Quantifying the Effects of Text Duplication on Semantic Models

Alexandra Schofield[1] **Laure Thompson**[1] **David Mimno**[2]
1 Department of Computer Science, Cornell University, Ithaca, NY
{xanda, laurejt}@cs.cornell.edu
2 Department of Information Science, Cornell University, Ithaca, NY
mimno@cornell.edu

Abstract

Duplicate documents are a pervasive problem in text datasets and can have a strong effect on unsupervised models. Methods to remove duplicate texts are typically heuristic or very expensive, so it is vital to know when and why they are needed. We measure the sensitivity of two latent semantic methods to the presence of different levels of document repetition. By artificially creating different forms of duplicate text we confirm several hypotheses about how repeated text impacts models. While a small amount of duplication is tolerable, substantial over-representation of subsets of the text may overwhelm meaningful topical patterns.

1 Introduction

Different discussions of the same subject tend to use similar words. Unsupervised models such as latent semantic analysis (LSA) (Deerwester et al., 1990) and latent Dirichlet allocation (LDA) (Blei et al., 2003) look for these statistical signatures of topicality in the form of repeated word co-occurrences. These methods have become increasingly popular because they are powerful and easy to apply to large unlabeled datasets. The apparent ease-of-use of LSA and LDA, however, makes it easy to overlook potential problems in text corpora. In this work, we focus on measuring the impact of one such issue: duplicate text.

Latent semantic methods look for patterns of repetition. But when text is repeated exactly, statistical methods that look for patterns may be diverted from more meaningful semantic groups: verbatim repetition looks, to the algorithm, more topical than actual topics. If not accounted for, repeated text can change measures of fitness to over-value fit on repeated texts, or even "leak" held out data that is duplicated in the training data. At best, duplication may cause us to overestimate the expressiveness and reliability of models. At worst, models skewed by text duplication may invalidate any conclusions drawn from them, and, by extension, the method itself.

Text replication is a persistent and difficult problem in natural language corpora. In social media settings, partial duplication due to quotation and threading is ubiquitous. Of the 20k posts in the 20 Newsgroups corpus (Lang, 1995), 1151 are exact duplicates, and 25% of the remaining tokens are quoted text from other newsgroup messages.[1] In literary corpora, different versions of the same document may also conflict: text files for Hamlet may differ slightly due to publisher information, line numbers, editorial changes between Shakespeare's folios, and footnotes. Removing exactly identical duplicates of texts is possible through direct lexicographic matching, but for lexical near-duplicates and partial textual overlap, we may need more careful heuristics to detect duplicates, forcing researchers to make judgments about what text to remove and what to keep.

Evaluating what level of duplication is "safe" can therefore not only reduce the risk of false conclusions but also save great amounts of work spent identifying and removing duplication. In this work, we investigate the effect of text duplication on LSA and LDA by experimentally amplifying the magnitude of text duplication in a variety of corpora. We look both at how models shift to over-represent repeated text and how that shift affects the model representation of documents without repetition. To account for the variety of types of duplication, we look at exact du-

[1] Computed using scikit-learn's 20 Newsgroups API: http://scikit-learn.org/stable/datasets/twenty_newsgroups.html

plication of whole documents as well as repetition of a text segment across many documents. Finally, we recommend what aspects of text deduplication one should focus on to successfully reduce negative effects, with different suggestions depending on the chosen model. [2]

2 Previous Work

Text duplication and reuse is a well-established problem in textual corpora. The web is filled with pages of near-duplicate content (Broder et al., 1997; Manku et al., 2007), journalistic reuse is common practice with the dissemination of information from news agencies to newspapers (Clough et al., 2002; Smith et al., 2013), and plagiarism is prevalent in student submissions (Clough et al., 2003). However, past work has focused on the *identification* of reuse instead of the *effects* that duplication has on semantic models.

The detection of text reuse relies on the ability to measure similarity between documents or passages. In general, these techniques measure the similarity of textual content, though other similarity metrics for reuse identification have been proposed (Bär et al., 2012). These measures can fall into two general groups: global and local. Global techniques measure the similarities of entire texts. These techniques are especially used for near-duplicate detection. A common approach of this form is fingerprinting (Potthast and Stein, 2008). This method involves transforming a document into a smaller representation (e.g. a set of n-grams) to measure similarity cheaply. Local techniques measure similarity at a finer granularity (e.g. paragraphs or sentences). In this setting, reuse may be mixed with text derived from other sources. These techniques often involve two steps: one aligning texts with some method (Lee, 2007; Smith et al., 2013)m and one scoring similarity of aligned sequences, e.g. based on cosine similarity of the bag-of-words vector. All of these techniques require choices of hyperparameters such as similarity threshold and n-gram size that affect what the technique considers duplicate text. Our work focuses on understanding what types of document deduplication are important so that practitioners can make better-informed choices about how to calibrate these models.

