Authorless Topic Models: Biasing Models Away from Known Structure

Laure Thompson
Department of Computer Science
Cornell University
Ithaca, NY 14853
laurejt@cs.cornell.edu

David Mimno
Department of Information Science
Cornell University
Ithaca, NY 14853
mimno@cornell.edu

Abstract

Most previous work in unsupervised semantic modeling in the presence of metadata has assumed that our goal is to make latent dimensions *more* correlated with metadata, but in practice the exact opposite is often true. Some users want topic models that highlight differences between, for example, authors, but others seek more subtle connections across authors. We introduce three metrics for identifying topics that are highly correlated with metadata, and demonstrate that this problem affects between 30 and 50% of the topics in models trained on two real-world collections, regardless of the size of the model. We find that we can predict which words cause this phenomenon and that by selectively subsampling these words we dramatically reduce topic-metadata correlation, improve topic stability, and maintain or even improve model quality.

1 Introduction

Unsupervised semantic models are a popular and useful method for inferring low-dimensional representations of large text collections. Examples of such models include latent semantic analysis (Deerwester et al., 1990) and word embeddings (Bengio et al., 2003), but for this work we will focus on statistical topic models (Hofmann, 1999; Blei et al., 2002), which are used to infer word distributions that correspond to recognizable themes. In practice, collections are often constructed by combining documents from multiple sources, which may have distinctive style and vocabulary. This heterogeneity of sources leads to a serious but rarely studied problem: the strongest, most prominent patterns in a collection may simply repeat the known structure of the corpus. Instead of finding informative, cross-cutting themes, models simply repeat the distinctive vocabulary of the individual sources. The model in this case is "correct" in that it has detected the strongest dimensions of variation, but it tells us nothing we did not already know.

As a motivating example, we focus on models trained on novels, where it is known that inferred topics are often simply names of characters and settings (Jockers, 2013). The words *Harry*, *Ron*, and *Hermione* look to the algorithm like the basis of an ideal topic because they occur very frequently together but not in other contexts. But this topic only tells us which books within a larger corpus are part of the *Harry Potter* series; themes like friendship, adolescence, and magic remain hidden. This phenomenon is not limited to fiction: we also include a case study of opinions from US state supreme courts. Unlike examples from fiction, Maine and Utah both exist in the same universe, but exhibit specific regional term use.

We begin by demonstrating that the problem of overly source-specific topics is both substantial and measurable. We present three metrics that provide related but distinct views of source specificity. These metrics are orthogonal to existing metrics of topic semantic quality: uselessly source-specific topics are often still highly coherent and meaningful. These metrics are also inversely related to commonly-used document classification evaluations. Learning 20 newsgroup-specific topics from 20 Newsgroups may be informative as an evaluation, but in practice users are rarely unaware of such structure.

Finally, we present a simple but effective method for reducing the prevalence of source-specific topics. This method relies on probabilistically subsampling words that correlate with known source metadata, and is related to subsampling methods that have been highly effective in word embeddings (Mikolov et al., 2013; Levy et al., 2015). The best of the proposed methods substantially reduces source-specific topics, increases topic differentiation without increasing model complexity, and improves topic stability.

2 Related Work

The common assumption of prior work on metadata-aware topic modeling has been that metadata provides valuable hints that can be used to improve topics. Several methods use document metadata to influence document-level topic distributions. The author-topic model (Rosen-Zvi et al., 2004), relational topic model (Chang and Blei, 2009), and labeled LDA (Ramage et al., 2009) extend LDA by directly incorporating a particular type of metadata (e.g. author information, document links, user-generated tags) into the model. Others, like factorial LDA (Paul and Dredze, 2012), Dirichlet-multinomial regression topic models (Mimno and McCallum, 2008), and structural topic models (Roberts et al., 2014) incorporate more general categories of metadata. All of these aim to *increase* dependence between topics and metadata. In contrast, our goal is to make topics *independent* of specified metadata.

Other research makes topic-word distributions sensitive to document-level metadata. The special words with background model (Chemudugunta et al., 2006) incorporates document-specific word distributions into LDA, while cross-collection LDA (Paul, 2009) incorporates collection level word distributions. The topic-aspect model (Paul and Girju, 2010) extends LDA to include a mixture of aspects of documents such that aspects affect all topics similarly. Although these models may be able to sequester author-specific words, there is no reason to expect that those words will not also drag along general, cross-cutting words.

In this paper we focus on ways to explicitly identify words that bias topics towards a specific metadata tag and modify the input corpus for an algorithm to reduce their effect. Researchers have often dismissed this sort of data curation as unprincipled and heuristic “preprocessing.” More recent work (Denny and Spirling, 2016; Boyd-Graber et al., 2014) emphasizes that *meta-algorithms* for data preparation can greatly affect the intrinsic model quality and human interpretability of topic models.

