Finding Topics in Health News Tweets with LDA

Oishani Bandopadhyay

Data Overview

- [UCI ML Repository]
- Health News in Twitter dataset (collected 2015)
- Tweets stored as text files with id, date & time
- Focusing on Reuters: global news, mid-sized

Data Cleaning

- Cleaned all text files:
  - Kept only tweet
  - Removed id, date & time
- Stored cleaned txt files in new folder for LDA

Goal

- Topic modeling using LDA:

  Identify topics within ‘documents’ (tweets in this case), and the most highly weighted words within each topic

Potential Issues

- Short documents (tweet length short)
- Already within a specific topic (health news)
- Coherence difficult to interpret and optimize

Questions

- Does LDA work on shorter text data?
- How many topics are ‘good’?
- Within a specific domain, eg, health news, what subtopics appear?
- How do topics in women’s, men’s, children’s health vary?

Pre-Processing

- Use regex to remove irrelevant parts of text
- Tokenize and remove stopwords
- Lemmatize tokens
- Create dictionary and corpora using gensim

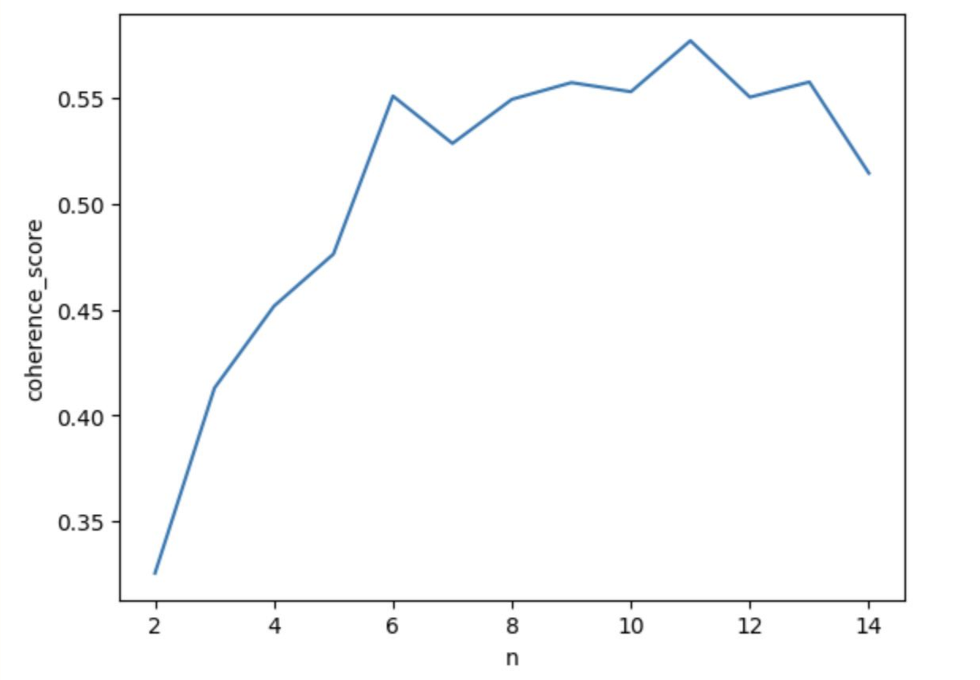
LDA Overview

- Bayesian inference model
- Assumes every document has a relatively small number of topics, topics in documents are in a probability distribution
- Terms within topic are also in a probability distribution

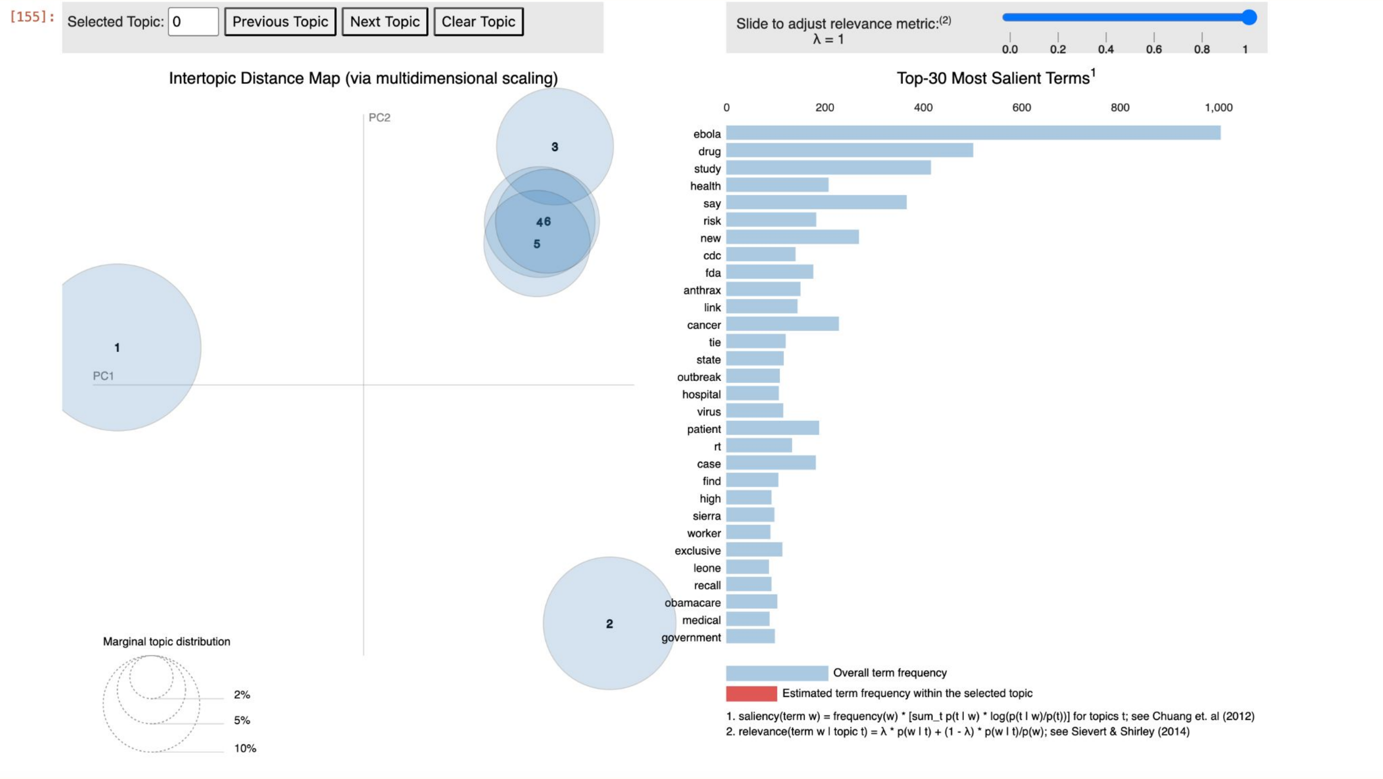
Coherence Overview

- Looks at co-occurrence of words together
- Each subset of words gets a conformation score based on vector similarity
- These scores are aggregated to get coherence for that number of topics