[2]Code for our experiments can be found at https://github.com/heraldicsandfox/semantic-text-duplication.

While it is possible to evaluate semantic models as features in a downstream supervised task, they are harder to evaluate intrinsically as unsupervised models of data exploration. For LDA, it is standard to consider held-out likelihood of a test set as a measure of model fit (Wallach et al., 2009b). One can also use human evaluations to judge the interpretability of topic summaries (Chang et al., 2009), though this measure can also be approximated with automated evaluations based on corpus statistics (Aletras and Stevenson, 2013; Lau et al., 2014; Mimno et al., 2011). One can also evaluate individual topics based on how much they diverge from corpus-wide distributional expectations (Al-Sumait et al., 2009).

Because LSA does not yield semantically meaningful dimensions, intrinsic approaches to evaluation are focused on the spatial aspects of the model's word embedding into the real domain. Word similarity tasks are perhaps the most common evaluation, which compare human "gold standard" judgments of word pair similarity to distances between the corresponding word vectors (Finkelstein et al., 2001; Bruni et al., 2012; Hill et al., 2016). However, the vagueness of definitions of "similarity" and the contextual dependency of similarity have cast doubt on these as gold standards of evaluation (Faruqui et al., 2016). Solving word analogies using vector arithmetic is also sometimes used to evaluate neural word embeddings, but LSA does not tend to produce this structure well (Pennington et al., 2014; Mikolov et al., 2013).

3 Theorized Impact

The fundamental problem with repeated text in a distributional semantic model is the over-representation of specific word co-occurrences to a model. To understand this, we consider the matrix factorization representation of these models. Borrowing notation from Arora et al. (2013), we consider a corpus with M documents and vocabulary size V over which we want to learn a K-dimensional representation of each document and vocabulary term. We can build an $M \times V$ matrix C to represent our corpus, where C_{di} is a function of the frequency of term i in document d. Both LSA and LDA represent factorizations of this matrix into two rank-K matrices, $C = WA^T$, where W is an $M \times K$ matrix and A is a $V \times K$ matrix. In the case of LSA, we apply tf-idf weighting

to C before producing a truncated singular value decomposition $C = U\Sigma A^T$, with Σ a diagonal matrix of dimension $K \times K$, and U and A column-orthogonal matrices. We can reduce this to the factorization above by multiplying Σ with one of the two outer matrix factors, e.g. $W = U\Sigma$. LDA performs a non-negative matrix factorization on a smoothed stochastic version of C, producing row-stochastic matrices W and A.

Duplicate text implies that more rows in C will contain a particular signature of word frequencies. This implies that a low-rank matrix factorization will increasingly devote representative power to this particular textual signature in order to minimize loss in its representation. We expect to observe two principal effects:

- As text is repeated more, to optimize model fit on the data, one or more topics/dimensions will converge to model the repeated text.
- Text that is not exactly or near-exactly repeated (or *singular* text) will be modeled less effectively both in terms of model fit and interpretability.

These effects are based on the incentive of the model to overfit repeated text: topics and dimensions modeling solely the repeated text will leave less representational power for the remaining text, and combinations of repeated and singular text will likely yield less coherent topics.

4 Evaluation Methods

We quantitatively examine several aspects of models with varying forms and degrees of duplication to determine the magnitude of the change produced by repeated text. It is important to note that our goal is simply to measure the difference between models, and not to make normative statements about the *quality* of topics. Indeed, many measures of topic quality such as word intrusion (Chang et al., 2009) and word co-occurrence (Newman et al., 2010; Mimno et al., 2011; Lau et al., 2014) may improve as a result of degenerate, single-document topics: most documents are internally coherent, so a single document's word distribution may appear to be a sensible topic.

Loss The first aspect is model loss. As stated in Section 3, as a segment of text is repeated more, we anticipate that the fit over documents containing repeated text will improve, while the fit over documents not containing repeated text will worsen. To evaluate this for LSA, we examine the Frobenius norm of the difference between the reconstruction WA^T and C for the rows corresponding to documents with and without repeated text. For LDA, we estimate the perplexity of both the training data and held-out data without repetitions from the same corpus.

Concentration Secondly, we examine component (e.g. topic/dimension) concentration. Repetition of a document amplifies the co-occurrence between the terms contained in the document. As this signal grows stronger, we expect models to begin "memorizing" these words. We anticipate that affected models will develop a simpler latent representation for the repeated document, one concentrated over a small number of components. For example, if a model is devoting topic k to a repeated document, then instances of that document should have a high proportion of topic k. Concentration measurements relate to loss, but focuses specifically on the document-component or document-topic patterns, while loss also includes information about the topic-word dynamics.

