM, Airoldi EM. Summarizing topical content with word frequency and exclusivity. In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Edinburgh, Scotland: Omnipress; 2012:9–16.
4. Taddy M. On Estimation and Selection for Topic Models. In: Neil DL, Mark G, eds. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research: PMLR; 2012:1184--1193.
5. Chang J. *lda: Collapsed Gibbs Sampling Methods for Topic Models; 1.4.2.* 2015.