Coherence Plot - Entire Reuters Data



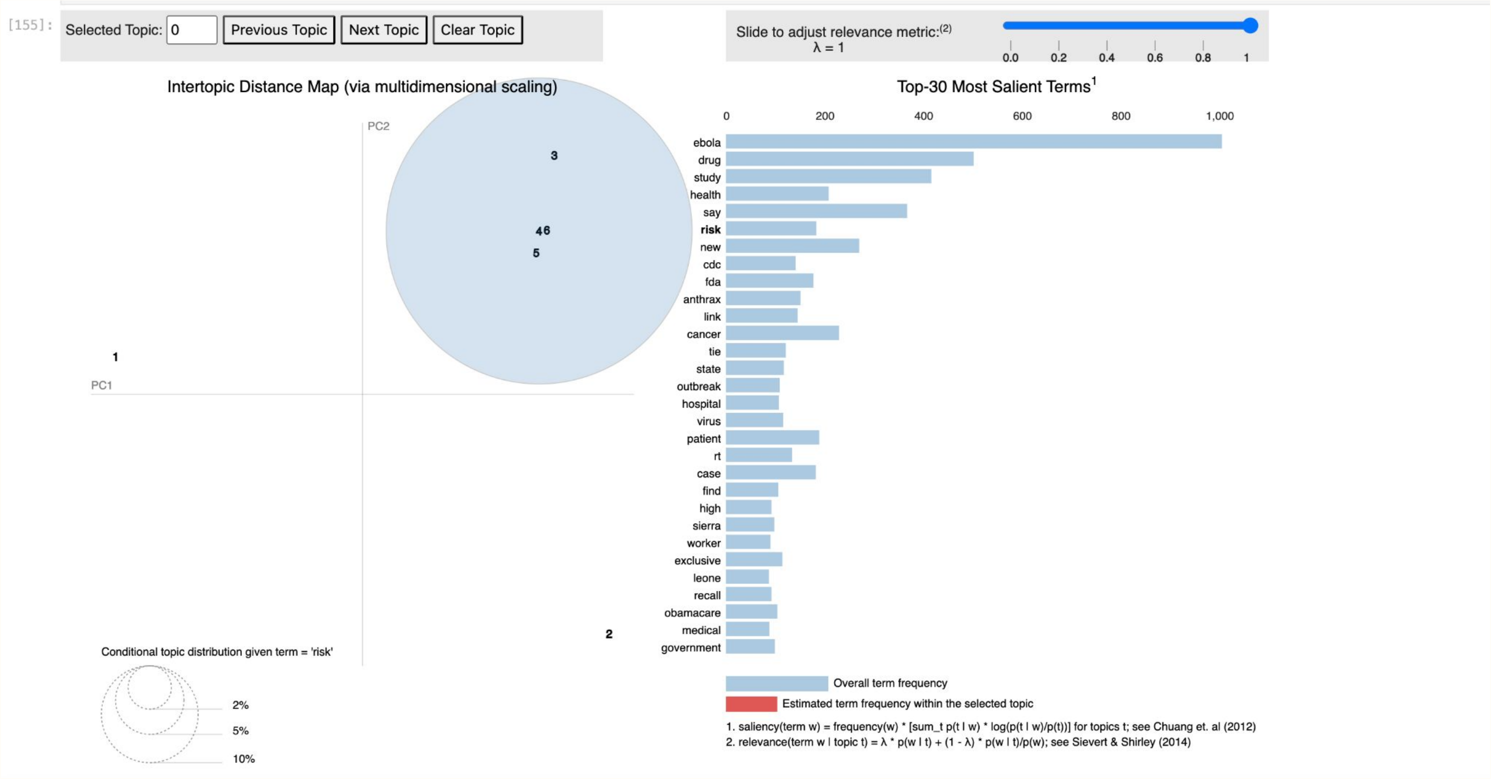
Topic Modeling - 6 Topics



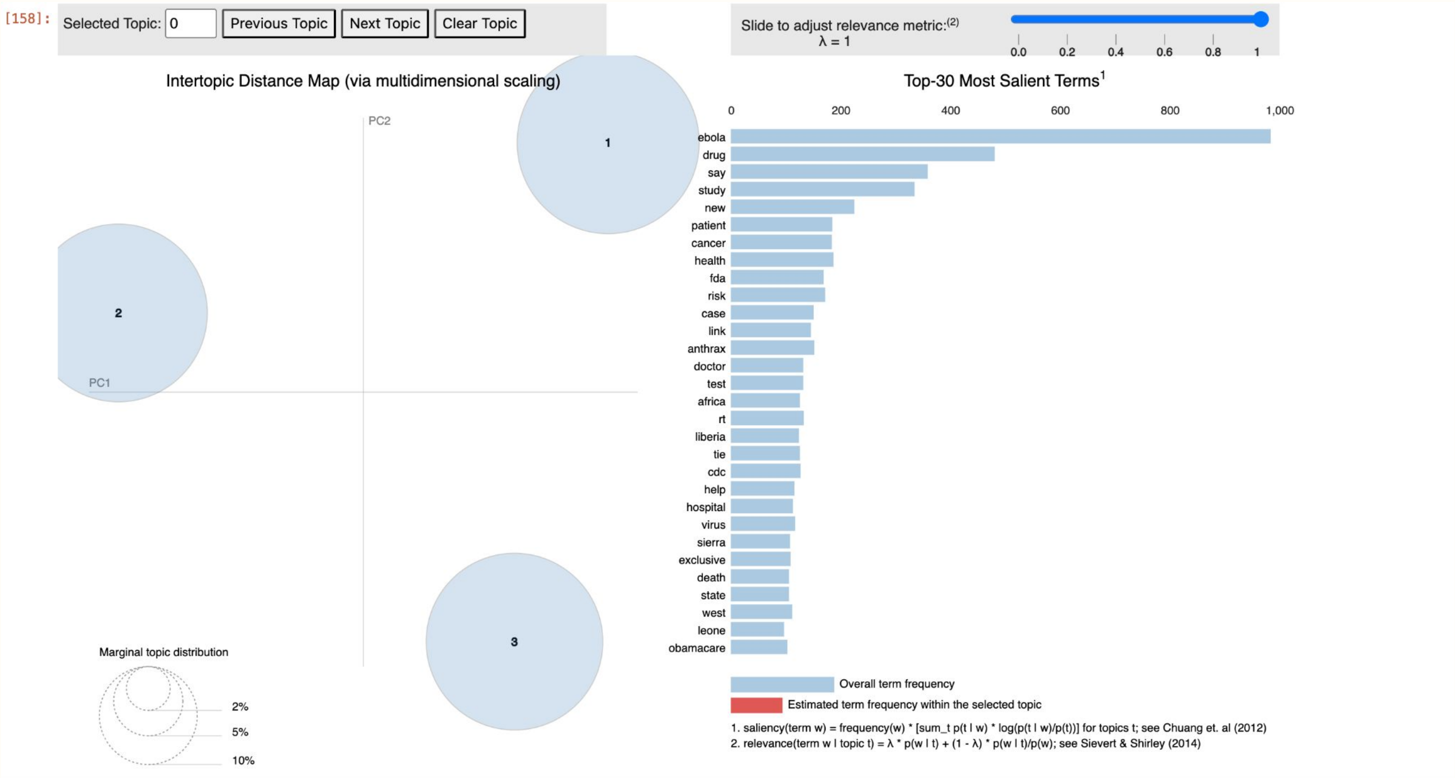
GPT-5 on Topics 3-6

- 🔹 Why they overlap
-
- These topics all cluster around infectious diseases, health policy, and public health crises, which explains why they appear close on the intertopic map.
-
- However, LDA splits them based on different contexts and frames of discussion:
-
- Topic 3: Disease outbreaks and crises with strong emphasis on fear, exposure, epidemics.
-
- Topic 4: Public risk perception, screening, and community-level response.
-
- Topic 5: Institutional health infrastructure and reporting.
-
- Topic 6: Mental health, pharma industry, and technology/lifestyle factors.
-
- Overlap happens because journalists often co-report these themes together (e.g., a hospital outbreak story also mentions risk, policy, and treatments), but the term co-occurrence patterns still form separate clusters.

