# Developing a Curated Topic Model for
# COVID-19 Medical Research Literature

**Philip Resnik**
Linguistics/UMIACS
University of Maryland
resnik@umd.edu

**Katherine E. Goodman**
Epidemiology/Public Health
UMD School of Medicine
kgoodman@som.umaryland.edu

**Mike Moran**
SoloSegment
mike@solosegment.com

## Abstract

Topic models can facilitate search, navigation, and knowledge discovery in large document collections. However, automatic generation of topic models can produce results that fail to meet the needs of users. We describe an effort in progress to develop a *curated* topic model for COVID-19 articles informed by subject matter expertise and the way medical researchers engage with medical literature.

## 1 Introduction

The language technology community has been responding in force to the coronavirus pandemic with sustained energy and creativity, and the medical research literature, facilitated by the CORD-19 dataset, is a major hub of activity (Wang et al., 2020). A crucial goal of all this technological effort is to support *non-technologist* users, namely the medical researchers, clinicians, and policy makers who are involved with the response to COVID-19. However, to date there has been little explicit discussion in the technology-development community about what functionalities will actually help these users in the trenches. Zhang et al. (2020) sum things up nicely when they write, "we don't actually know how our systems . . . can concretely contribute to efforts to tackle the ongoing pandemic until we receive guidance from real users who are engaged in those efforts . . . [The] challenge [is] how to build improved fire-fighting capabilities for tomorrow without bothering those who are trying to fight the fires that already raging in front of us".

In this paper we describe a cross-disciplinary collaborative effort that is intended to help close that gap by developing a *curated topic model* for COVID-19 medical research literature, informed both by subject matter expertise about the domain and by the way that medical researchers typically engage with medical literature. Our efforts make use of a human-in-the-loop platform for topic model development informed by the research literature on interactive topic modeling as well as by practical experience developing topic taxonomies for large scale document collections. The outcome of this work in progress is not a new algorithmic contribution; rather, we plan to publicly release a curated model that can be used downstream by the community as a resource and as a starting point for further work.

## 2 Desiderata

The core idea in topic modeling is to take in a collection of documents, discover a latent set of topics characteristic of the collection, and represent each document as a mixture of those topics. Typically each latent topic is itself represented as a distribution over the collection's vocabulary. Latent Dirichlet Allocation (Blei et al., 2003) is the ancestor of a whole variety of topic model variants, adding hierarchical topics, supervision, temporal structure, discrete covariates, and more, as well as more recent neural variations. In most applications of topic models, the topic-word distribution — i.e. the representation of each topic by its most probable $n$ words, for small $n$ — provides human consumers of the topic model with intuitions about what each topic is about, as a way of understanding the contents of the collection.

However, automatically discovered topic models are often subject to noise and poor quality (Boyd-Graber et al., 2014). In addition, the usual products of topic modeling can be a poor fit for real-world use cases involving non-technical subject matter experts (SMEs) within a particular domain. From our own experience and discussions with medical researchers, we have identified a set of desiderata for topic modeling in the context of the COVID-19 literature.

- SMEs are unlikely to be satisfied with purely bottom-up discovery of topics. In order to obtain buy-in, topic models need to be substantially consistent with experts' pre-existing view of the world — even if one advantage of topic models is also to discover categories that were previously unknown.

- Related, SMEs need to trust the resources they use. In medicine there is a long tradition of painstakingly constructed category systems. Sometimes these involve boundaries that are *too* crisp; for example, over-rigidity in DSM diagnostic categories has motivated development of the multidimensional RDoC system in psychiatric research (Owen, 2014). However, the goal should be to augment rather than replace human training and expertise.

- Topics need readable labels. Reading a list of words to intuit an underlying topic or concept will not suffice.

- There is likely to be high value in a special purpose category system for COVID-19, particularly for epidemiologists and specialists in clinical infectious diseases. General purpose categories (e.g. MeSH, Lipscomb (2000)) are unlikely to be sufficient for a rapidly emerging space with its own characteristics and concepts.

We find that these observations are largely in line with the principles for automated text analysis articulated by Grimmer and Stewart (2013) for the social sciences. In particular, they emphasize the role of automated analysis in expanding human capabilities, that models should be evaluated not intrinsically but in terms of their ability to support scientific goals, and they advocate strongly against the use of automated models without some form of validation.

We envision many uses for a curated topic model, but our driving use case involves the way many if not most medical researchers engage with the literature. A typical researcher is intimately familiar with Boolean searches in PubMed, including use of wildcards (e.g. `bacteri*`), quoted multi-word phrases (e.g. `''white blood cell''`), fielded searches (e.g. `aromatherapy[TIAB]` specifying to look in the articles' title, abstract, or keywords), and MeSH taxonomy headings and subheadings (e.g. `hypertension[mh] AND toxicity[sh]`). When reviewing search results, users will typically look at the article list and then, for a specific article, it is extremely common to drill down by looking at lists of articles surfaced by PubMed as *similar to* and/or *citing* the current article being looked at. Clicking through to other articles can be interleaved with new or refined Boolean searches.