The effect of components converging to a single piece of repeated text should be easily observed by examining how close topics are to the unigram language model induced by the repeated text. If we repeat multiple documents independently, however, we may also expect to see distinct components correlated with disjoint subsets of the repeated texts. To account for this, we evaluate component concentration separately for documents with repeated text and without repeated text.

For LDA, we examine the *entropy* of document vectors. Information entropy represents the expectation of the representation length of a given outcome as a function of the probability distribution over outcomes:

$$E_d = \sum_k \theta_{dk} \log \theta_{dk}$$

where θ_{dk} is the probability of a token generated in document d having topic k. Entropy is inverse to concentration: the entropy of text should lower as the text is repeated more, as all of their topical mass would be concentrated in topics converging to modeling duplicate texts. Conversely, documents not containing repeated text may also have their entropy increase as text repetition increases, as topics will less adequately fit to the behavior of the singular documents.

In LSA, entropy is not as directly applicable: vectors in $W = U\Sigma$ can be arbitrarily real-numbered. However, we still want to access a similar basic concept, the amount a vector representation of a document is concentrated in a few dimensions. So, we examine the *absolute dispersion* of each row vector d in W:

$$D_d = \sum_k \frac{|d_k|}{\|d\|_1} \log \frac{|d_k|}{\|d\|_1}$$

where d_k is the kth component of d. Absolute dispersion measures the entropy of the L1-normalized masses of the vectors in W.

Expressivity The final aspect is expressivity of topics. If one topic converges to the unigram language model of repeated documents, the resulting model has effectively lost one topic worth of expressive power by focusing on overly-specific themes. Someone looking to learn generalized semantic corpus patterns from a topic model will therefore have one fewer topic of interest available. The frequency of terms in the repeated text may also overwhelm the most probable terms in many of the topics, again reducing the ability to interpret these topics or to understand their content through a summarized representation. While expressivity in the form of topic summaries makes little sense for LSA, using LDA models, we may examine topic summaries, obtained as the top s most probable terms of a topic where s is a fixed parameter. We may select the same number of terms s from the most probable in a unigram language model of the repeated text, and determine what proportion of the tokens obtained from concatenating topic similarities are the top terms of the repeated text language model.

5 Experimental Setup

Data We use two corpora: a sample of articles from the New York Times Annotated Corpus (Sandhaus, 2008) and a collection of Reuters newswires from the Spanish Language News Text Corpus (REUSL) (Graff, 1995). We choose news corpora because they provide well-curated text with repeated subjects but few exact document-level duplicates, though quotes and templated text may still cause text duplication. We can use these as a testbed for general duplication behaviors we see across a variety of corpora. Text is lower-cased and tokenized to only include tokens of three or more characters, allowing for contractions or hyphenations as single tokens. New York Times articles average 494.5 words in length, while Reuters newswires average 201.5 words.

To ensure our experiments are the only cause of exact duplication of text in our corpora, we use strict methods of text deduplication. When two or more documents have more than 70% unigram overlap, we remove all but the longest document. In addition, we delete 7-grams that appear in more than 10 documents based upon existing thresholds for plagiarism detection (Citron and Ginsparg, 2015). To account for stopwords, we remove all terms appearing in more than 80% of documents. Finally, we remove documents with fewer than 7 tokens after processing. We perform this process on a random sample of 30,000 documents from each corpus to ensure we may obtain a sample of 25,000 curated documents for each of our two corpora. We also produce 10% samples of these corpora, containing 2,500 documents each, to measure the effect of corpus size.

Text Duplication Treatments We use our deduplicated news corpora to construct datasets with artificial text duplication. We examine two different duplication scenarios: exact document duplication and template string duplication.

In *exact document duplication*, we randomly sample $p\%$ of the documents in the dataset and include c copies of each sampled document in our final corpus along with one copy each of the remaining documents, which we refer to as *singular* documents. To test the extremes of this effect, we also perform *single document* tests for large c with only one repeated document. From these synthetically duplicative corpora, we can determine whether effects are triggered by the sheer volume of duplicated text or if they are influenced by the diversity of the copied documents.