3 Collections and Models

We collected two real-world corpora that combine text from multiple distinct sources: science fiction novels and U.S. state supreme court opinions.[1]

Science Fiction (SCI-FI). We selected 1206 science fiction novels by 245 authors based on award nominations and curated book lists hosted on Worlds Without End.[2] We consider each author as a source, and treat collaborations as distinct sources. We augmented the corpus with other established authors to increase the diversity of author gender and ethnicity. The novels span from the early 1800s to the present day. Most of these works are currently protected by copyright, so rather than full text we obtained page-level word frequency statistics from the HathiTrust Research Center’s Extracted Features Dataset (Capitanu et al., 2016). This data indicates, for example, that page 227 of *Dune* contains one instance of the word *storm* as a noun. Following previous work (Jockers, 2013) we divide volume-length works into page-level segments, omitting headers and footers.

Corpus	Authors	Docs	Types	Avg Len
SCI-FI	245	327K	132K	153
COURTS	50	52K	89K	1039

Table 1: Corpus statistics for the number of authors, documents, and word types, as well as average document length. Document and word type counts are listed in thousands (K).

U.S. State Supreme Courts (COURTS). Each U.S. state has a supreme court that decides appeals for decisions made by lower state courts. In this collection each document is a court opinion, written by the court after the completion of a case, summarizes the case and judgment. We treat each state court as a source, expecting that courts use geographically specific language (e.g. Colorado, Denver, Colo., Boulder)

[1] Code and data is available at https://github.com/laurejt/authorless-tms.

[2] https://www.worldswithoutend.com/lists.asp

that is not relevant to the legal content of opinions. We examine court opinions for all 50 state supreme courts for cases filed from 2012 through 2016.[3]

Data Preparation. We apply the same initial treatment to both corpora. Tokens are three or more letter characters with possible internal punctuation (excluding em- and en-dashes). Words are lower-cased. To deal with globally frequent terms, we remove words used by more than 25% of documents in a corpus. To reduce the computational burden of a large vocabulary, we remove words occurring in fewer than five documents. We remove all documents with fewer than 20 tokens. This process removes 706 pages and 9192 court opinions from our starting science fiction and state courts corpora.

We train LDA models using Mallet (McCallum, 2002) with hyperparameter optimization occurring every 20 intervals after the first 50. We set the number of topics to be on the same order as the number of sources, so for SCI-FI we use $K \in [125, 250, 375]$ and for COURTS we use $K \in [25, 50, 75]$.

4 Evaluating Topic-Author Correlation

We introduce three ways to measure the source-specificity of topics. For concreteness we will use the terms "source" and "author" interchangeably, but a document's source could be any categorical variable. We want to identify topics that are used by relatively few authors, and more specifically topics whose "meaning" is unduly influenced by the contributions of relatively few authors.

Given a collection of D documents written by A authors such that each document d is written by a single author a, we train an LDA topic model with K topics. Then for each word token i in document d we have both a word type w_{id} and a posterior distribution over its token-level topic assignment z_{di}. For clarity of presentation we can assume a single topic assignment for each token and view the corpus as a data table with three columns: word type w, topic z, and author a. By summing over rows of this table we can define marginal count variables for authors $N(a)$ and topics $N(k)$ as well as joint count variables for the count of a word in a topic $N(w, k)$, a topic in an author $N(k, a)$, and a word in a topic in an author $N(w, k, a)$. A maximum likelihood estimate of the probability of word w given topic k is $P(w \mid k) = \frac{N(w,k)}{N(k)}$.[4]

We note that these statistics must be defined at the token level. As in Mimno and Blei (2011) we are looking for violations of the assumption that $\Pr(w \mid k) = \Pr(w \mid d, k)$. Gibbs sampling algorithms typically preserve token-level information in the form of sampling states, but EM-based algorithms often preserve only document-topic distributions θ_d and topic-word distributions ϕ_k. We can estimate the posterior distribution over topic assignments for each token in document d with word type w as $\Pr(z \mid d, k) \propto \sum_k \phi_k(w)\theta_d(k)$, and generate sparse representations by sampling from this distribution.

Author Entropy. We begin by measuring a topic's author diversity—how evenly its tokens are spread across authors—using the conditional entropy of authors given a topic (Eq. 1). Topics whose tokens are largely concentrated within a few authors will have low entropy, while topics more evenly spread across many authors will have high entropy. With asymmetric hyperparameter optimization we find that the most frequent topics (large α_k) have high author entropy, but topics with high author entropy can have a wide range of frequencies: topics can be both rare and well-distributed.

$$H(A \mid k) = \sum_a \Pr(a \mid k) \log_2 \Pr(a \mid k) = \sum_a \frac{N(a,k)}{N(k)} \log_2 \frac{N(a,k)}{N(k)} \tag{1}$$

While author entropy provides a general sense of author diversity, it does not take into account the expression of topics by authors. Content-based evaluation is especially important because many collections are not well balanced across authors. The fact that a topic is not balanced across authors does not necessarily imply that it is problematic. A novel about the voyages of a ship captain may contain a large proportion of words about sea travel and ships, while a novel that contains one minor character who is a ship captain may contain a small proportion of the same language, used in the same way. We therefore

[3] https://www.courtlistener.com

[4] We do not use Dirichlet smoothing for the purposes of this work for simplicity and to make more reliable comparisons across varying vocabulary sizes. Results using smoothing are similar.

need to be able to distinguish two cases: first, a topic that is consistent across authors but that is used at different rates by different authors, and second, a topic that is not only used at different rates but has different contents across authors. In the first case we can accurately use a topic to "stand for" a particular concept of interest, while in the second case we would get a false impression of the contents of documents, because the expression of the topic in the minority authors differs from the topic as a whole.