The overall picture here is a very structured kind of search. In their day to day experience navigating the medical literature, it would be a fairly drastic shift from this way of doing things to typing in full-sentence questions in the style of question-answering systems or queries in the style of web search engines. When thinking about topic models, this motivates our thinking less in terms of, for example, improvements to ranking or visualizing topic-based clusters, and more in terms of discrete topic labels and how they could be introduced into the user's familiar, discrete experience, more analogous to MeSH terms. We would anticipate that one particularly valuable use of curated, labeled topical categories will be in helping researchers to overcome information overload when navigating results. For example, augmenting documents in the collection with discrete topic category labels would make it possible to organize related-article and citing-article lists into subcategories, and to include topical categories in follow-up searches.

## 3 Human-in-the-loop process

We use an interactive, human-in-the-loop topic modeling platform to produce a curated topic model. The process we will be following is informed by prior experience using this platform to develop curated taxonomies for large document collections.[1]

We begin with preprocessing, including conventional steps such as tokenization (including identification of relevant multi-word expressions), lowercasing, removal of stopwords, and down-selection of the vocabulary to high-value words based on frequency and other statistical properties.[2] This is

---

[1] As an example, a topic taxonomy was created for hundreds of thousands of documents in the content management system for a major national professional organization, producing human-readable topics that were integrated into faceted search on the organization's web site. Because we are using a commercial system and the model itself is the intended contribution, we provide an overview rather than sharing full technical details.

[2] We will use *word* to refer to both unigrams and multi-word tokens.

followed by creation of multiple initial topic models of differing granularities. We do not optimize the number $K$ of topics automatically, since doing so typically relies on automatic approximations to human judgment such as normalized pointwise mutual information (NPMI, Aletras and Stevenson, 2013; Lau et al., 2014). Rather, we will use human judgments directly by constructing models across a range of $K$ and assessing how promising each model is as a starting point, via a combination of qualitative assessment and by comparing human quality ratings for a random sample of topics as assigned by two independent SMEs.

Having selected an initial starting point, the human-in-the-loop process includes drilling down to better understand the model (including, for example, identifying documents that are highly representative of a given topic, or visualizing topic similarity), interleaved with human-feedback operations of the kind investigated by Lee et al. (2017) and Smith et al. (2018). For example, they discuss within-topic feedback such as removing or adding words as good signals for a topic, as well as model-level feedback such as merging equivalent topics or removing topics that are too incoherent to be worth refining. Topics are also assigned labels manually, and labels can be revised at any time as the SME's understanding of a topic evolves. This feedback is provided in batch, and then the model is recomputed using the feedback to provide inductive bias. After the recalculation, the SME inspects the resulting model, and can either continue another iteration of refinement or designate the model as final. Typically two to four iterations of refinement using this process are sufficient to produce a high quality model.

Once a final model has been produced, we will evaluate its quality using multiple SMEs in epidemiology or clinical infectious diseases. We plan to use both subjective topic coherence ratings on a Likert scale (Aletras and Stevenson, 2013) and word intrusion (Chang et al., 2009), And to look not only at summary measures of agreement (e.g. correlation) but at specific sources of disagreement.

## 4 Preliminary analysis

We conducted a preliminary analysis using the titles and abstracts in the May 1, 2020 release of the CORD-19 dataset (`metadata.csv`).[3] This exploratory modeling used spaCy (`spacy.io`) for tokenization and identified phrase chunks in preprocessing as meaningful semantic units (Mimno, 2020). Although spaCy's phrase chunking performed very well, Mimno observes that "for text with lots of technical terms, a carefully curated list of multi-word terms can make a huge difference", and in subsequent exploration of initial models we will integrate COVID named entity resources to identify biomedical multi-word expressions, e.g. entities in Kroll et al. (2020).

We constructed and inspected initial models with $K = 50, 100, 150$ topics. On drilling down into the models, it became apparent that the collection of documents contained a high proportion of articles that had been retrieved by the CORD-19 dataset search terms but were not directly relevant to COVID-19. In addition, we found topics aggregating words in non-English languages. On further inspection, we came to the conclusion that a cleaner set of topics for the COVID-19 research literature is likely to be created if the inclusion criteria for the document collection are stricter, limiting to articles sufficiently "about" COVID-19 that one of the inclusion search terms used in creating the CORD-19 dataset (Wang et al., 2020) appears in the title or abstract.

