



Best selling Books: Data Analysis

Weekly Update

<https://github.com/sruthipsuresh/best-sellers-analysis>



Weekly Summary

- Built all datasets from first draft and created better data pipeline to get ALL books from Google Books API.
- Figured out how to get more results from GBooks

Data Collection and Cleaning

Book List and Terms

gBooksfinal.ipynb > gbooksprocessing.ipynb



API Queries

title	author	isbn (h)
How to be an Anti-Racist	Ibram X. Kendi	9.78053E+12
White Fragility	Robin Di Angelo	9780807047408
Between the World and Me	Ta-Nahisi Coates	9780812993547
Stamped: Rascism, Anti-Rascism and You	. Reynolds and Ibram X. Kendi	9781568584638
Say Her Name (Poems to Empower)	Zetta Elliott	9781368045247
And She Was by Jessica Verdi	Jessica Verdi	9781338150537
Drag Teen	Jeffrey Self	9780545829939








```
term
transgender
transgender ideology
transgender rights
microaggressions
```

Terms either queried as:

- **In title:** Returns results where the text following this keyword is found in the title
- **subject:** Returns results where the text following this keyword is listed in the category list of the volume.



Data Collection

 covid_terms_category_clean.csv	Created all datasets	54	2 minutes ago
 covid_terms_title_clean.csv	Created all datasets	268	2 minutes ago
 pc_terms_category_clean.csv	Created all datasets	78	2 minutes ago
 pc_terms_title_clean.csv	Created all datasets	573	2 minutes ago
 specific_books_clean.csv	Created all datasets	7	2 minutes ago
 transgender_terms_category_clean.csv	Created all datasets	50	2 minutes ago
 transgender_terms_title_clean.csv	Created all datasets	125	2 minutes ago

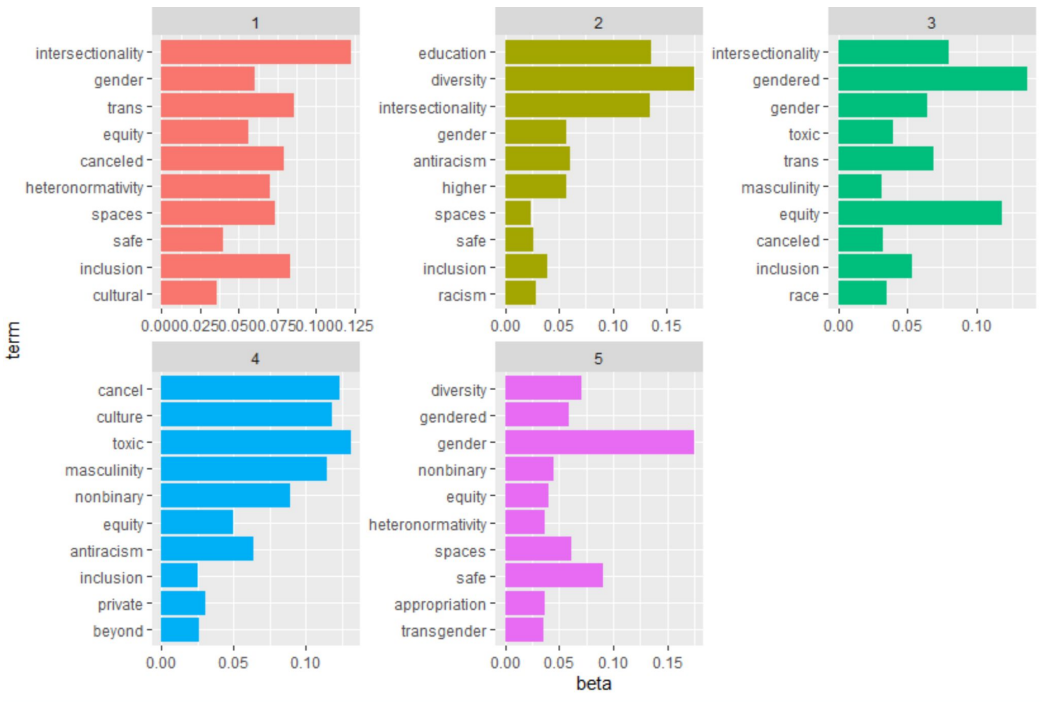
Modeling using R

ldamodeling.R

stmmodeling.R



Tested out LDA with PC terms (intitle)