To differentiate expected author imbalance from pathological cases, we calculate Jensen-Shannon divergence between a topic's word distribution as estimated from the full collection $\Pr(w|k)$ and two distributions that have been transformed to reduce the influence of the most prominent authors. If the topic has low author correlation then there will be little divergence between the original distribution and its transformation. This method mimics a technique for identifying "junk" topics by AlSumait et al. (2009).

Minus Major Author. The first transformed distribution M (Eq. 2) recalculates the probability of words based on all documents *except* those written by the majority author. If a topic is consistent across authors then the presence or absence of its largest author contribution (labeled a_{major}) should have little effect on the topic's word distribution. The larger the resulting divergence, the more influence the major author has over the topic. Unlike author entropy, this technique does not inherently favor balanced distributions of authors; a very author-imbalanced (low entropy) topic can still have a low minus major author divergence if the dominating author's contribution agrees with the remaining topic tokens.

$$\Pr(w \mid M_k) = \Pr(w \mid \neg a_{major}, k) = \frac{N(w,k) - N(w, a_{major}, k)}{N(k) - N(a_{major}, k)} \tag{2}$$

Balanced Authors. The second transformed distribution B (Eq. 3) treats the contribution of each author equally, no matter how many words in that topic the author produces. The minus-major metric is most sensitive to the case where a single author dominates a topic, but does not handle the case where a small group of authors dominates. Using the balanced transformation we measure the similarity of each author contribution. The larger the resulting divergence between the original and transformed word distributions, the larger the variance in contributing author token usage.

$$\Pr(w \mid B_k) \propto \sum_a \Pr(w \mid k, a) = \sum_a \frac{N(w,k,a)}{N(k,a)} \tag{3}$$

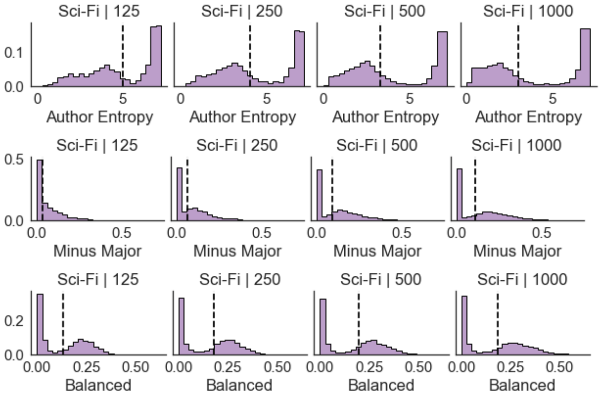

Figure 1: Author entropy, minus major author divergence, and balanced author divergence for topics in topic models trained on SCI-FI. Dashed lines indicate medians. Increasing the number of topics in a model does not reduce the proportion of author-specific topics.

We check the validity of our metrics by evaluating topic models trained on SCI-FI for a wide range of topic sizes (125–1000). As seen in Figure 1, all three measures produce bimodal distributions for all topic sizes, combining highly author-specific topics and more general cross-cutting ones. The proportion of cross-cutting topics remains fairly constant across topic sizes: for all of these models, over 50% of topics fall in the source-specific range. We emphasize that source-specific topics are not necessarily "bad". If the structure of the corpus were not known, these topics would provide a highly useful and coherent insight into that structure. But if, as is typical, the structure *is* known, more than half of the statistical capacity of these models is wasted learning distributions that simply reiterate known structure, regardless of the number of topics.

While all three measurements produce similarly shaped distributions, they do not always agree in detail. Table 2 shows example topics that provide intuition for these differences. At the extremes, Topic A is a general, cross-cutting topic while Topic G is dramatically author-specific. While all three metrics score well for Topics A and B, in Topic B the word

Topic	Entropy	Minus Major	Balanced	Top Words
A	6.79	0.00067	0.017	school professor work university years research science students student college study class year history scientific theory young new field physics
B	6.67	0.0047	0.032	doctor paul hospital nurse patient medical patients doctors room ward bed drugs treatment clinic drug case mental sick therapy medicine
C	5.44	0.043	0.17	jack emma **malenfant** trip janet michael ing **wireman leonard nemoto** sally **jeannine reynolds render manekato mccann** runners thi **joshua**
D	5.31	0.027	0.13	sand **pirx** mars desert roger dust rock **bass** dunes crater martian **jeffries kirov** dune **sweeney eileen** rocks canyon lava camp
E	3.42	0.080	0.16	robot robots andrew human **cully** susan **calvin** brain being **powell donovan** law **moldaug** sir **drake positronic bogert lanning** humans three
F	2.32	0.067	0.083	old night yes cried town last men **rocket** god years hands house upon stood wind boy shut door let dark
G	0.28	0.35	0.32	**f'lar lessa weyr robinton hold dragon f'nor lord dragons benden rider bronze harper thread mnementh brekke ramoth fax fort queen**

Table 2: Topics from a 250-topic model trained on Sci-Fi and their corresponding measures of author entropy, minus major author, and balanced authors. Underlined values indicate poor quality scores and bolded terms indicate word types with low (< 1) author entropy within the topic.

paul seems out of place, but it is common enough in several authors that its word-level author entropy is not low. Topics E and G both score poorly in all three metrics, and both are highly specific to single authors (Isaac Asimov and Anne McCaffrey). But while G is clearly and exclusively names and settings, E contains the common terms *robot*, *robots*, and *human*, and could be confused for a general topic on artificial intelligence.