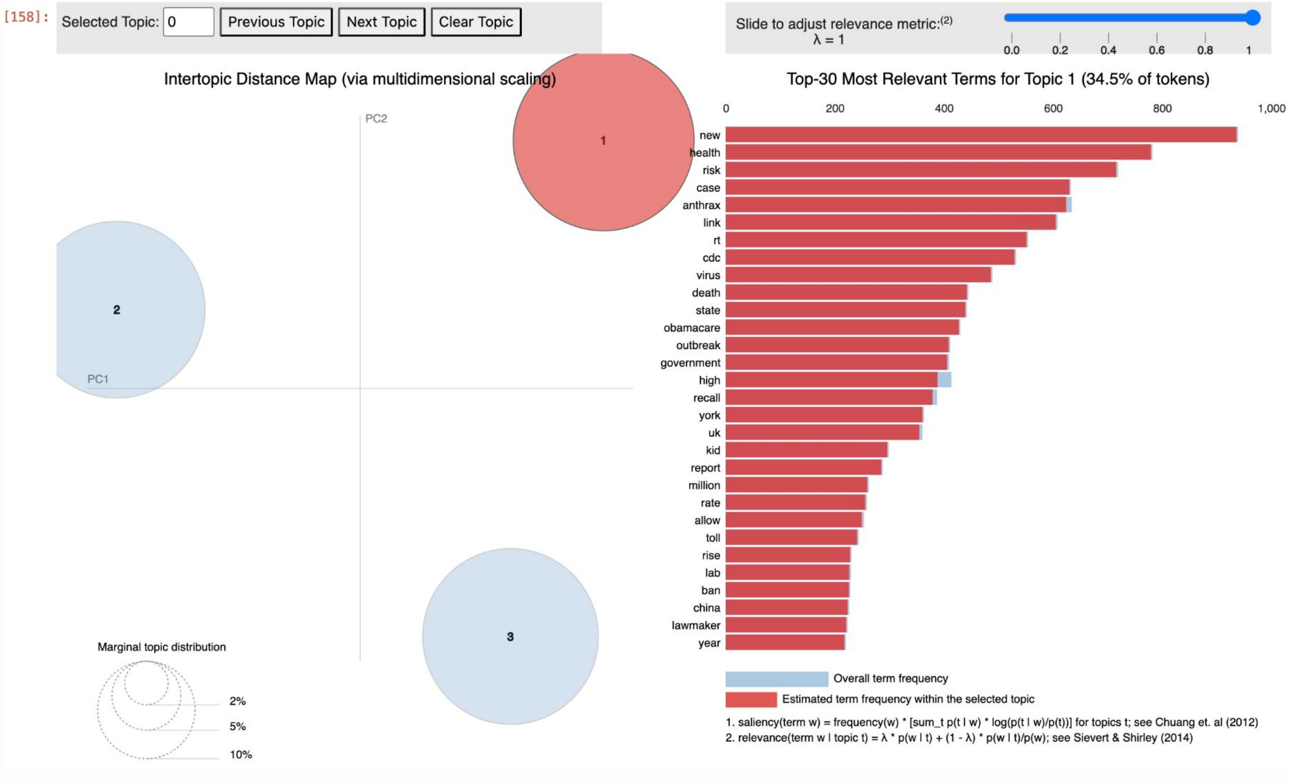
Topic Modeling - 6 Topics



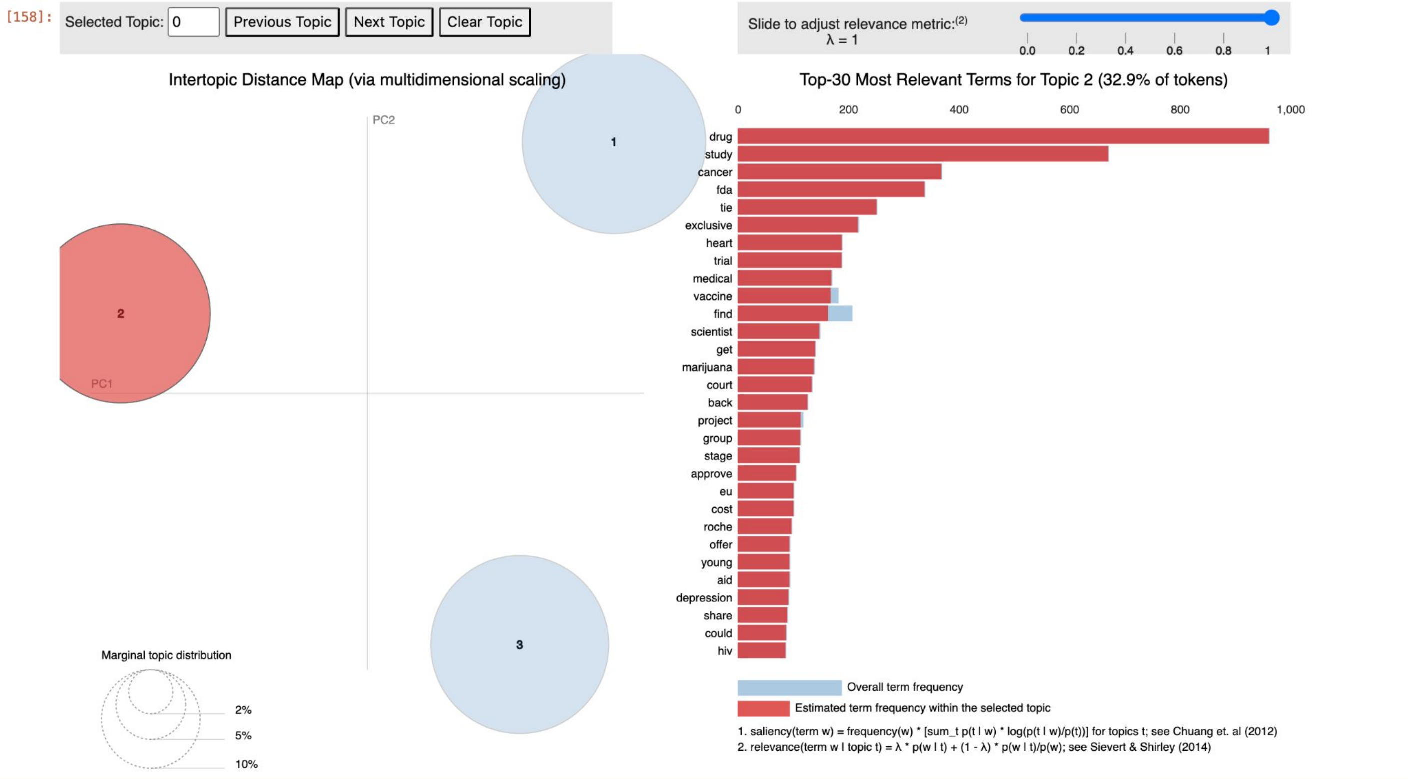
Topic Modeling - 3 Topics



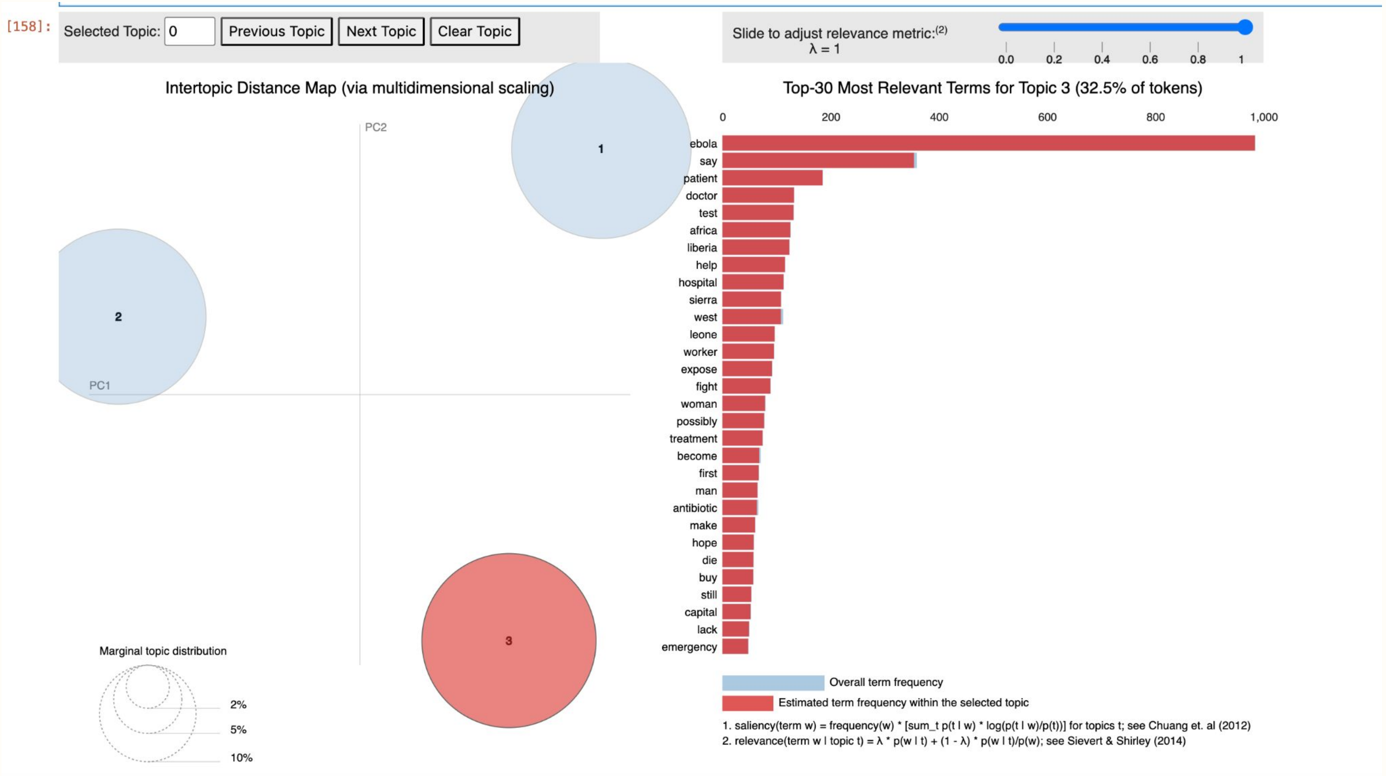
Topic Modeling - 3 Topics



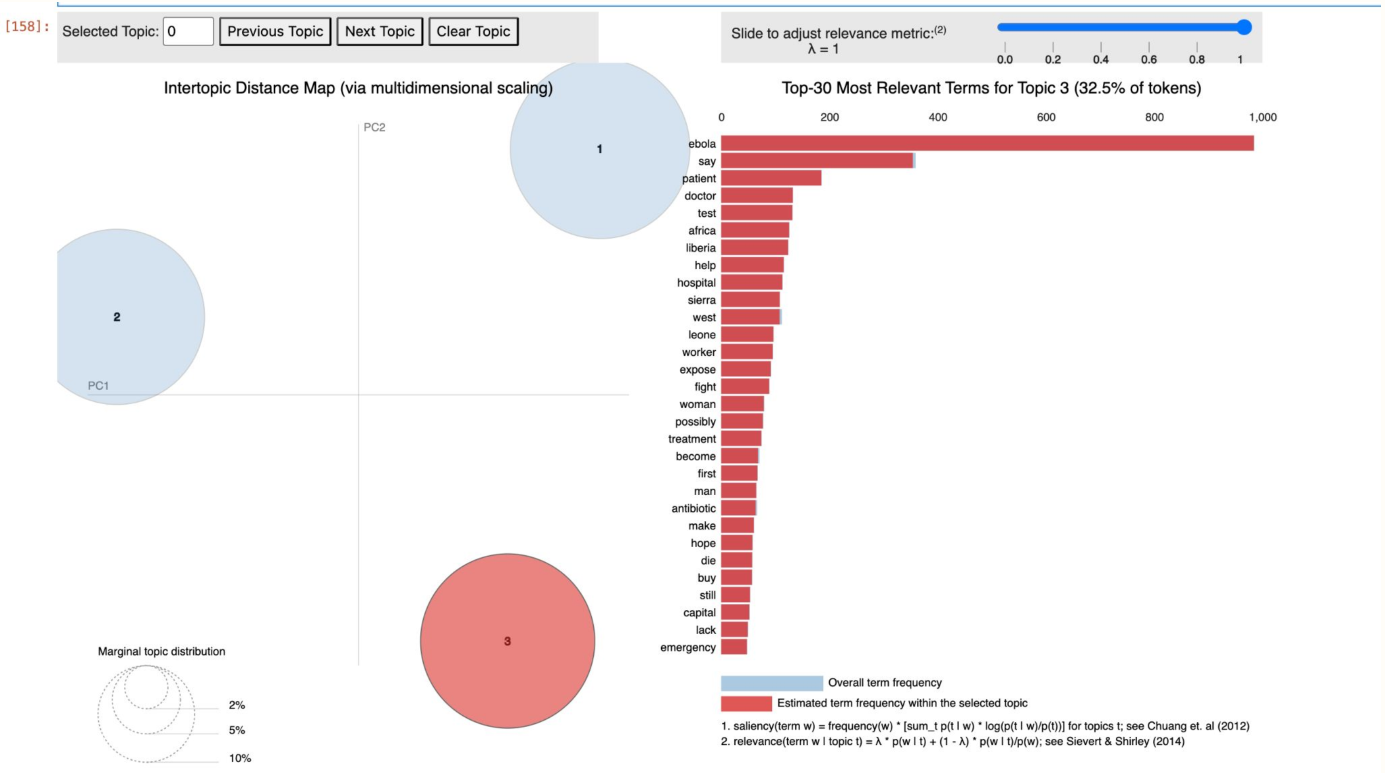
Topic Modeling - 3 Topics



Topic Modeling - 3 Topics



Topic Modeling - 3 Topics



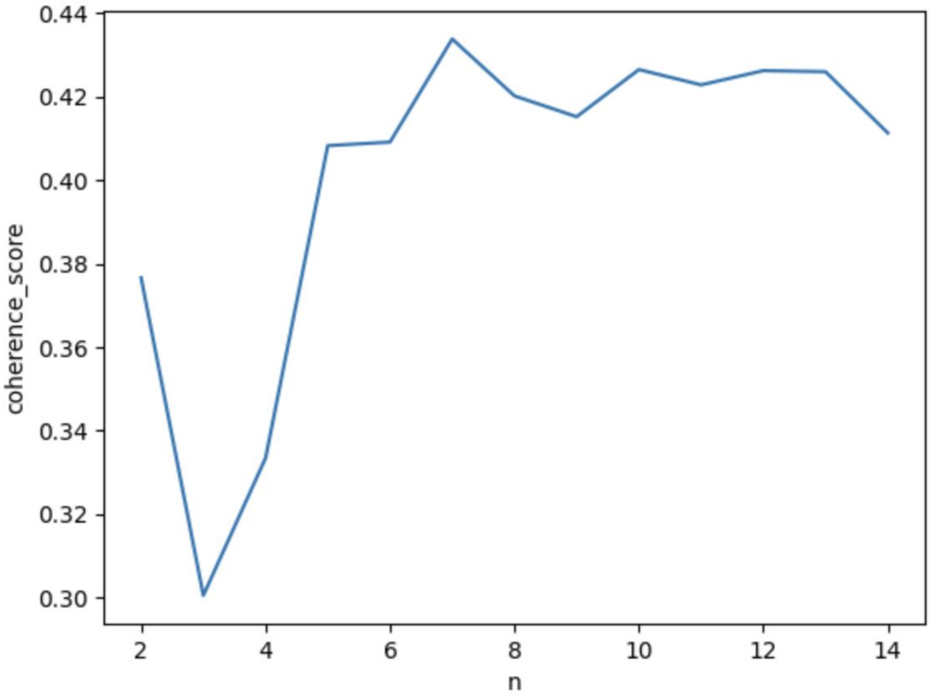
Women’s Health Filter

- Create a new corpus keeping only tweets with keywords about women’s health