In *template string duplication*, rather than duplicating the sampled $p\%$ of documents, we prepend a fixed string to each document in the $p\%$ sample, producing what we refer to as *templated* documents or texts. As repeated text may be lexically similar or different from the non-repeated text of the corpus, we consider two different types of prepended string. The first is a randomly-sampled document from the deduplicated corpus but not included in the training set (*Sampled Template*), simulating repeated text that is lexically similar to the document content. The second is

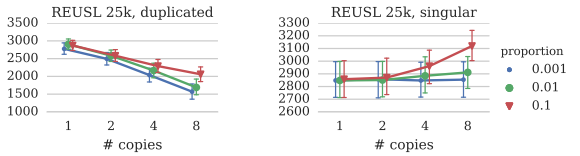

Figure 1: Training perplexity with LDA models trained on the REUSL 25k corpus with 80 topics. Perplexity decreases significantly for the duplicated documents with repetition, but the effect on singular documents is negligible with repeated proportion of the corpus smaller than 0.1.

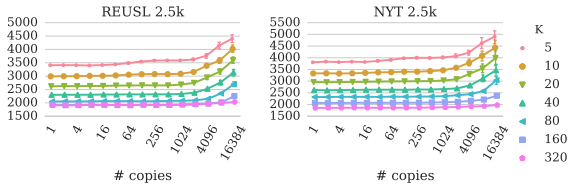

Figure 2: Perplexity from duplicating a single document remains largely unaffected for singular documents until the number of repetitions is $c = 4096$, when duplicate texts outnumber singular texts. There is also a subtle inflection point for smaller numbers of topics K at $c = 256$, approximately 1/10th of the corpus, but this effect is not visible with more topics.

the first 100 words of the classic Lorem Ipsum filler text (*Lorem Template*), simulating repeated text with little lexical overlap with the documents. Because we are investigating bag-of-words models, we do not worry about grammatical errors in the nearly-duplicated text, so the segmentation of this repeated prefix should not be a concern.

Training We analyze two types of semantic models: LSA and LDA. LSA models are trained using tf-idf weighting on word-document matrices using custom Python code.[3] LDA models are trained using Mallet (McCallum, 2002) with fixed hyperparameters $\alpha = 50/K$ and $\beta = 0.01$ for ease of comparison. To compute perplexity, we use log likelihood estimates from Mallet's built-in left-to-right estimation (Wallach et al., 2009a).

6 Results

Because of the exponential combination of different experimental settings available, it would be unfeasible to examine all our metrics for all data. Instead, we focus our analysis on specific examples

[3]Code uses `scipy`, `numpy`, and `scikit-learn`.

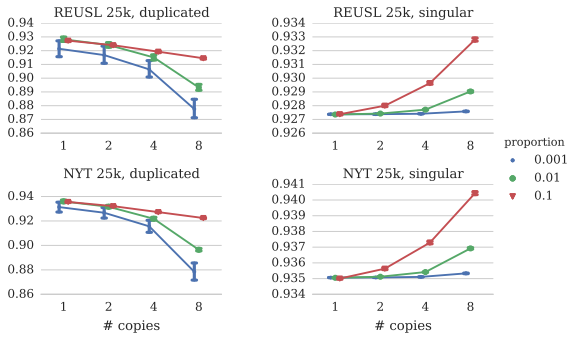

Figure 3: Model loss for LSA models with 80 components. Loss for duplicated documents decreases as the number of repetitions c increases. The frequency of replication affects loss at a much smaller scale for singular documents.

that we believe demonstrate the effects seen in the rest of the corpus. We use smaller sets of 2.5k documents for examining the effect of heavy duplication and sets of 25k documents otherwise.

6.1 Loss

We begin with the case of exact document duplication. In Figure 1, the perplexity of LDA decreases substantially as documents are duplicated. This reduction is due to better fit to the duplicated documents. As fit improves in duplicated documents, however, we do not see a meaningfully worse fit for singular documents. These documents increase in perplexity, but the increase is not significant at low levels of duplication, such as when $c = 2$ or $p = 0.001$. In the single document case in Figure 2, this effect is emphasized: likelihood on singular documents remains level even with heavy repetition in short corpora. The sheer volume of duplicated text does not by itself damage model fit, likely because the duplicated text can be easily modeled by a single topic.

This effect is not solely due to LDA's specific probabilistic model. We see a similar pattern in LSA. In Figure 3, we see that loss for duplicated documents decreases as duplication increases. However, the amount of decay depends on the proportion of the corpus replicated: the smaller the proportion size, the more dramatic the decay. In contrast, the loss for singular documents increases only slightly with more copies, though more for higher proportions of duplication.