The metrics are also enlightening when they disagree. Topic C has high author entropy, but only because it mixes highly author-specific words from several different authors. Since each author's contribution differs from the others it scores poorly on the two content-based metrics. Topic D is partially about Mars, but also contains author-specific character names from stories set on Mars. No single author dominates, but the contributions of each author look different. Topic F is so highly correlated with Ray Bradbury that its entropy is low and it looks different when his contribution is removed, but its words are sufficiently general that Bradbury's use of the topic is close to the other authors' (minimal) use.

5 Contextual Probabilistic Subsampling

In this section we present interventions that predict the effect of words and contexts, and modify an input corpus to reduce the number of overly author-specific topics in resulting models. We hypothesize that this problem is due to *burstiness* (Doyle and Elkan, 2009): words that are globally rare, but locally frequent. Dampening the author-specificity of individual word types may reduce their connection to document sources. We therefore evaluate context-specific subsampling prior to modeling, with parameters defined based on tail probabilities of word-specific parametric models.

In selecting this particular approach we follow three design principles that we believe maximize use in actual practice. First, we want interventions to be minimal and have the least possible disruption to current work processes. We therefore choose to focus on meta-algorithms for data preparation that are compatible with but independent from existing, widely implemented inference algorithms. Second, we want any user-specified parameter choices to be simple and intuitive. Although we find that entropy is a useful diagnostic metric, information theoretic metrics such as mutual information are difficult for non-experts to interpret correctly, and critical values can differ widely across collections and dimensionalities. Third, we want both the choice of interventions and the effects of interventions to be transparent to users. We initially considered methods such as adversarially trained autoencoders, but we find that directly subsampling words is much faster, simpler, and easier to explain.

Identifying Author Specific Terms. The simplest way to find author-specific terms is to find terms unique to an author. The Sci-Fi collection contains an unusual number of author-specific coinages, but words used by many authors can still be highly correlated with a particular author. We therefore estimate parametric distributions for each term and compare author-specific term proportions to this

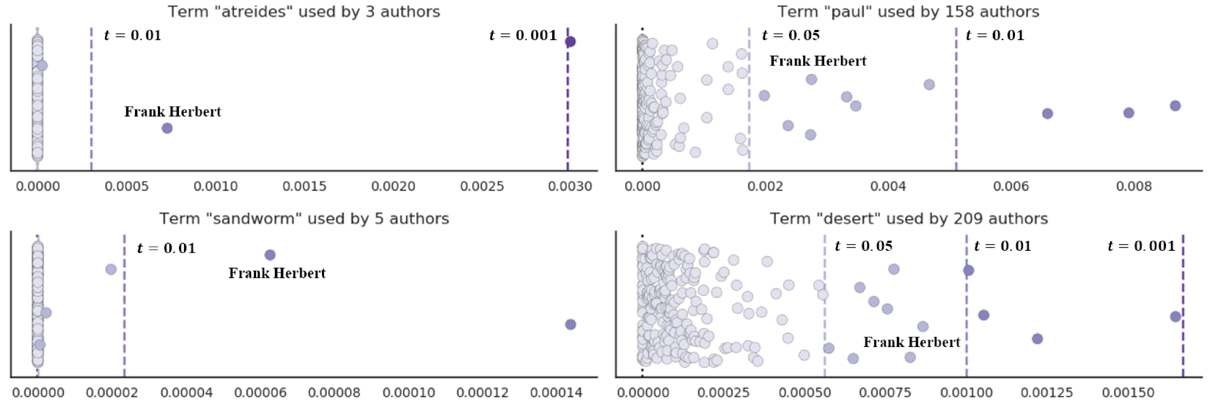

Figure 2: Reasonable threshold values t flag both rare words (left) and common words being used in author-specific ways (right). Each point represents the relative frequency of a term (x-axis) for an author (y-axis) in SCI-FI.

distribution. For each word type w, we calculate the sample mean $\bar{x}_w$ and variance s_w^2 and construct a gamma distribution Γ_w with shape $k = \bar{x}_w^2/s_w^2$ and rate $\theta = s_w^2/\bar{x}_w$. Similar to a significance test, given a user-specified probability threshold t we can define a critical term proportion value under Γ_w

$$\Pr[\Gamma_w \leq f_w^*] \leq 1 - t. \tag{4}$$

A word w is thus considered too specific to an author a if a's usage is too unlikely to occur according to Γ_w. Specifically, this occurs when the frequency $f_{w,a}$ is larger than the cutoff frequency f_w^* defined in Eq. 3. This method satisfies our design goals of simplicity and transparency: the threshold is intuitive and can be adjusted to change how aggressively words are flagged for curation.