- keywords = ['women', 'woman', 'female', 'menstrual', 'menstruation', 'abortion', "women's", 'girl', 'lady', 'birth', 'menopause', 'mother', 'mom', 'childbirth', 'ladies', 'ovulation', 'uterus', 'breast', 'ovar', 'ovary', 'ovarian']

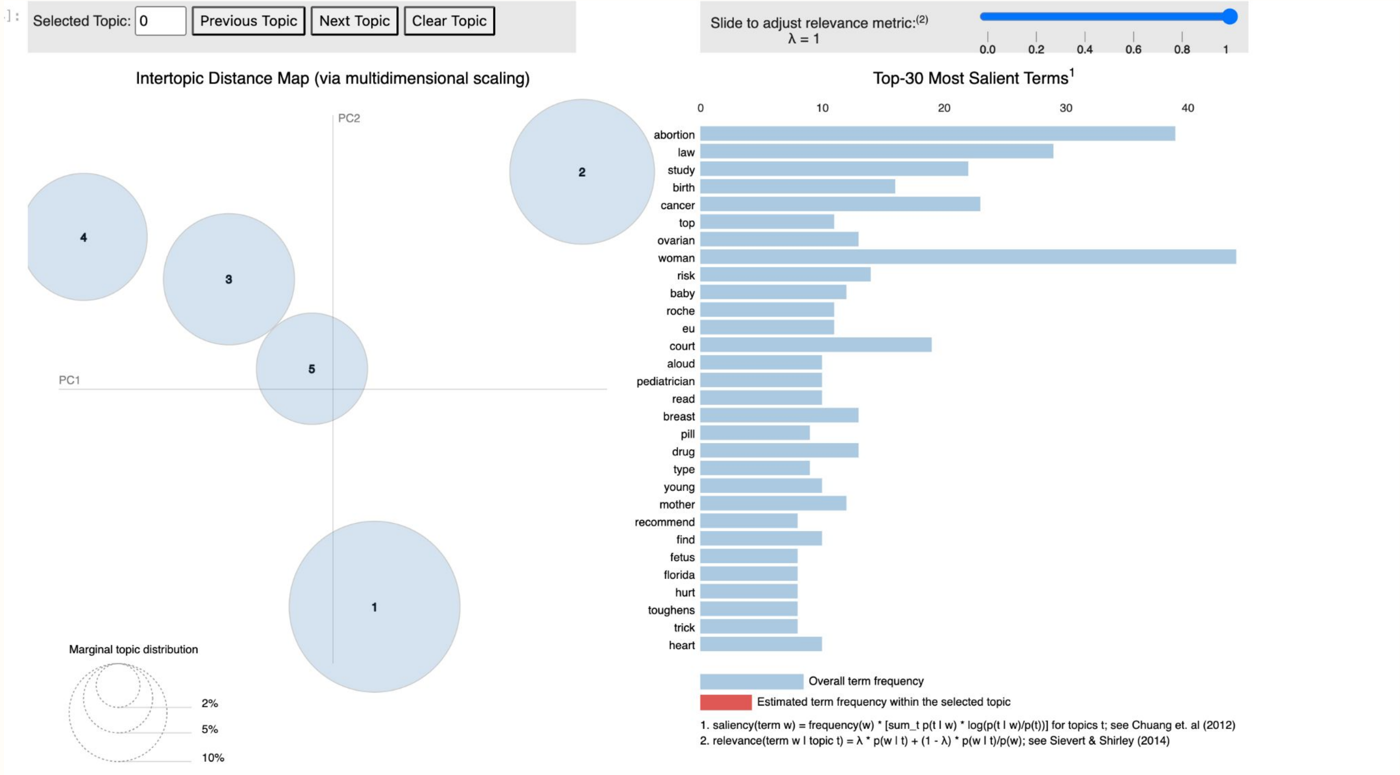
Women’s Health Filter

- Create a new corpus keeping only tweets with keywords about women’s health
- keywords = ['women', 'woman', 'female', 'menstrual', 'menstruation', 'abortion', "women's", 'girl', 'lady', 'birth', 'menopause', 'mother', 'mom', 'childbirth', 'ladies', 'ovulation', 'uterus', 'breast', 'ovar', 'ovary', 'ovarian']