TITLES

- Makes the case for structural topic modeling stronger.
- Matches “manual” assignment



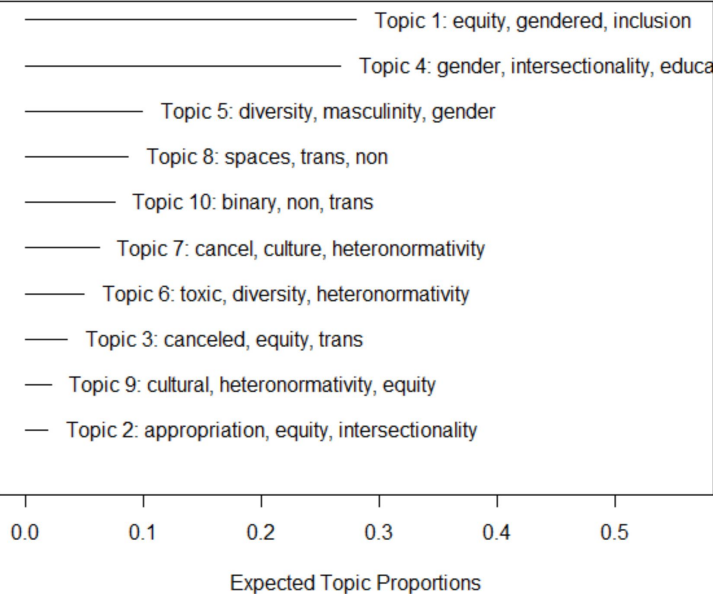
Tested out LDA with PC terms (intitle)

DESCRIPTION

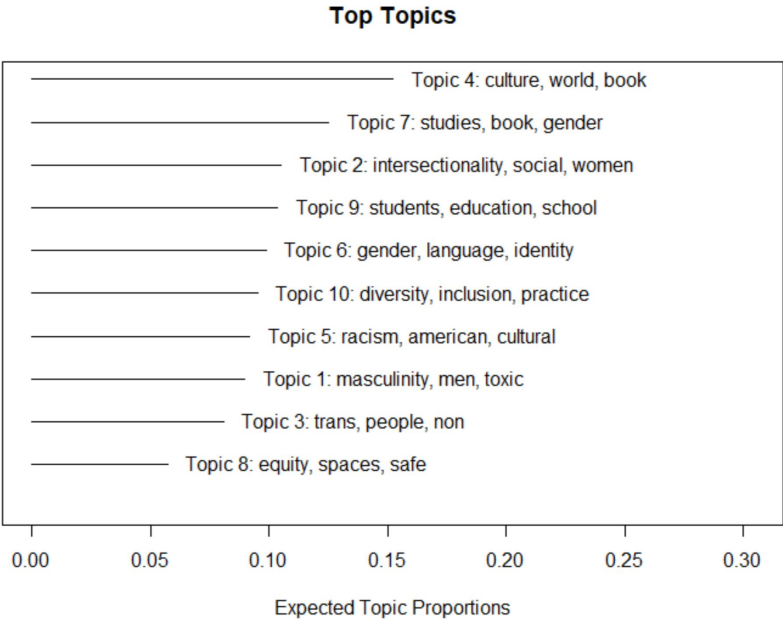




No metadata (STM)



titles



description



Metadata Ideas

- Author?
- Date published?
- “STM is very similar to LDA, but it employs meta data about documents (such as the name of the author or the date in which the document was produced) to improve the assignment of words to latent topics in a corpus”

Conclusion



Plans for this week:

- Build dataset as discussed.
- Run all other analysis notebooks.
- Add metadata to columns and also plot change in the prevalence of topic over time.
- Model “prevalence” of topics on “author” and “year”.