Figure 2 shows two character names and nouns from Frank Herbert's *Dune*, where one name and noun are rare and the others are frequent. We see that the rare words *atreides* and *sandworm* are significant to Frank Herbert for $t = 0.01$: there is essentially no "normal" level of use of these words in other authors. Herbert also uses the more common terms *paul* and *desert* more than expected, but to a lesser extreme.

Determining Stop Rates. How we choose to dampen author-specific words is as important as how we detect them. If we globally removed these words using a traditional stoplist, we would lose a substantial portion of the vocabulary. A more sophisticated approach is to construct a stoplist for each author. In this case, words are only removed from contexts in which they are statistically overrepresented. For rare terms, where there is no middle ground between significant use and no use at all, this contextual treatment is effectively the same as a traditional stoplist. But for a word with more widespread use, that word would disappear only from contexts with abnormally high usage.

While this technique avoids erasing the majority of a collection's vocabulary, it leads to a paradoxical situation where a word that is thematically central to a work occurs *less* frequently in that work than in other works. Entirely removing *desert* from Frank Herbert or *robot* from Isaac Asimov would reduce the model's ability to identify relevant themes.

To find a middle ground, we use probabilistic subsampling to reduce outlier author use to something more in line with the collection's overall usage. We use the same threshold t to set subsampling rates. For a word type w and author a the probability of stopping a token of type w in a is

$$\Pr(\text{Stop } w \text{ in } a) = 1 - f_w^*/f_{w,a}. \tag{5}$$

The threshold t specifies when an author's use of a word is too extreme for our model Γ_w. If we reduce these outlier frequencies to their corresponding cutoff frequencies f_w^*, they will be set to the largest below-threshold frequency dictated by Γ_w. We construct our subsampling rates such that in expectation new author frequencies will equal their corresponding threshold frequency from the original distribution.[5]

[5] Iteratively reevaluating Γ_w leads to an unstable "race to the bottom."

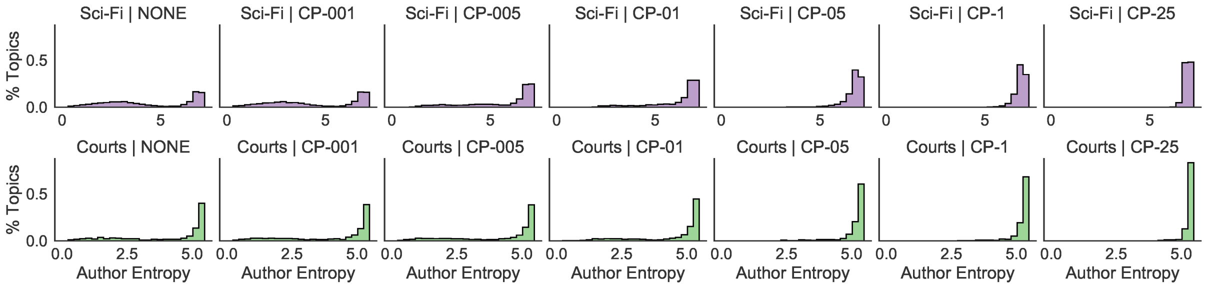

Figure 3: Increasing the threshold t for contextual probabilistic (CP) subsampling results in more topics with high dispersion over authors.

6 Results

Unless otherwise noted, we refer to models with a topic size of 250 for SCI-FI and 50 for COURTS, and set the hyperparameter t of context-based methods to 0.05. We refer to a treatment with no intervention beyond standard stopword removal as **NONE**. We compare these models to three classes of curation methods, each with varying parameters. **AF-[n]** removes all terms that are used by at most n authors. **C-[t]** removes any term from author a's context whose frequency $f_{w,a}$ exceeds significance threshold t with respect to distribution Γ_w. **CP-[t]** subsamples terms according to Eq. 5. We train 10 runs with random initializations for each parameter setting.

Subsampling reduces topic-metadata correlation. We begin by measuring how well the curation techniques reduce the formation of author-correlated topics. We find that while removing words with low author frequency has little effect (not shown), contextual methods greatly reduce the formation of "bad" topics according to all three measures. As expected, the value of the threshold t affects performance of the context-based methods. In Figure 3, we see that lowest values of t are ineffective; $t = 0.001$ is hardly distinguishable from NONE and $t = 0.005$ is on par with low author frequency stoplists. We observe that settings of $t \geq 0.05$ perform very well, and choose this value as a default in our public code release.

Subsampling before inference does more than change the appearance of topics, it changes the content of the inferred topics. To test whether subsampling after inference has the same effect we construct ten additional models by *post hoc* stopping the 250-topic trained models for NONE-treated SCI-FI to match token-for-token the CP-05 curated versions. We find that *post-hoc* removal has little effect on topic-metadata correlation; over twenty percent of topics are dominated by a single author with the worst having 96.4% of tokens contributed by one author.