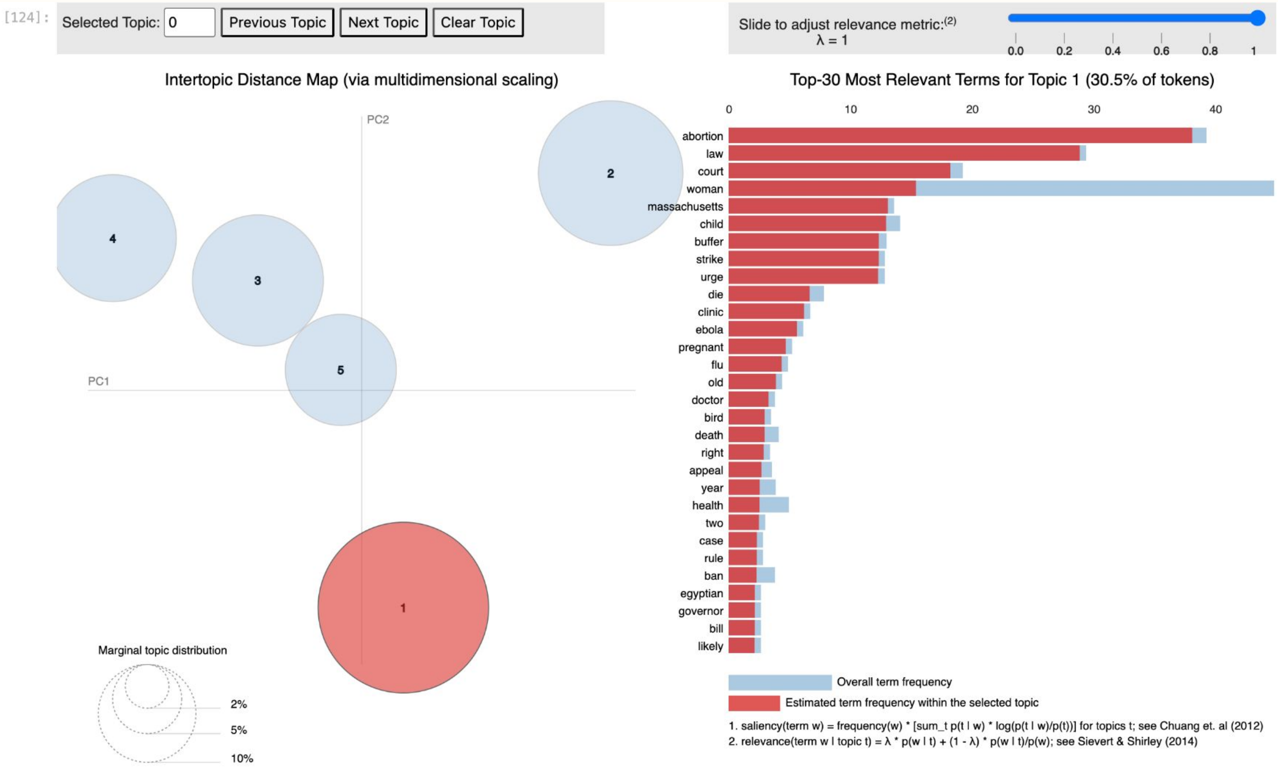
Coherence Plot - Women's Health Data



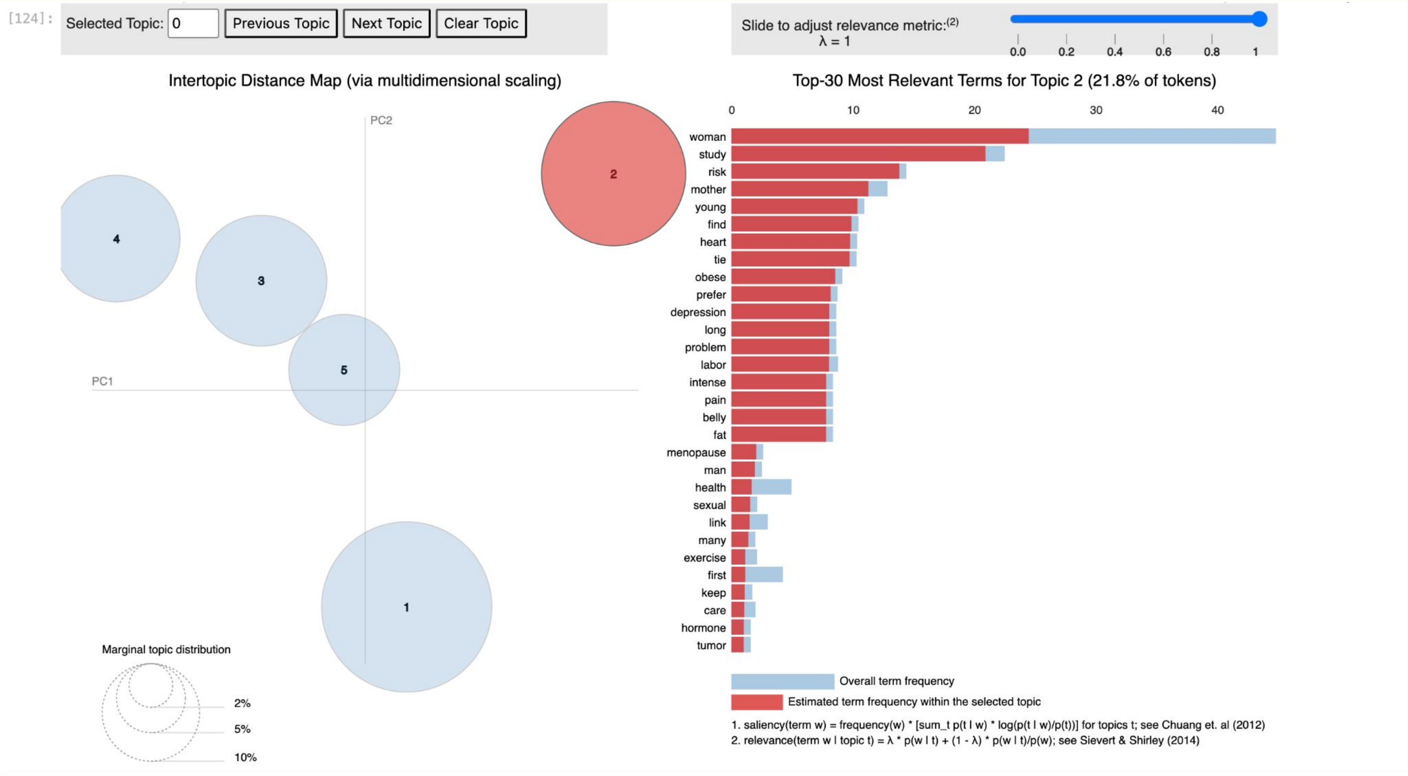
Topic Modeling - 5 Topics



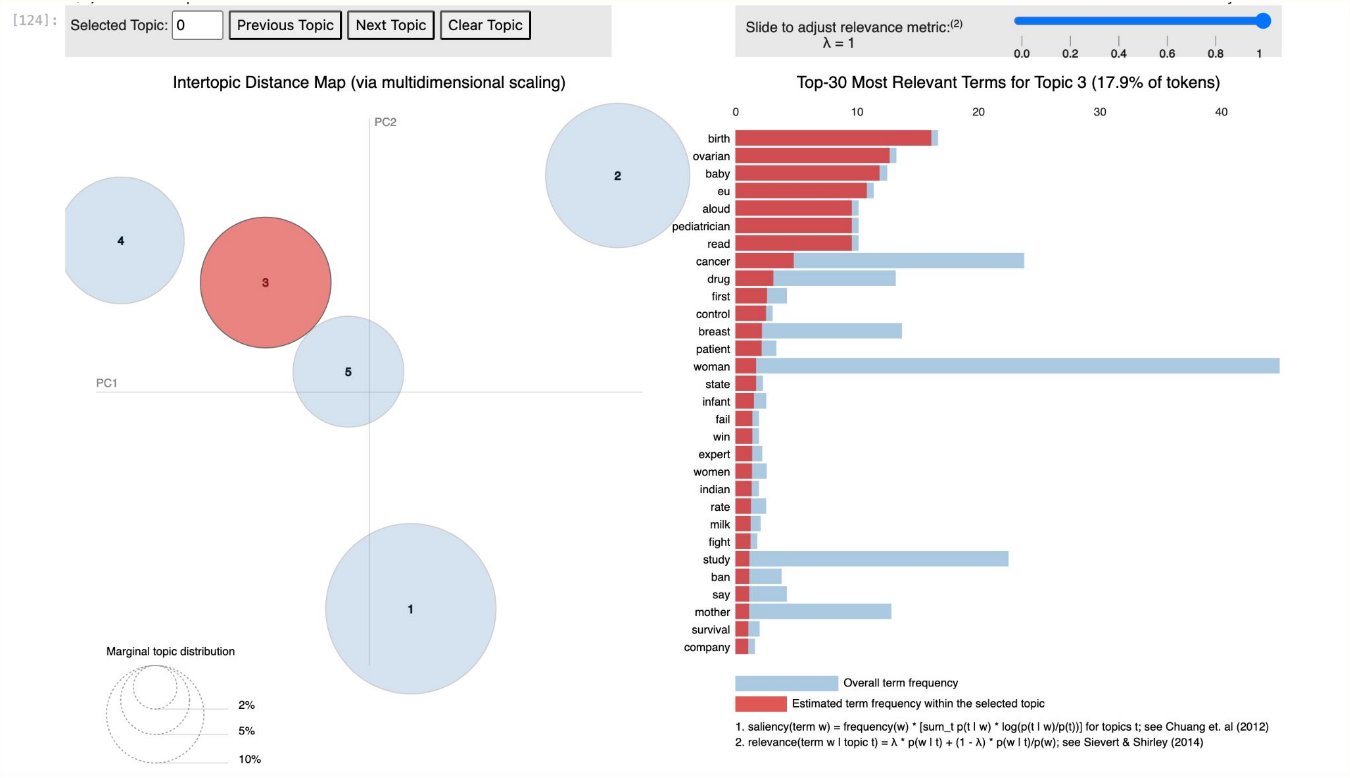
Topic Modeling - 5 Topics



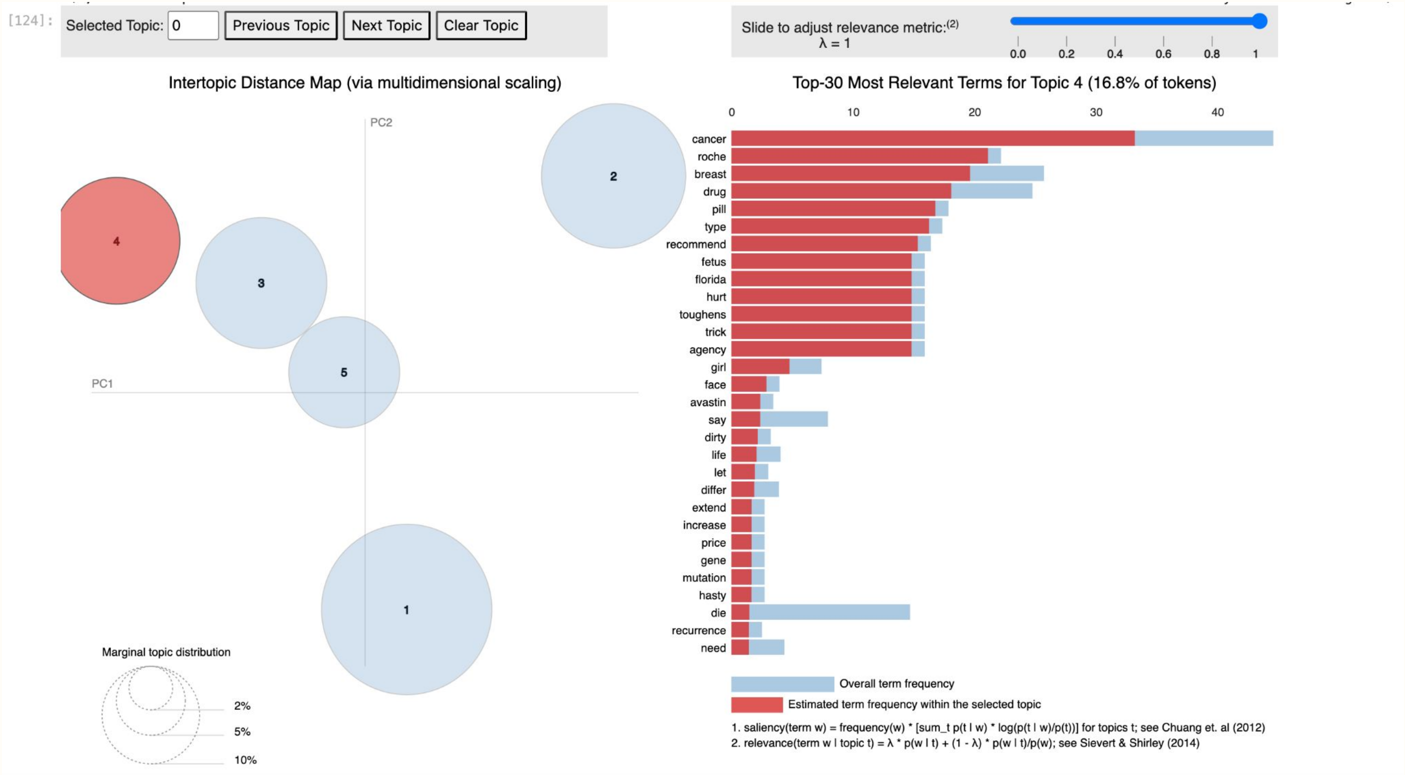
Topic Modeling - 5 Topics



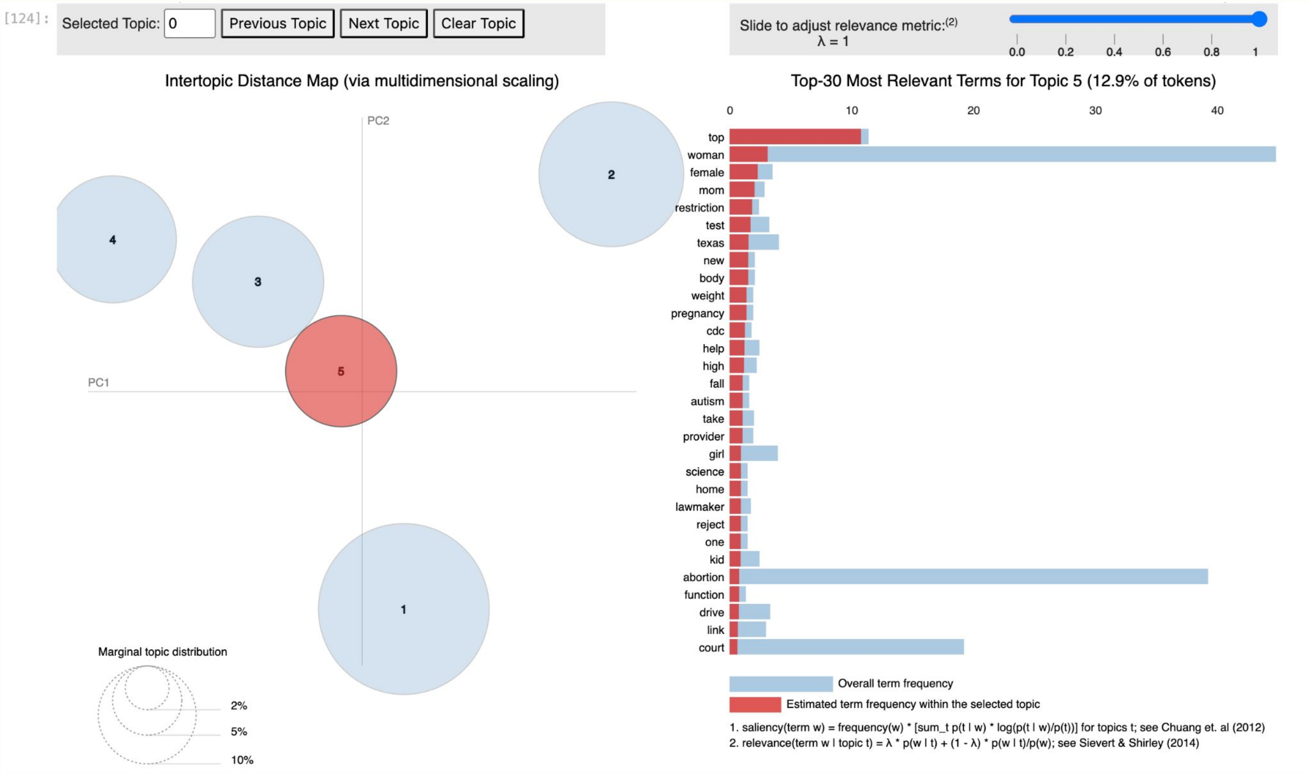
Topic Modeling - 5 Topics



Topic Modeling - 5 Topics