To gain a better understanding of how duplication affects LSA, we look at the effects of repeating a single document an extreme number of

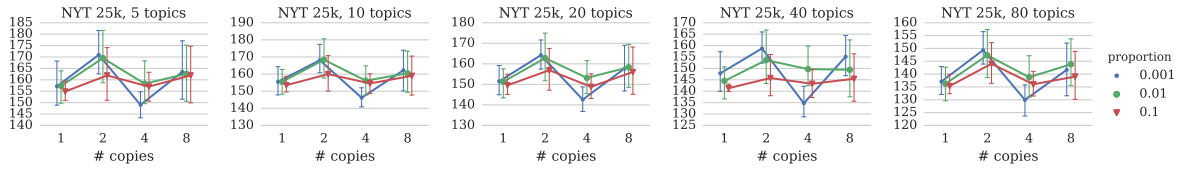

Figure 4: Held-out data perplexity (in thousands) for different the NYT 25k corpus with varying numbers of topics K. Increasing the proportion of repetition for exact duplicate documents does not increase test perplexity. With repeated corpus proportion $p = 0.001$, however, repeating documents exactly 4 times (but not 2 or 8 times) significantly improves perplexity, potentially because it induces a new topic to model it. Held-out data contained no repeated documents.

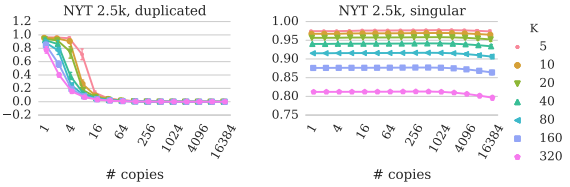

Figure 5: LSA model loss for the NYT 2.5k corpus with one duplicated document. Model loss for singular documents is again unaffected by repetition, while loss for the duplicated document quickly falls to zero as repetition increases. The fall in loss signals the start of model "memorization."

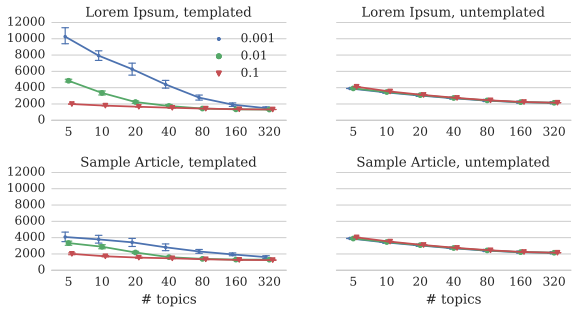

Figure 6: LDA training perplexity for REUSL 2.5k with different types of templated text repetition. The effect of duplication is prominent for small numbers of topics but diminishes with more topics to sufficiently model the missing text. With the fraction of the corpus that contains duplicates $p = 0.1$, the perplexity of template documents is below that of untemplated texts.

times. From Figure 5, we see that the loss of singular documents does not meaningfully change as the number of copies increases. We can observe the steep decline in loss for duplicated documents as the signal of when the top K components begin to "memorize" the duplicated document. The more components K, the fewer repetitions need for the overfitting to begin.

In the LDA case, we may also look at held out perplexity. Figure 4 shows that the fit for held-out test data is not generally significantly affected by increased repetition. There is a pattern within the data, in which repeating documents 4 times seems to produce better perplexity for singular documents than 2 or 8, significantly so for a small fraction of the corpus. A theory for this is that at a sufficient level of repetition, LDA fits the repeated text to its own topic instead of trying to conflate it with other document contents, producing better topics. However, additional repetition further saturates these topics and adds noise to the meaningful co-occurrence signal.

Figure 6 demonstrates that, as before, perplexity is significantly higher as template repetition increases when there is a small number of topics $K = 5$. However, as the number of topics increases, this disparity ceases to be significant. Interestingly, however, with high enough proportion p of documents containing templates, the perplexity drops below that of documents not containing the duplicates at all numbers of topics.

For LSA, templated repetition has no apparent effect on the loss of untemplated texts. However, the effect for templated texts is less straightforward. Figure 7 shows that for proportions $p = 0.1$ and $p = 0.01$ the loss of templated texts is smaller than for untemplated texts for all K component sizes. For proportion $p = 0.001$, though, templated loss is only smaller than untemplated loss when K is large, while templated loss is never significantly lower than untemplated loss for $p = 0.0001$. *Lorem Template* and *Sample Template* also exhibit different behaviors templated texts when $p = 0.001$ and K is large: loss is is significantly smaller for *Lorem Template* and has a larger drop in loss from $K = 80$ to $K = 160$.

