The Effect of Preprocessing on Bayesian Stochastic Blockmodels for Topic Identification

Brandon SEPULVADO

13 July 2020

Methodology and Quantitative Social Sciences
NORC at the University of Chicago

Overview

1. Approaches to topic identification
2. Bayesian Stochastic Blockmodels
3. Data and Analytical Approach
4. Results

Topic Identification

- Text analysis/NLP
  - Topic modeling (e.g., LDA, STM)
  - Cluster on embeddings
- Network-based approaches
  - One-mode
    - Community detection on semantic networks
  - Two-mode
    - Bipartite community detection
    - Bayesian Stochastic Blockmodels

Bayesian Stochastic Blockmodels

- Why Bayesian Stochastic Blockmodels?
  - Superior performance to LDA
  - Avoid artefacts of single-mode projections
  - Cluster simultaneously documents and words
  - Hierarchical topic structure
- How does preprocessing impact network topology?
  - # nodes
  - # edges
  - Density

How do preprocessing-induced topological changes impact the topics recovered?

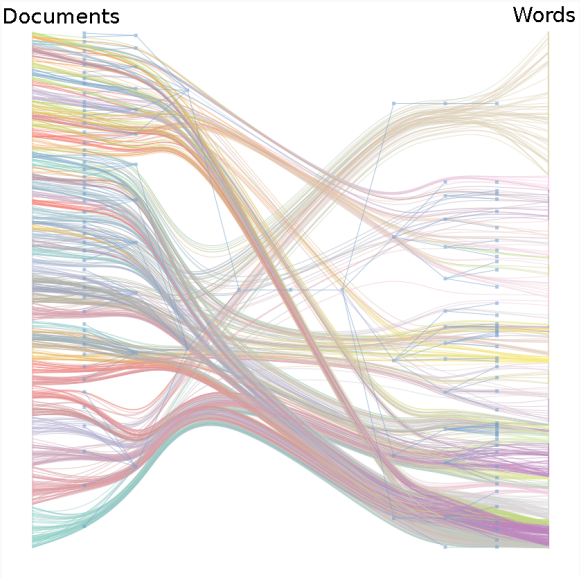
Data and Analytical Approach

- Data
  - Synthetic biology ethics
    - Keyword-based query from Web of Science
    - 15,152 synthetic biology records
    - 574 ethics-related records within this larger set (abstracts used)
  - Education thought leaders
    - List of 1,022 education-related associations, institutions, foundations, individuals curated by education experts
    - 930 Twitter accounts found
    - 1,514,352 tweets from them
    - Randomly sampled 10,000 tweets
- Analytical approach
  1. *Minimum* preprocessing: remove punctuation and case standardization (lowercase)
  2. *Medium* preprocessing: minimum + remove stop words
  3. *Maximum* preprocessing: medium + lemmatize

Results: Synthetic Biology Ethics

Processing: **minimum**
words: 11,970
documents: 574
topic levels: 3

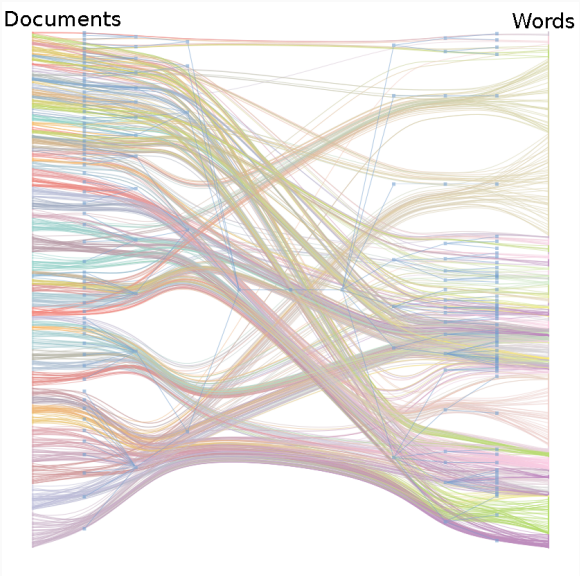
- Level 0: 48 topics
- Level 1: 15 topics
- Level 2: 5 topics



Results: Synthetic Biology Ethics

Processing: **medium**
words: 11,716
documents: 574
topic levels: 3

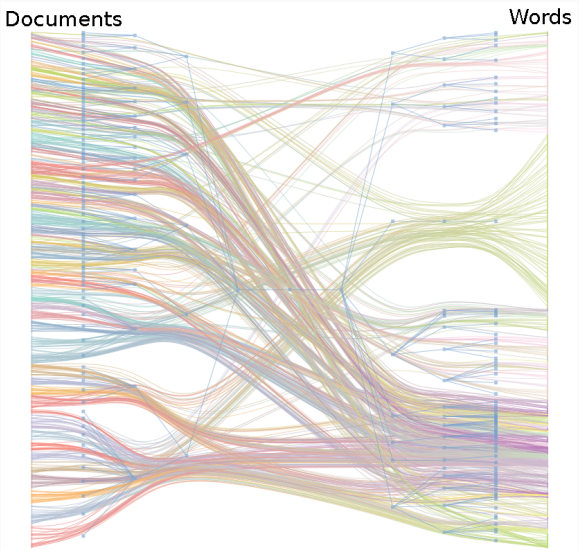
- Level 0: 64 topics
- Level 1: 21 topics
- Level 2: 7 topics