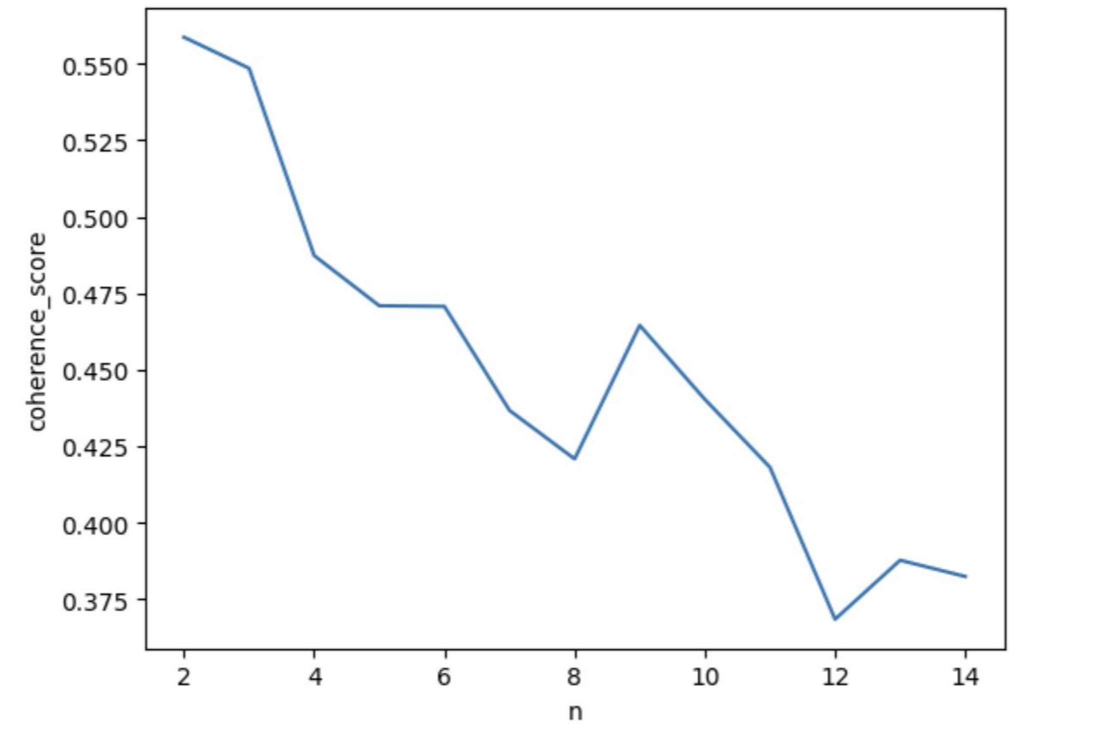
Topic Modeling - 5 Topics



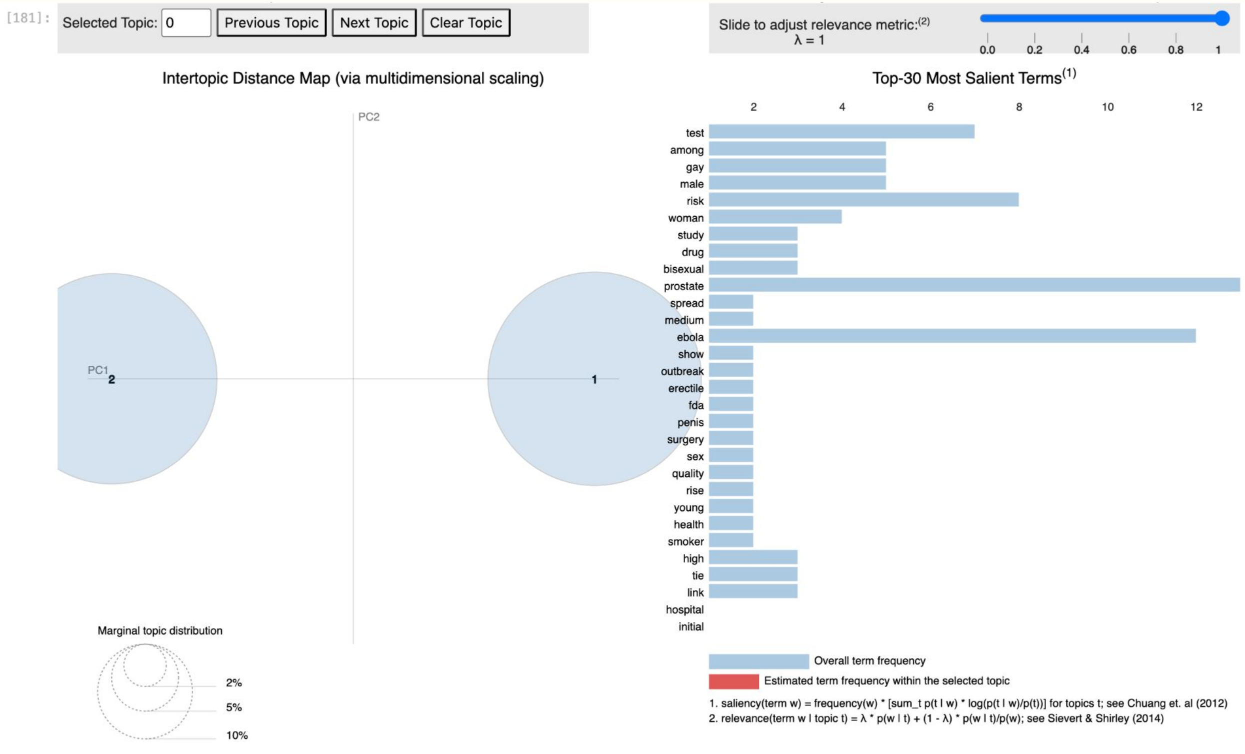
Men’s Health Filter

- Create a new corpus keeping only tweets with keywords about men’s health
- keywords = ['men', 'man', 'male', 'testicular', 'prostate', 'sperm', "men's", 'erectile', 'gentleman', 'semen', 'penile', 'penis', 'vasectomy', 'gentlemen', 'erection', 'testes']

