Name: Quang-Anh Ngo Tran (Alex Tran)
Class: Data Science and Text Analysis
Date: 08/03/2023

Homework 3

Question 1.

The corpus of documents I will be conducting my analysis on will be a sample of the Corpus of Historical American English (COHA) - the largest structured corpus of historical English, with a balanced set of genres by decade. My sample of the whole corpus has 1144 documents with a total of about 48 million words (4,833,585 words). I got this sample by downloading from their website (https://www.english-corpora.org/coha/).

For preprocessing in Question 2:

No use of n-grams: Previous research on stereotypes about men and women tend to be in a single word, with very little phrases (e.g., dominant, analytical, muscular, dictatorial, affectionate, imaginative, gorgeous, subordinates; Cejka & Eagly, 1999; Diekman & Eagly, 2000; Williams & Best, 1982; Deaux & Lewist, 1984). Thus, if I am examining stereotypes about men and women in these texts, I may not miss much by including only 1-gram.

Remove stopwords and meaningless words (self-created): Function words (e.g., a, an, the, etc.) are not meaningful for my question of examining stereotypes. Other than that, base on my frequency plots done in Homework 1, I am also eliminating the following terms because they do not make sense but they appear a lot for some reasons: "s", "p", "n't", "--", "d", "z", "ll", "m", "ve", "nbsp." Finally, I am keeping “he” and “she” (default as stopwords) instead of eliminating them because they are pertinent to my examination of gender stereotypes.

Remove punctuations and numbers: Punctuations and special characters are also not very interpretable for my research question. Sometimes, punctuation comes with numbers (e.g., 12,345) but since number is also often not associated with gender stereotypes, removing punctuation and numbers would not be significantly removing meanings.

Lowercase all words: I am not examining the gender of proper nouns. Thus, it is not a significant contributor to the examination of gender stereotypes. Thus, lowercasing all words should not affect the research question results.

Stemming: This helps reduce words to their linguistic stems, which significantly reduce our vocabulary without losing too much meaning (besides losing their prefixes and suffixes)

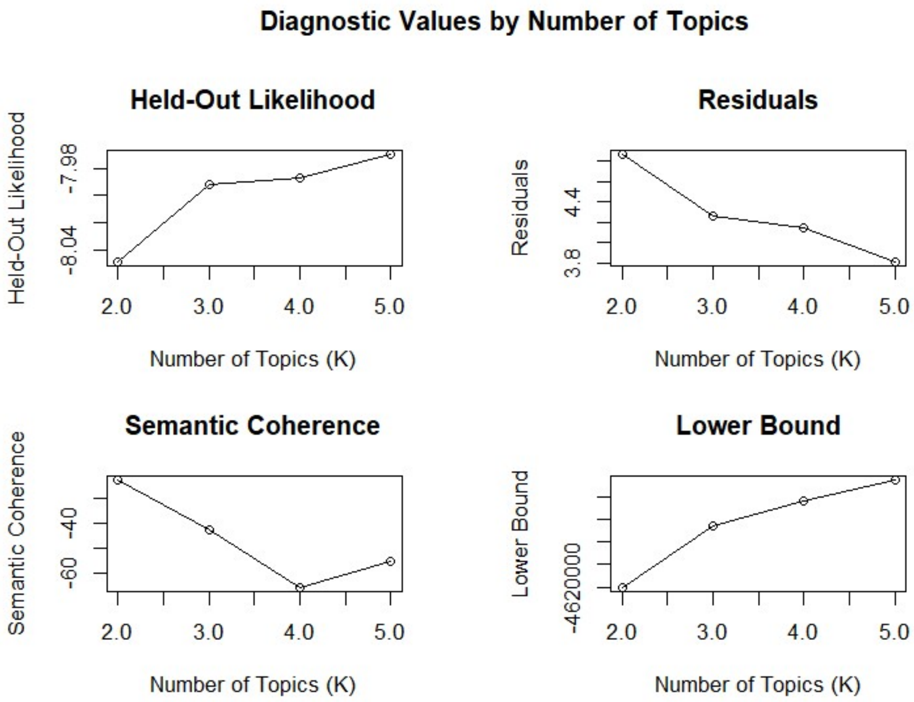
Remove infrequently used terms at the 5% threshold: Help significantly reduce vocabulary size. Furthermore, gender stereotypes have been shown to be somewhat prevalent in everyday language (Bailey et al., 2022; Charlesworth et al., 2021). Thus, removing infrequently used words should not affect the words that are associated with gender stereotypes.

All these preprocessing steps aim to boost the “signal in the noise” because they keep information about gender stereotypes intact while eliminating redundant information.

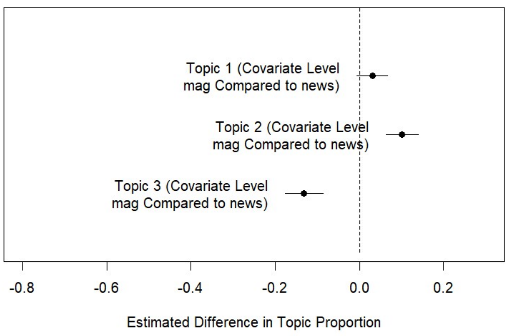
For preprocessing in Question 3: since I am using word embeddings, I am only lowercasing my texts. As I said before, I am not examining the gender of proper nouns. Thus, it is not a significant contributor to the examination of gender stereotypes. This preprocessing step should not alter my results, even though I am willing to redo the analysis without lowercasing.

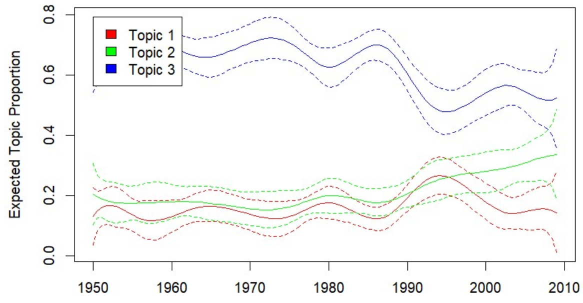
Question 2.

Choosing the number of topics: Based on these tradeoffs between held-out likelihood + residuals and semantic coherence, 3 topics seem to be most efficient. Thus, I will try running the STM for 3 topics. I included year and genre as the covariates to see if the topics have different prevalence across the years and genres.



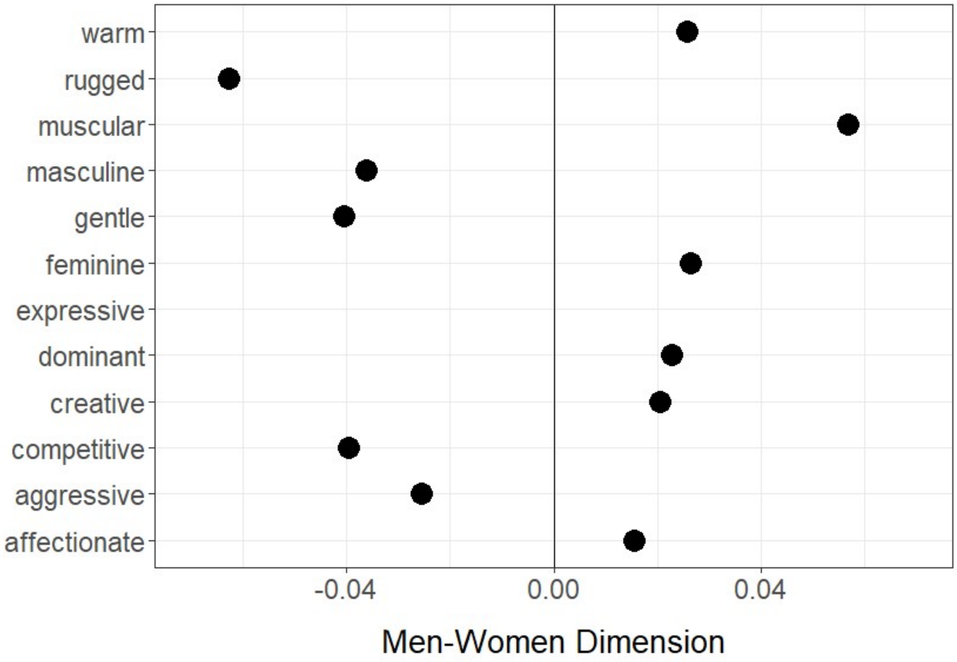
For 3 topics: the models converged. Topic 1 I named science (FREX: sortal, protein, amino, telescop, particl, miner, vitamin). Topic 2 I named Medieval English (FREX: fabian, dubin, dial, fanni, kitti, bebb). Topic 3 I named politics (FREX: judd, arti, vote, democrat, tax, kennedi, carter). Please see the following results in the R Markdown/html file. Topic 3 is not doing so well in terms of exclusivity. However, considering how prevalent politics are in fiction, non-fiction, news, and magazines, this is not surprising. Topic 3 is indeed the most prevalent topic. All genres significantly mentioned topic 1 science and topic 2 Medieval English. However, only news and magazines significantly mentioned topic 3 politics (which makes sense). There are also some time differences in the prevalence of each topic. Magazines uses Medieval English significantly more than news sources, but news sources mention politics significantly more than magazines (which makes sense). My preprocessing decisions mentioned in 1 probably boosted the "signal in the noise" and draw out 3 pretty clean patterns of topics. I removed the infrequently used terms but topic2 Medieval English still appears, which indicates a lot of texts must have used such language.





Question 3.

I preprocessed by lowercasing the words. For the hyperparameters, I left rank as 300 (instead of 200), maximum number of occurrences as 15, number of iterations before convergence tolerance (0.001) as 100. My learning rate is 0.15. This is per the recommendations in Rodriguez & Spirling (2022). I created a gender dimension spanning from “men”-related words (e.g., he, male, boy) to “women”-related words (e.g., she, female, girl). Then, I look at the following terms to map onto the created dimensions: masculinity, femininity (as checks for my model), competitive, aggressive, dominant, muscular, rugged, affectionate, gentle, warm, expressive, creative, gorgeous (these stereotypes include personality, cognitive, and physical stereotypes).



Most of these findings are consistent with previous research on gender stereotypes (and common sense). Masculine and Feminine fall on the right sides, which provides some credibility for the model. Men are more associated with rugged, aggressive, competitive. Women are more associated with warm, creative, and affectionate. There are a few cases that are surprising though: muscular closer to women, gentle closer to men, dominant closer to women. I guess the latter 2 is possible – in an androcentric world, men may get all the positive stereotypes - but muscular to women is hard to explain.