Results: Synthetic Biology Ethics

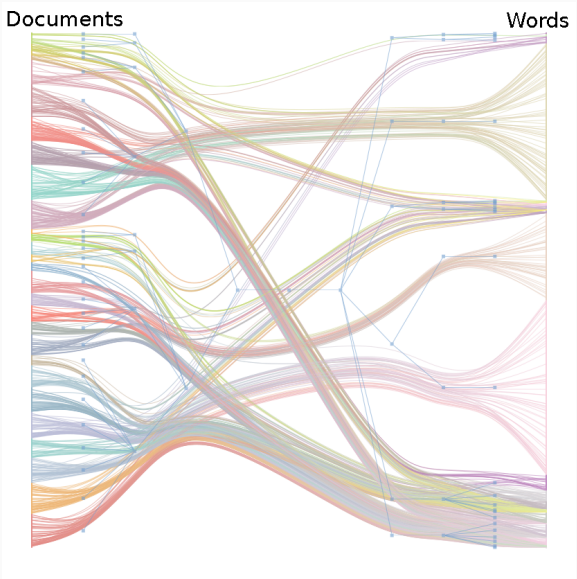
Processing: **maximum**
words: 9,698
documents: 574
topic levels: 3

- Level 0: 77 topics
- Level 1: 23 topics
- Level 2: 8 topics



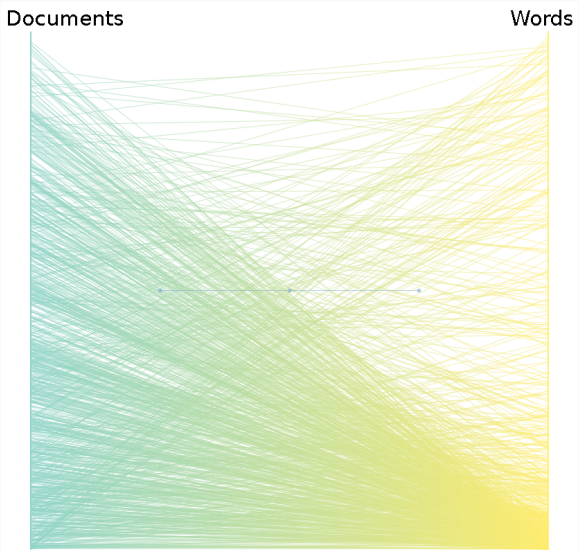
Results: Education Thought Leaders

Processing: **minimum**
words: 43,742
documents: 10,000
topic levels: 3



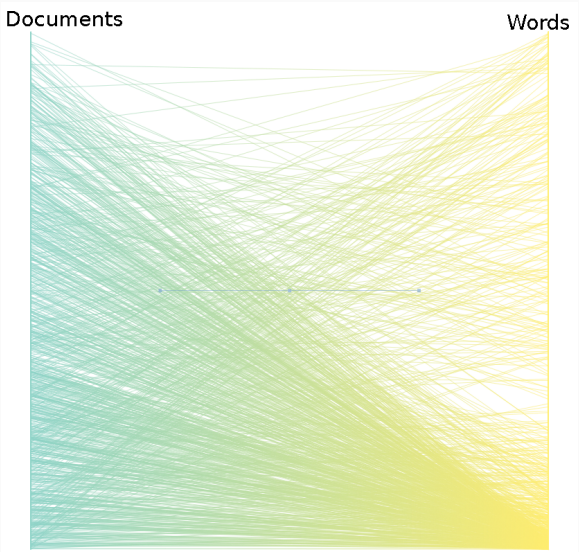
Results: Education Thought Leaders

Processing: **medium**
words: 43,429
documents: 10,000
topic levels: *none*



Results: Education Thought Leaders

Processing: **maximum**
words: 43,429
documents: 10,000
topic levels: *none*



Conclusions

Document type matters!

- Synthetic biology ethics: Longer documents, complex writing
  - Number of topic levels remains constant
  - Change in topics (both number and substance) within levels
  - Minimal change in topic quality
- Education thought leaders: short documents, simple writing
  - Total loss of utility with increased preprocessing

Acknowledgements

Synthetic biology:
This work was supported by the National Science Foundation Office of Advanced Cyberinfrastructure's Harnessing the Data Revolution program under grant no. 1939887.

Education thought leaders:
This presentation is based on research funded by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.

Thank you!

If you have questions or comments or want to check out the code, do not hesitate to reach out.

Brandon SEPULVADO
Data Scientist
Methods and Quantitative Social Sciences Department
NORC at the University of Chicago

Twitter: @brsepulvado