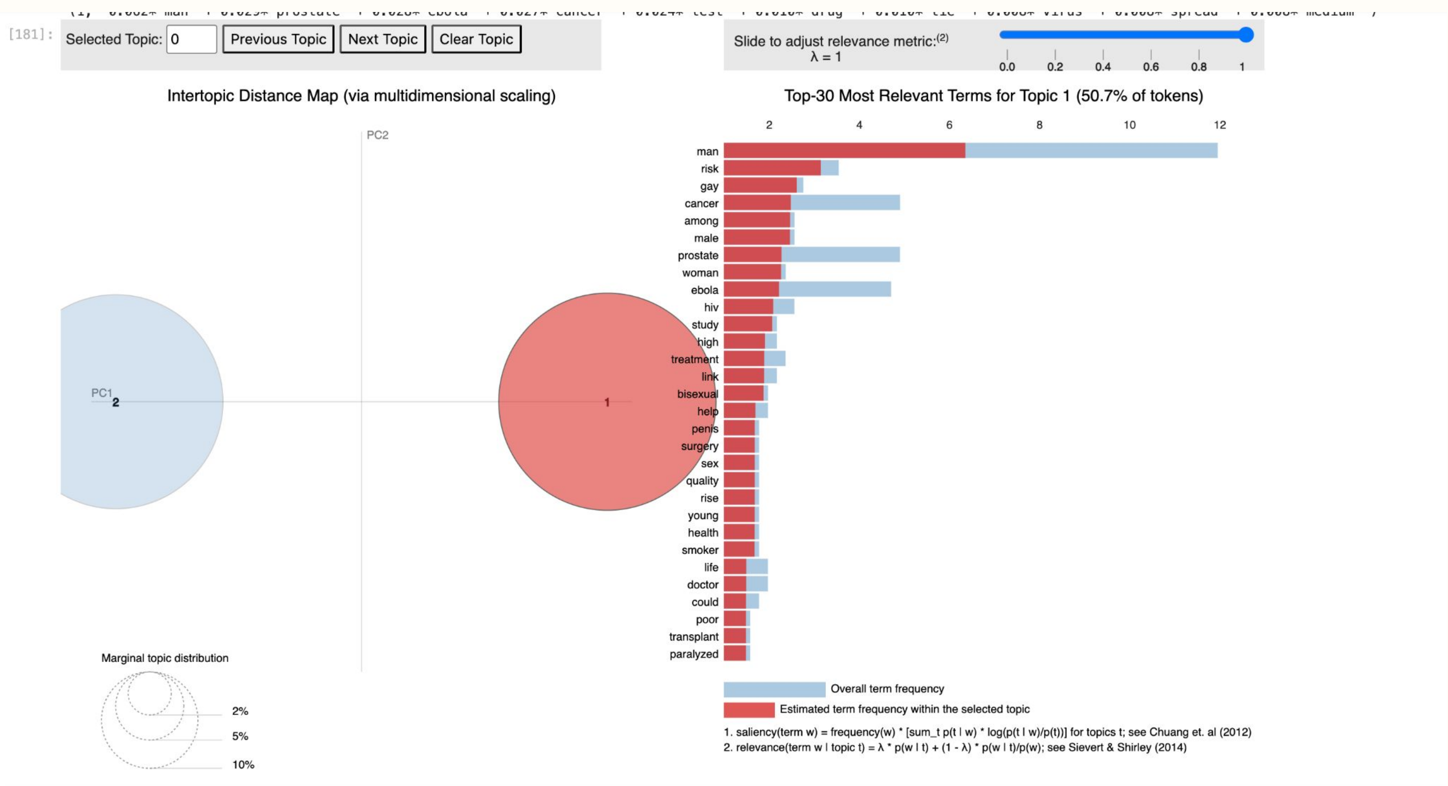
Coherence Plot - Men’s Health Data



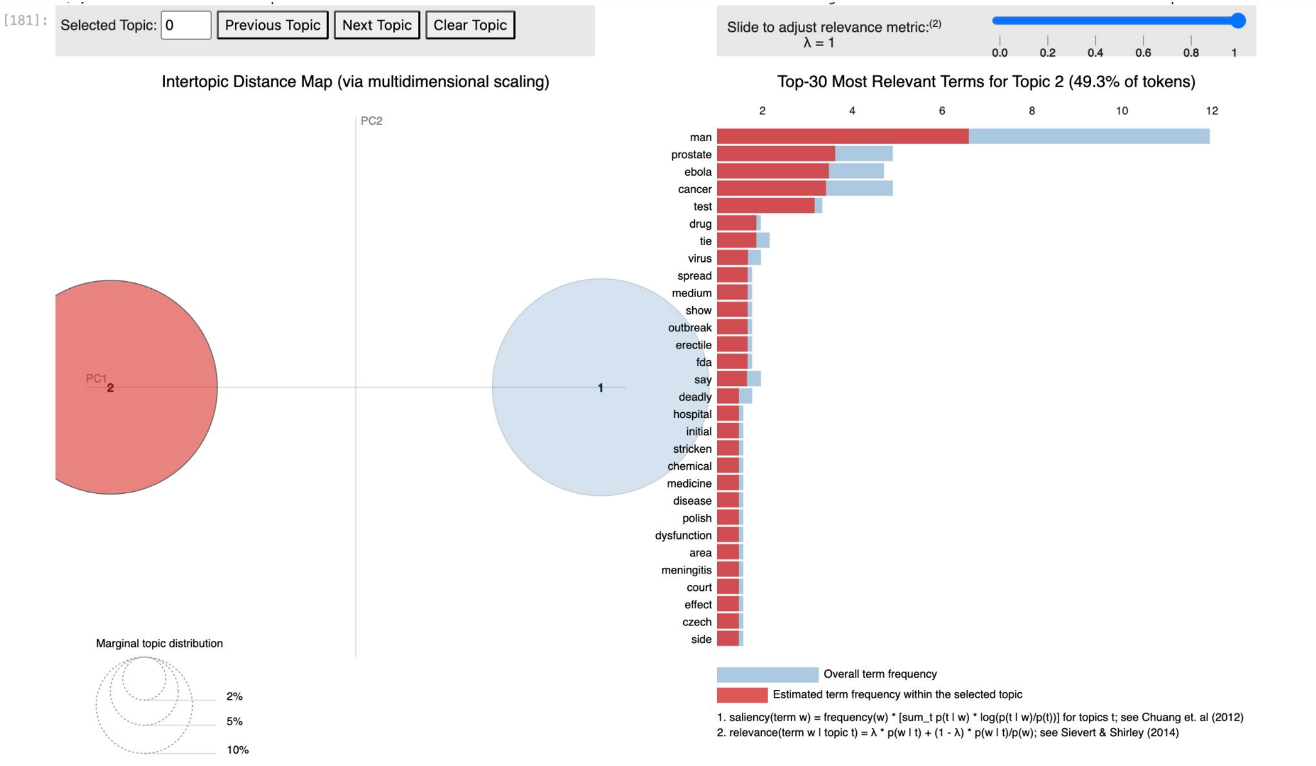
Topic Modeling - 2 Topics



Topic Modeling - 2 Topics



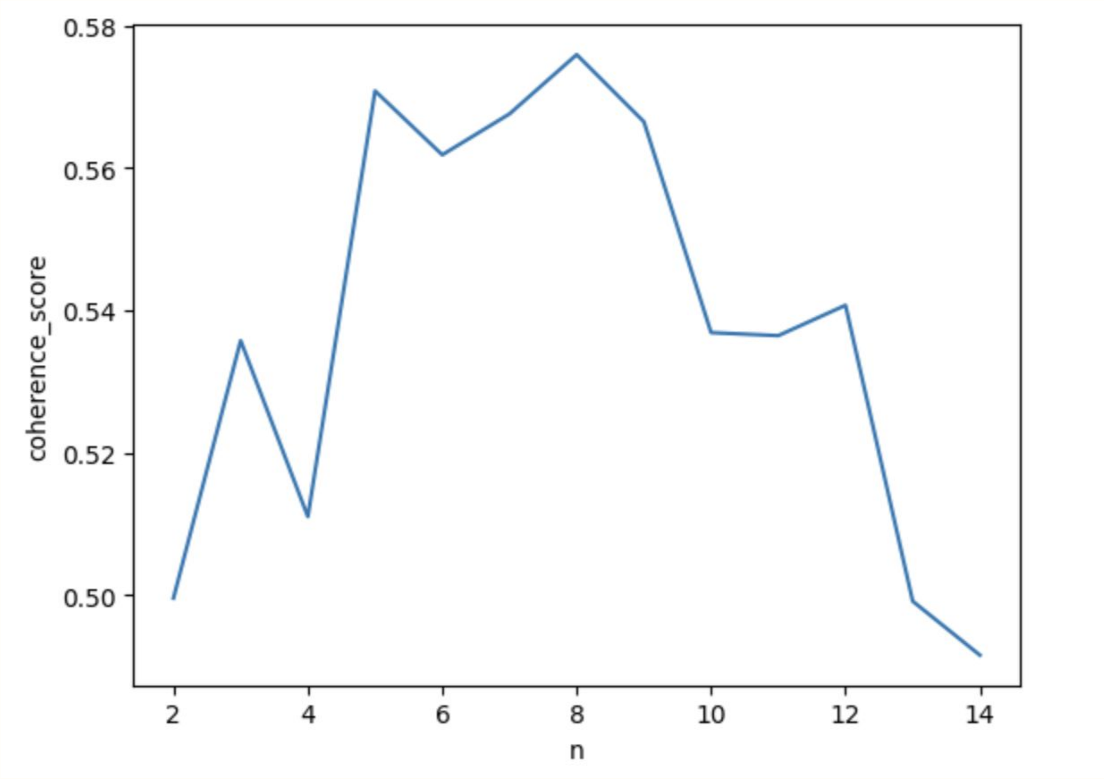
Topic Modeling - 2 Topics



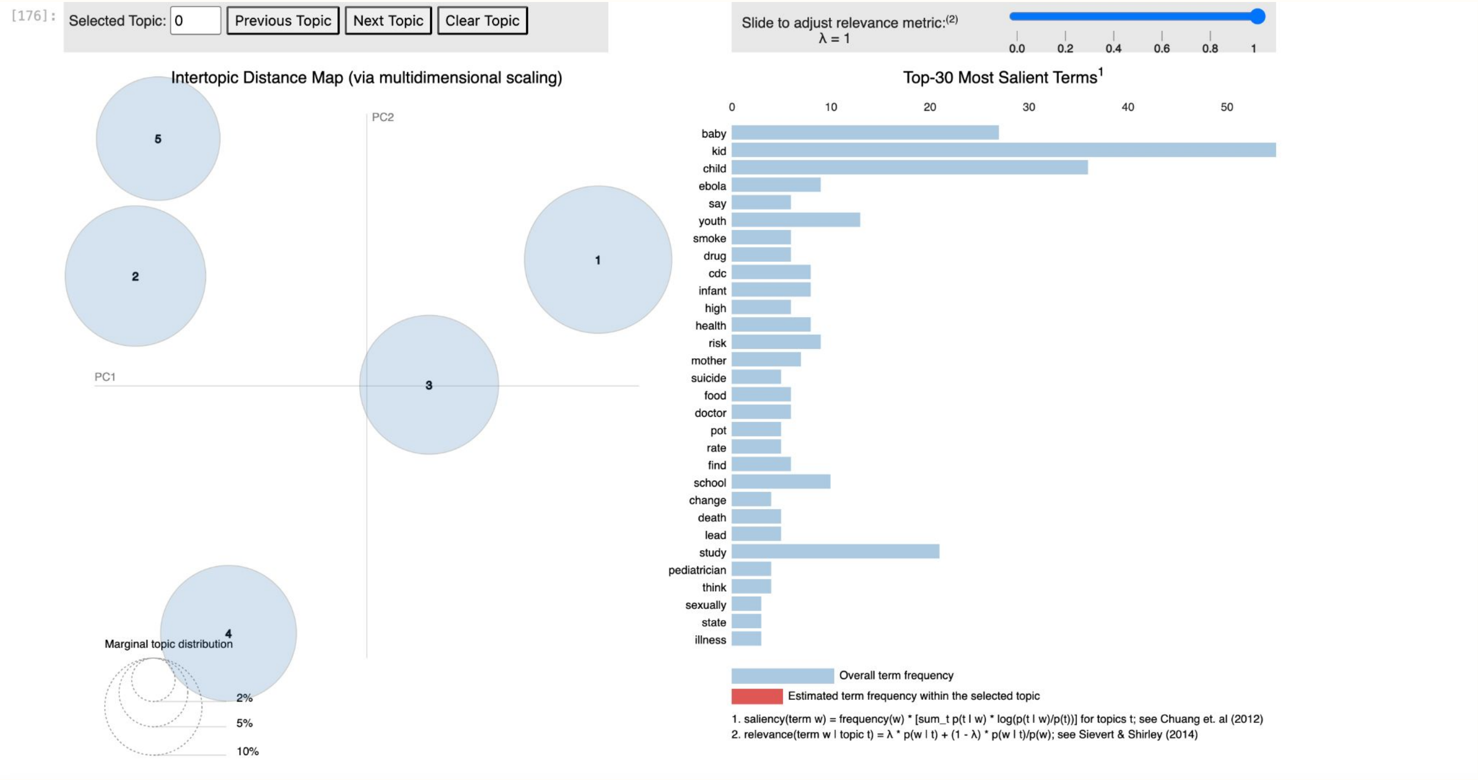
Children’s Health Filter

- Create a new corpus keeping only tweets with keywords about children
- keywords = ['men', 'man', 'male', 'testicular', 'prostate', 'sperm', "men's", 'erectile', 'gentleman', 'semen', 'penile', 'penis', 'vasectomy', 'gentlemen', 'erection', 'testes']

Coherence Plot - Children Health Data



Topic Modeling - 5 Topics



Answers to Questions

- Subsetting documents by keywords shows interesting spread of topics
- One of the women's health topics has more political terms than the entire data, men's data, or children's data
- Kids' salient terms include mother, not father

Answers to Questions

- Topic modeling can work for smaller documents within a specific subdomain!
- Choosing number of topics doesn't rely on coherence graph elbow
  - Too similar documents
  - Fewer topics more interpretable

Issues

- Coherence doesn’t work as well to find topic numbers because of semantic similarity in the entire data
- LDA is generally worse for smaller documents
- Subsetting data changes the number of documents drastically