Semantic quality is preserved. We define author-specificity as a property orthogonal to model quality: there is nothing fundamentally wrong with a topic full of character names outside the context of specific user needs. But ideally in reducing the prevalence of overly author-specific topics we would replace them with equally meaningful ones. We measure semantic quality of topics using Mimno et al. (2011)'s topic coherence metric as reported by Mallet. This metric measures the tendency for the most probable (top) words of a topic to cooccur. A topic k with m top words $w_{k,1}, \ldots, w_{k,m}$ has topic coherence

$$\sum_i \sum_{j<i} \log \frac{D(w_i, w_j)}{D(w_i)} + \beta, \tag{6}$$

where D represents the number of documents containing a word or word pair and β is the LDA hyperparameter for topic-word smoothing. Large negative values indicate that the top words of a topic seldom cooccur, while values close to zero indicate that the top words frequently coccur.

We find that despite substantial changes in topic content, corpus modification has no consistent effect on the semantic quality of topics. In Figure 4, we find that all curation methods except CP-001 have significantly higher mean topic coherence than NONE for SCI-FI. Contextual methods with $t \geq 0.05$ have the highest coherence. For COURTS, topic coherence is maintained across treatments, except for the most aggressive interventions C-05 and C-1.

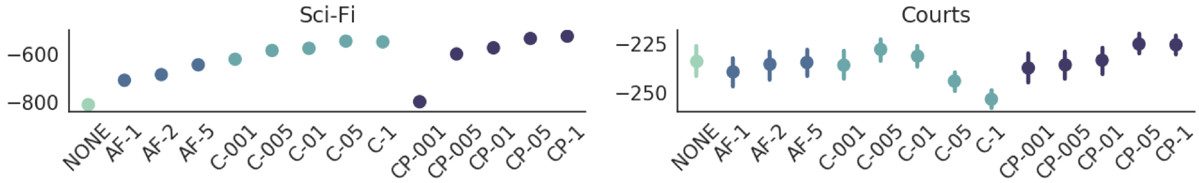

Figure 4: Contextual probabilistic subsampling improves mean topic coherence for SCI-FI despite the removal of frequent words. Coherence degrades under context curation for COURTS.

Corpus damage is reduced. All things being equal, we want to modify the input collection minimally, both in terms of vocabulary and actual document content. Figure 5 confirms that contextual curation has the highest type and token loss across corpora, because it completely removes all instances of a word type in a context. This may partially explain the dramatic loss of model quality for these specific treatments.

Contextual probabilistic subsampling removes more tokens than author frequency cut-offs, but better preserves the vocabulary. For thresholds $t \leq 0.01$, contextual probabilistic subsampling removes fewer word types than any of the author frequency cut-off methods. However, there is less agreement across corpora for $t \geq 0.05$. For SCI-FI, these methods remove more types than AF-5, while the reverse is true for COURTS. This discrepancy might arise from differences in the relative size of collection sources—some authors write more than others, some courts issue more opinions—and vocabulary use.

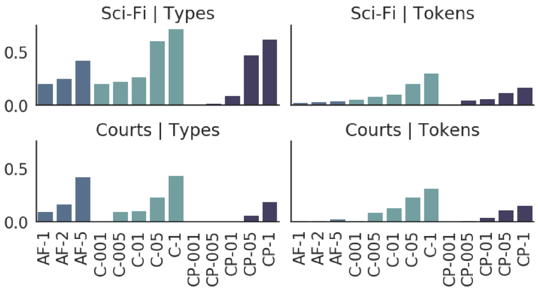

Figure 5: Proportional loss of removed word types and tokens. Contextual probabilistic subsampling does substantially less damage than contextual curation.

Subsampling affects more than names. Character names are the most prominent motivating example for this work, so it is reasonable to ask whether named-entity tagging or even a simple part-of-speech (POS) filter would be sufficient. To check whether we are just removing proper nouns, we compare the frequency of four general POS categories: common nouns, proper nouns, verbs, and adjectives. These make up 37%, 10%, 27%, 13% of all tokens respectively in SCI-FI. Figure 6 shows the proportion of tokens removed from each category for each curation method. Unsurprisingly, proper nouns make up a large proportion in all cases, but contextual methods also remove substantial numbers of tokens across all word groups.

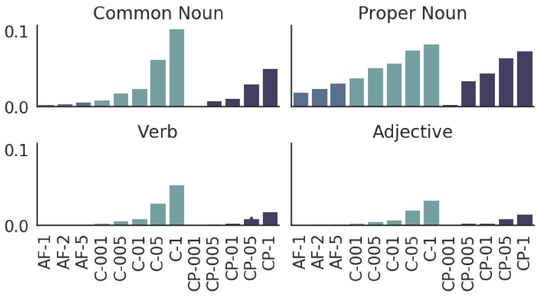

Figure 6: Proportion of SCI-FI tokens removed across part-of-speech groups. Contextual methods remove tokens from all groups.

Subsampling increases stability and specificity. We find that removing author-specific terms using contextual probabilistic subsampling greatly mitigates the formation of author-correlated topics, but what do these models learn instead? Are they augmenting the set of uncorrelated topics found within the untreated models, or are they perhaps identifying entirely new structure? More importantly, what are the characteristics of the newly formed or persisting author-correlated topics? To answer these questions, we perform pairwise comparisons of the topic-word distributions from different models using Jensen-Shannon divergence to find the most likely of topic correspondences. By linking these topics together, we can gain a sense of which topics persist across treatments, which are refined or split, and which are lost entirely. We focus on SCI-FI since it has larger models, but we will highlight similar analysis for COURTS.