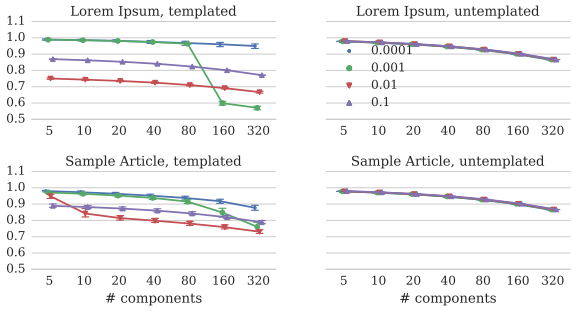

Figure 7: LSA loss of templated text for the REUSL 25k corpus. Higher levels of templating p result in smaller model loss for templated texts than for untemplated texts. For $p = 0.001$, templated loss becomes smaller than untemplated loss for $K = 160$ but more dramatically for the *Lorem Template*.

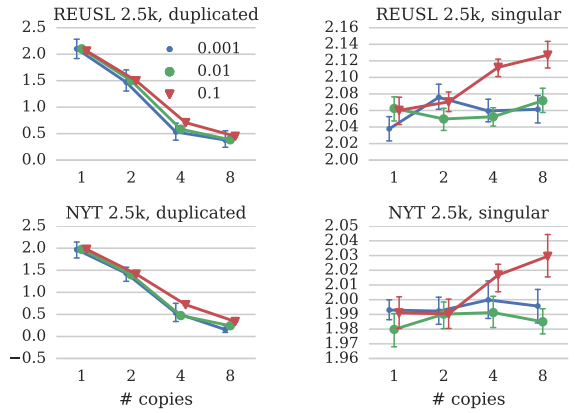

Figure 8: Entropy for LDA with 80 topics decreases for duplicated documents as the frequency of those documents increases, has little initial effect on the entropy of the singular documents.

This may indicate that LSA is able to more effectively model templated text when the templates have a distinctively different language model than the original documents.

6.2 Concentration

We expect the effect of duplication on entropy will be inversely correlated with its effect on model loss. As we increase the proportion of the corpus that is repeated, the model will devote more resources to duplicate text, leaving less modeling power for the remaining text. We therefore expect dispersion to increase with p for duplicate documents and decrease with p for singular documents. In Figure 8, the first effect clearly holds for LDA, but the second does not: there is a negligible

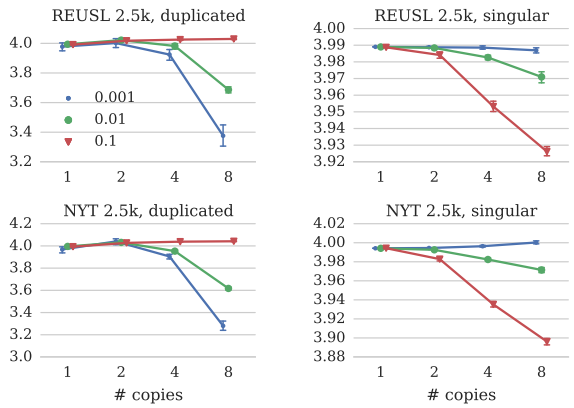

Figure 9: Absolute dispersion for LSA with 80 components increases slightly when first producing duplicates ($c = 2$), but falls off for smaller proportions of repetition $p = 0.01$ and $p = 0.001$ at higher frequencies. Increasing c has a comparatively small effect on the absolute dispersion of singular documents.

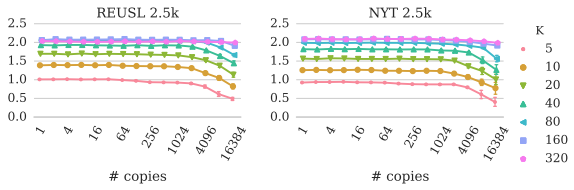

Figure 10: When a single document in the short corpora is repeated enough to comprise the majority of the corpus, the LDA entropy decreases over singular documents.

change in entropy with the number of repetitions of documents. Figure 9 shows a subtler version of the same effect for LSA. Notably, the decrease in absolute dispersion for repeated documents is only visible in the short corpora.

We can examine the extreme effects of the change in component concentration for singular documents by looking at its behavior in the *single document* treatment. In Figure 10, we see that while entropy remains level for repetitions comprising smaller portions of the corpus, eventually the entropy drops for both repeated and singular documents. This may be because most topics describe the repeated document, leaving few to model the remaining singular documents.