We have also determined that a simple heuristic filter for English documents works well: requiring that the title or abstract contain at least one of the most common words in English.[4] Of the roughly 3000 items filtered using this heuristic (about 5,500 in the June 7 release), only a tiny number include abstracts in English; the heuristic picks out non-English as well as articles where a title is provided but the abstract is empty.

Our initial impressions of topic granularity suggest that $K = 50$ topics is going to be too coarse grained for a collection of this size. This is consistent with our prior experience with human-in-the-loop modeling for collections with tens of thousands of documents, where we have found that 100-150 topics was a good initial starting point. For the interactive refinement process, it makes sense to err in the direction of fine-grained topics in the initial model, since the refinement process makes it easy

---

[3]We observe that use of titles and abstracts, rather than full text, is consistent with how article similarity is calculated in PubMed (PubMed Help, 2020).

[4]Specifically *the, be, am, are, is, was, were, being, been, to, of, and, a, an, in, that, have, i, it, for, not, on,* and *with.*

to whittle down the set by identifying, sharpening, and merging topics that cover the same conceptual territory and by removing manifestly incoherent topics.

To provide an impression of the topic model obtained in our initial process we show three randomly chosen topics from the 100-topic model for the May 1 release, noting that initial model results are expected to change once we implement the stricter inclusion criteria and medical entity processing:

- traditional_chinese extract plant plants extracts compounds activities medicine biological_activities tcm chinese natural_products chinese_medicine natural glycyrrhizin flavonoids quercetin traditional_chinese_medicine action inflammatory

- mental_health anxiety life stress depression individuals self post women physical disorder psychological chronic_diseases scores scale exercise quality status health_status risk_factors

- membrane endoplasmic_reticulum plasma_membrane ifitm3 membrane_proteins membrane_protein cell_surface golgi transmembrane_domain golgi_complex golgi_apparatus transport domain membranes channel terminal transmembrane_protein transmembrane release ion_channel

## 5 Conclusions

We have argued for the importance and utility of a curated topic model for COVID-19 medical research literature, proposed criteria that such a model should meet, and we have sketched out an interactive topic modeling process to construct a model meeting those criteria.

This is work in progress, but preliminary analysis suggests that initial topic models are high quality, we have identified several steps for improving initial automatic modeling, and our process of human-in-the-loop refinement is designed to further refine and curate a model whose content will be driven both bottom-up by the data and top-down by subject matter expertise.

At the end of our curated model development process, we will make publicly available topic-word distributions with corresponding labels, document-topic distributions for the input documents, the vocabulary, and a script that enables preprocessing consistent with ours for new documents. This will provide the community with materials needed to visualize the topics and documents, to manually organize topics into a hierarchy, to compute topic posteriors for new documents, to treat a document's most-prevalent topic(s) as discrete labels, and to integrate topic-labeling of documents into users' experience in search engines.

More generally, one of our key aims is for this curated model to serve as a useful starting point for further model development, e.g. using our resource to construct informative priors in development of further models. In addition, we are seeking resources that would support making the human-in-the-loop platform available to a wider range of subject matter experts.

## Conflict Statement

Mike Moran is Chief Product Officer of SoloSegment, a private company that is commercializing the interactive topic modeling platform we use. Philip Resnik is an advisor to SoloSegment.

## Acknowledgments

## References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning research*, 3(Jan):993–1022.

Jordan Boyd-Graber, David Mimno, and David Newman. 2014. Care and feeding of topic models: Problems, diagnostics, and improvements. In *Handbook of Mixed Membership Models and Their Applications*, CRC Handbooks of Modern Statistical Methods, Boca Raton, Florida. CRC Press.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Hermann Kroll, Jan Pirklbauer, Johannes Ruthmann, and Wolf-Tilo Balke. 2020. A semantically enriched dataset based on biomedical NER for the COVID19 open research dataset challenge. *arXiv preprint arXiv:2005.08823*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

David Mimno. 2020. Using phrases in Mallet topic models. http://www.mimno.org/articles/phrases/. [Accessed: 2020-06-29].

Michael J Owen. 2014. New approaches to psychiatric diagnostic classification. *Neuron*, 84(3):564–571.

PubMed Help. 2020. PubMed Help. https://www.ncbi.nlm.nih.gov/books/NBK3827/. [Accessed: 2020-06-29].

Alison Smith, Varun Kumar, Jordan L. Boyd-Graber, Kevin D. Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. *23rd International Conference on Intelligent User Interfaces*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. CORD-19: The COVID-19 open research dataset. *ArXiv*.

Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly deploying a neural search engine for the COVID-19 open research dataset: Preliminary thoughts and lessons learned. *arXiv preprint arXiv:2004.05125*.