Before making any inter-treatment comparisons, we examine topic stability internally within each

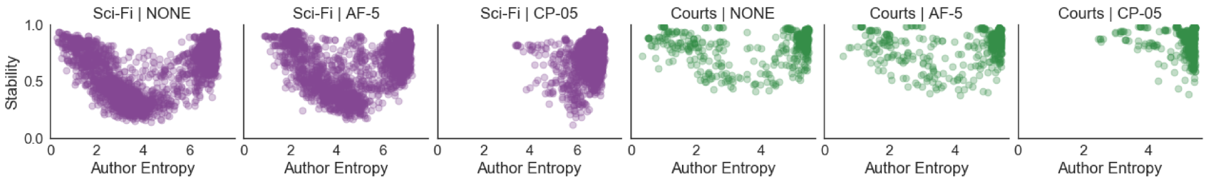

Figure 7: Topic Stability and Entropy for SCI-FI ($K = 250$) and COURTS ($K = 50$). AF-5 has little effect. Many of the low-entropy topics avoided by CP-05 are highly unstable.

treatment. We define stability as the average similarity between a topic and its nearest equivalent from each of the nine other trained models for a treatment. More formally, the stability of topic k_i from the ith instance of a model is

$$\text{Stability}(k_i) = 1 - \frac{1}{9}\sum_{j \neq i} \min_{k_j} JSD(P(w \mid k_i), P(w \mid k_j)) \tag{7}$$

where JSD is Jensen-Shannon divergence. A topic stability close to one implies that a topic persists across runs, while a value close to zero implies that a topic is *ephemeral*—observed once and unlikely to be seen again across random initializations.

High stability does not imply author-specificity. In Figure 7, we see that the most stable topics tend to have either maximal or minimal author entropy, while the most unstable topics have middling values. The unstable topics tend to capture a mixture of disjoint structures as we saw in topics C and D from Table 2. This also occurs (but to a lesser extent) in COURTS with topics containing many distinct regional terms (S1: *s.w oklahoma tenn kan ind n.e indiana app tennessee o.s*) or containing a mixture of a general and state-specific concept (S2: *school wyo miss wyoming mississippi ann education students hill student*). Thus, the most stable topics are the most apparent by being very context specific or most cross-cutting.

Now that we have evaluated the stability of topics under the baseline NONE treatment, we can use minimum divergence to align those topics with topics trained under the CP-05 subsampling treatment. Unstable NONE topics are generally very distant from their nearest CP-05 counterparts. Of our example topics in Table 2, C and D are the most unstable at 0.39 and 0.42 respectively. Topic C diverges heavily (0.87) from its closest CP-05 match, while aspects of D are echoed in its nearest match *sand desert rock mountains mountain dust land surface plain water* (0.53). COURTS topic S2 is also more distantly associated (0.63) with an education/administration topic: *board school commission administrative agency plan department board's education regulations*. Over 95% of NONE topics with high stability and high author entropy are linked to a CP-05 topic with divergence less than 0.5. Topic A has a close match (*professor university college student students research school science work years*) at 0.23. A appears to have become more specific in CP-05 by splitting into two additional topics that echo other aspects, namely teaching children (0.54) and scientific research (0.55).

The case of stable, low entropy NONE topics is harder to interpret. While half of these topics are far from their CP-05 match (> 0.7), 16% have divergences of less than 0.4. Topic G matches well to *lord hold between master queen star enough turns high good* (0.3) which is both very stable and CP-05's lowest author entropy topic (64.1% from Anne McCaffrey).[6] While these topics have not been prevented entirely, they have been largely mitigated.

The topics in CP-05 that are the most dissimilar from topics within NONE demonstrate that this treatment adds differentiation. We find that overall 50% of CP-05 topics have a large divergence (> 0.7) with the NONE topics. Some of these divergent topics consist of names, but these groupings might indicate regional or temporal naming patterns. In other cases, we encounter new and interesting topics such as an authentic robots topic (*machine robot machines robots human mechanical metal brain men built*), which matches to both a general computer topic and example topic E (Asimov). We also find a new topic on magic and witchcraft (*magic ghost demon evil witch demons power spell magician ghosts*) whose

[6] The topic is common words used in specific ways: a *hold* is a fortified settlement, dragons teleport by going *between*.

closest match is a general religion topic *god gods religion world religious ancient temple people faith these*. In fact, the term *witch* never appears as a top-20 term for any topic within the 250-topic NONE models. These topics may appear for NONE when we increase the topic size to $K = 1000$, but at the cost of a much larger model and with no guarantee against intruding character names.

Subsampling produces cross-cutting topics. While our topics score well quantitatively, how humanly interpretable and useful are the resulting topics? Are they actually cross-cutting in nature? We address these questions by more closely examining topics generated by the CP-05 subsampling treatment. We can explore the collection by sorting authors and individual novels within topics.

The highest frequency topics from the *NONE* treatment are largely preserved by *CP-05*. These topics by their nature are very cross-cutting and filled with frequent, general words. Despite this extreme generality they can provide a way to analyze passages representing high-level discourse concepts such as inquiry (*why asked ask answer question want questions should does because*) and the description of events and time (*during such most these course because happened effect period result*).

The mid-frequency topics are more concretely thematic in nature. We find a topic describing empire, politics, and history (*empire world power people war new government history political under*) which is associated with Doris Lessing's *Canopus in Argos* series, Isaac Asimov's *Foundation* series, and Kim Stanley Robinsons's *The Years of Rice and Salt*. In line with the science fiction genre, these novels focus on expansive future and alternative histories. We also find a topic on language (*language words english speak word understand spoke speech languages talk*). The most prominent authors in the topic—Robert A. Heinlein, Robert Silverberg, and Poul Anderson—are among the five most prolific authors in SCI-FI, which suggests the generality of the topic. Notably the most prominent volumes are by none of these authors: *Babel-17* by Samuel R. Delany, *Native Tongue* by Suzette Haden Elgin, and *Changing Planes* by Ursula K. Le Guin. All three include the social and political language as a major plot point. These three works are fundamentally tied confirming that this topic embodies a cross-cutting linguistic theme.

Looking more closely at the lower frequency robots topic (*machine robot machines robots human mechanical metal brain men built*), we find that it is both topically cohesive and cross-cutting. The five most-represented authors all have works heavily related to artificial intelligence: Isaac Asimov, Robert Silverberg, Stanisław Lem, Clifford D. Simak, and Philip K. Dick. The most-represented volumes tell a similar story with *Men and machines* by Robert Silverberg, *The complete robot* by Isaac Asimov, and *The Humanoids* by Jack Williamson holding the top three ranks. Reassuringly, there are well-represented novels by less-represented authors such as *The Starchild Trilogy* by Fredrick Pohl and Jack Williamson. The low frequency of this topic is surprising given the presence in the collection of robot-related novels, especially works by Isaac Asimov. This discrepancy revealed that an Asimov-specific topic (*human being law might must such without may robot beings*) has persisted. Many authors receive a non-negligible token representation, but Asimov's token count is still a factor of ten larger than the second most prominent author (Robert A. Heinlein).

7 Conclusion

We present a formal definition of the problem of overly source-specific topics, three evaluation metrics to measure the degree of source-specificity, and a simple text curation meta-algorithm that dramatically reduces the number of source-specific topics. This approach has immediate practical application for the many collections that combine multiple distinct sources, but it also has important theoretical implications.

We view this work as a preliminary step towards predictive theories of latent semantics, beyond purely descriptive models. Despite ample practical evidence that interventions such as stoplist curation can have significant effects, most previous work has focused on algorithms for identifying a single "optimal" low-dimensional semantic representation. Our results indicate that there are potentially many interventions in text collections that each have distinct but predictable effects on the results of algorithms. Just as biologists use multiple stains to view different aspects of microorganisms using the same microscope, users of text mining algorithms should be able to choose multiple distinct text treatments, each with its own predictable effects, to meet distinct user needs.

Acknowledgements

We would like to thank our reviewers for their helpful feedback and suggestions. We would like to thank Ishaan Jhaveri and Ricardo A. Wilson II for directing our attention to science fiction, as well as Jack Hessel and Alexandra Schofield for their reviews and feedback. The work was supported by NSF #1526155, #1652536, and the Alfred P. Sloan Foundation; and through resources provided by the Hathi Trust Digital Library and the Hathi Trust Research Center.

References

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of LDA generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 67–82.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2002. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 14, pages 601–608.

JL Boyd-Graber, DM Mimno, and D Newman. 2014. Care and feeding of topic models. *Handbook of Mixed Membership Models and Their Applications*, pages 225–254.

Boris Capitanu, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, and J. Stephen Downie. 2016. The HathiTrust Research Center extracted feature dataset (1.0) [Dataset]. http://dx.doi.org/10.13012/J8X63JT3.

Jonathan Chang and David M Blei. 2009. Relational topic models for document networks. In *AIStats*, volume 9, pages 81–88.

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in Neural Information Processing Systems*, volume 19, pages 241–248.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391.

Matthew James Denny and Arthur Spirling. 2016. Assessing the consequences of text preprocessing decisions.

Gabriel Doyle and Charles Elkan. 2009. Accounting for burstiness in topic models. In *International Conference on Machine Learning*, volume 26, pages 281–288.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence*, volume 15, pages 289–296.

Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119.

David Mimno and David Blei. 2011. Bayesian checking for topic models. In *Empirical Methods in Natural Language Processing*, pages 227–237.

David Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, volume 24, pages 411–418.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Empirical Methods in Natural Language Processing*, pages 262–272.

Michael Paul and Mark Dredze. 2012. Factorial LDA: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems*, volume 25, pages 2582–2590.

Michael Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI Conference on Artificial Intelligence*, volume 24, pages 545–550.

Michael Paul. 2009. Cross-collection topic models: Automatically comparing and contrasting text. Master's thesis, University of Illinois Urbana-Champaign.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Empirical Methods in Natural Language Processing*, pages 248–256.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Uncertainty in Artificial Intelligence*, volume 20, pages 487–494.