For LSA, absolute dispersion remains level for all repetitions tested for the *single document* treatment. This result highlights a key difference between LDA topics and LSA components: while changing the number of topics in LDA influences

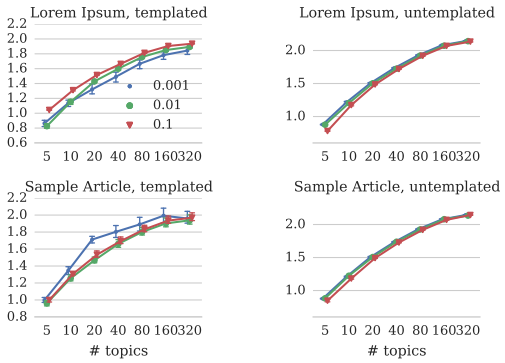

Figure 11: LDA entropy for the REUSL 25k corpus with *Sample Template* and *Lorem Template* treatments. With few topics, templated documents have lower entropy than untemplated documents, but with many topics, their entropy is higher. In the mid range of topics for *Lorem Template*, higher proportions of sampled text p produce higher entropy, but for *Sample Template*, lower p produces higher entropy.

the prior to raise or lower entropy over topics, the components of LSA are fixed. Increasing the number of components increases dimensionality, but never alters preexisting dimensions.

The effect is more subtle when templated text is repeated within documents. Figure 11 shows that with $K = 20$ LDA topics, if we apply the *Sample Template* to a small fraction of documents ($p = 0.001$), it produces a higher entropy than corpora with larger template inclusion proportion p. This is not surprising: though the template text and the original document are similar in style, with high probability they will still have different topics, which the model will have trouble fitting well without more observations. The *Lorem Template* has the reverse effect: the language is sufficiently disjoint from the content of the documents that few topics or even a single topic can model the repeated text fully, leading to low entropy. When the language model of duplicated text is disjoint from that of the text of interest, the template can be modeled by one or a few topics or components without significantly affecting other text.

6.3 Expressivity

Quantitative analyses of model fit and topic uncertainty are helpful in analyzing the effect of different settings, but do not necessarily tell us whether topics from corpora with repeated documents are useful. Analysis of expressivity helps us fill in some of the gaps in our explanations above as to what is happening at the individual topic level. In Figure 12, we see that for a moderate number of topics, increased repetition of documents impacts a substantial portion of the top-ranked words, or most probable terms of topics. The saturation effect has some relation to the number of repeated documents. With a single document repeated, as in Figure 13, as the number of topics increases, the ratio of top-ranked words belonging to the unigram language model drops. We also notice that with few topics, there is a clear "saturation point" where the topic begins to be represented more, which remains level until half the short corpora are represented by the duplicate document. The pattern overwhelmingly shows that single texts are easily fit by single topics.

In the case of the *Lorem Template* input, where little textual overlap exists between the template and original text, a few topics quickly fill in the repeated text, producing a limited effect on most topics. In Table 1, the number of topics containing "lorem" and "ipsum" remains small as the number of topics grows. Regardless of topic count of proportion, topics containing "lorem ipsum" are entirely broken Latin: the top probable terms of an example 320-topic model with $p = 0.1$ are *est justo donec iaculis sit ipsum quam lorem tristique sed amet eget pharetra curabitur fringilla non consequat mattis nec nascetur*, a direct sample of words from the template text.

7 Conclusion

The presence of duplicated strings, either documents or duplicated text within documents, is a serious but not insurmountable problem. Duplicate text can substantially alter the dimensions learned by distributional semantic models. The effect of duplication depends on several factors: the number of distinct repeated strings, the similarity of repeated strings to the rest of the corpus, and the

Proportion	5	10	20	40	80	160	320
0.001	0	0	0	0	0	0.2	0.8
0.01	0	0	0.2	0.56	1	1	1
0.1	1	1	1	1	1.78	2.3	3

Table 1: As the number of total topics increases, the average number of topics fitting the *Lorem Template* duplicate text remains stable, only rising above 1 with repeated proportion of the corpus $p = 0.1$ and at least 80 topics.

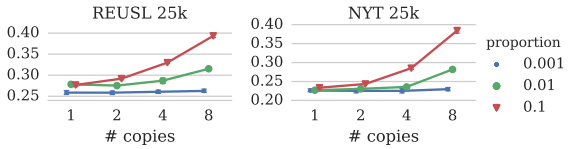

Figure 12: With 80-topic LDA models of our larger datasets, we see that increased repetition leads to significant increases in the amount of representation of repeated text in the top keys of topics.

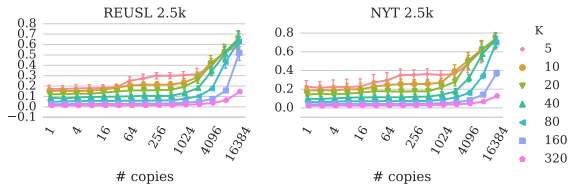

Figure 13: Top keys of LDA topics for only a single repeated document remain concentrated in only a few topics in models with $K > 5$, negligibly impacting the top keys of remaining topics.

number of repetitions. We find that different algorithms are affected in different ways, but that there are methods to alleviate the effect of duplication without exhaustively removing all duplicated documents. We provide the following specific conclusions and recommendations:

LDA accommodates low rates of document duplication for many documents. We find that with more frequent repetition, the algorithm is able to sequester repeated text into small numbers of topics if certain conditions hold. To handle this case, the model must have many topics available relative to the number of repeated strings, and the language of the repeated text must be sufficiently distinct. If these conditions are met, repeated text will affect a small number of topics that can be identified by their similarity to specific documents, or automatically based on lower than expected inter-document variability within a topic (Mimno and Blei, 2011) or distance from specific corpus-word or document-word distributions (Al-Sumait et al., 2009). We therefore suggest training a model first with slightly more topics than desired, then evaluating if there are any signs of repeated texts overwhelming several topics due to low coherence or corpus statistics. If no such indications occur, or if the duplication remains in one or two topics, then there is no need to modify the corpus or retrain the model, as the duplicate-capturing topic may be ignored.

LSA permits high rates of document duplication so long as few unique texts are repeated. Repeating one document will likely only affect one or a few components regardless of how many repetitions occur. However, if there are many different repeated documents, more components will be used to model them, which worsens the model fit more as the number of unique repeated texts increases. In this case, it may be preferable to look for near-duplicate documents more aggressively and worry less about exact duplicates. Unigram-count-based deduplication may be appropriate in this case, using a simple threshold of cosine similarity between the vectors of unigram counts between two documents to deduplicate.

Repeated text templates for LSA and LDA are sequestered by the model so long as they do not overlap heavily with topics of interest. In a topic model, it may be easy to identify the templated text based upon it appearing in one topic. However, if there is a concern that there is systematic use of text templates in documents (such as page headers or publication information) that may be too close to the language model, the n-gram removal approach inspired by Citron and Ginsparg (2015) is an expensive but straightforward way to ensure these strings are detected and deleted. The combination of unigram deduplication, n-gram deletion, and the inherent ability of semantic models to separate co-occurring text should reduce the negative effects of text duplication.

Acknowledgments

We would like to thank our reviewers for their helpful feedback and suggestions. We would also like to thank Jack Hessel, Måns Magnusson, and Ana Smith for their reviews and feedback. This material is based on work supported by the DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a; National Science Foundation grants #1526155, #1652536 (CAREER), and #DGE-1144153 (GRFP); and a faculty research fellowship from the Alfred P. Sloan Foundation.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 13–22.

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. Springer.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, pages 280–288.

Daniel Bär, Torsten Zesch, and Irnya Gurevych. 2012. Text reuse detection using a composition of text similarity measures. In *Proceedings of the 24th International Conference on Computational Linguistics*, volume 1, pages 167–184.

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

Andrei Z Broder, Steven C Glassman, Mark S Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. *Computer Networks and ISDN Systems*, 29(8-13):1157–1166.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 136–145. Association for Computational Linguistics.

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.

Daniel T Citron and Paul Ginsparg. 2015. Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences*, 112(1):25–30.

Paul Clough, Robert Gaizauskas, Scott SL Piao, and Yorick Wilks. 2002. Meter: Measuring text reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 152–159.

Paul Clough et al. 2003. Old and new challenges in automatic plagiarism detection. In *National Plagiarism Advisory Service*. Http://ir.shef.ac.uk/cloughie/index.html.

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.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, page 30.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on the World Wide Web*, pages 406–414. ACM.

Gustavo Graff, David amd Gallegos. 1995. Spanish news text. *Linguistic Data Consortium*, DVD: LDC95T9.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, pages 331–339.

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 52nd Annual Meeting of the Association for Computational Linguistics*, pages 530–539.

John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 472–479.

Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th International Conference on the World Wide Web*, pages 141–150. ACM.

Andrew K McCallum. 2002. MALLET: a machine learning for language toolkit. Available at: http://mallet.cs.umass.edu.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

David Mimno and David Blei. 2011. Bayesian checking for topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 227–237. Association for Computational Linguistics.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In

Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 262–272. Association for Computational Linguistics.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processin*, volume 14, pages 1532–1543.

Martin Potthast and Benno Stein. 2008. New issues in near-duplicate detection. In *Data Analysis, Machine Learning and Applications*, pages 601–609. Springer.

Evan Sandhaus. 2008. The New York Times annotated corpus. *Linguistic Data Consortium*, DVD: LDC2009T19.

David A Smith, Ryan Cordell, and Elizabeth Maddock Dillon. 2013. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *Proceedings of the IEEE International Conference on Big Data*, pages 86–94.

Hanna M Wallach, David Mimno, and Andrew K McCallum. 2009a. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009b. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM.