# Understanding Text Pre-Processing for Latent Dirichlet Allocation

**Alexandra Schofield**[1] **Måns Magnusson**[2] **Laure Thompson**[1] **David Mimno**[3]
1 Department of Computer Science, Cornell University, Ithaca, NY
{xanda, laurejt}@cs.cornell.edu
2 Department of Statistics, Linköping University, Linköping, Sweden
mans.magnusson@liu.se
3 Department of Information Science, Cornell University, Ithaca, NY
mimno@cornell.edu

## Abstract

To apply natural language modeling techniques to new corpora requires users to convert documents to data using various pre-processing treatments. However, the effects of these transformations are still poorly understood. We describe several studies that quantify the impact of pre-processing in different forms, focusing on topic modeling applications. We find that many common practices either have no measurable effect or have a negative effect after accounting for biases induced by feature selection. Finally, we provide recommendations as to how to pre-process text for novice users of topic models looking to investigate their own text corpora.

## 1 Introduction

As NLP methods become more popular as an exploratory tool outside machine learning, it becomes increasingly important to provide recommendations for practitioners for how to apply them effectively. This work quantitatively evaluates pre-processing treatments in order to help users make informed decisions about which to use, as pre-processing effects can substantially impact model output (Denny and Spirling, 2016; Boyd-Graber et al., 2014). Our work differs from existing work in that it specifically focuses on quantitative comparisons of different pre-processing treatments that might not be directly comparable. We focus on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) as it has enabled large corpus exploration in diverse fields such as literature (Goldstone and Underwood, 2012; Rhody, 2012), archaeology (Mimno, 2011), classics (Mimno, 2012), history (Newman and Block, 2006), and political science (Gerrish and Blei, 2012).

In this work, we explore projects that look at three different steps in a pre-processing pipeline: document de-duplication,[1] stopword removal (Schofield et al., 2017), and stemming (Schofield and Mimno, 2016). In order to study the relationship between text treatments and models, we first build a methodology for comparing models trained with a pre-processing treatment to those where the treatment was applied *after* training. Second, because typical evaluations are sensitive to vocabulary reduction and corpus modification, we present new and modified metrics to evaluate topic model quality in the presence of such confounding factors. Finally, we provide recommendations for standard pre-processing.

## 2 Methods

Usually in machine learning research, we keep the data fixed and compare models produced by different algorithms. Here, we keep the algorithm fixed and compare models produced by different corpora. But evaluation metrics such as model perplexity and between-term mutual information are sensitive to data volume and vocabulary size, so it becomes crucial to define metrics that make comparative evaluation meaningful across slightly different sets of documents.

**Pre-processing vs. Post-processing** Our primary strategy is to compare pairs of models, one trained on a corpus that has had a particular pre-processing procedure applied before training, and one trained on a corpus that has had the same procedure applied *after* model inference. This allows the comparison of the trained model output with a given treatment to output without it, while accounting for the differences in evaluative metrics caused by changes in the underlying documents. We evaluate topic models trained using

[1] This work is currently under submission at EMNLP 2017

a collapsed Gibbs sampler with Mallet (McCallum, 2002), in which each token has an associated topic variable. We then perform whatever corpus transformation would have occurred in pre-processing: for stopword removal, tokens and their topic assignments are deleted, while for stemming, tokens are stemmed but their topic assignments are left intact. We may then re-infer document-topic and topic-word distributions ($\theta$ and $\phi$) from MAP estimates as if Gibbs sampling were originally performed on this transformed corpus. The corpus will be identical in appearance to the one pre-processed before inferring the model, allowing us to isolate the effect of pre-processing on inference.

**Bad Models vs. Bad Representations** The classic method of previewing the contents of a topic model is to display each topic as a list of the highest-probability words in the topic. Through this lens, the choice to reduce vocabulary size has clear advantages for the representation quality of these probable terms. The benefit is the increase in the apparent content of these topic summaries: removing stopwords increases the number of remaining terms that convey clearly topic-specific semantic content, while morphological conflation combines near-duplicate terms. However, it is important to distinguish these apparent effects of summarized topic representations "looking bad" from the actual effect of these treatments on the inference of the document-topic and topic-word distributions themselves.

To do this correctly requires normalization to counter the apparent improvements in evaluation metrics that would be produced by any vocabulary reduction technique. Held-out predictive likelihood tasks are particularly vulnerable to spurious results. We observe that stronger stemming treatments, those that more aggressively reduce vocabulary size, improve model fit. However, much of this improvement can be accounted for by the reduction of the probability space of the model produced by a reduced vocabulary. If we normalize by the probability of a unigram language model on the same text (a measure of how complex the text is in the first place) we can calibrate the improvement of a more complicated model. This calibration allows us to draw a distinction between improvement due to the decrease in vocabulary size and improvement due to intelligently constraining the model to conflate words sharing a stem.

**Metrics** We use a variety of metrics to understand topic quality, principally estimated held-out likelihood (Wallach et al., 2009) and automatic coherence metrics (Bouma, 2009; Mimno et al., 2011). In addition, we consider variation of information (Meilă, 2003), or VI, as a measure of how much topics change over different stemming treatments. For stopword analysis, inspired by Dredze et al. (2008), we evaluate topic representations by their top words based upon the accuracy of a classifier trained to identify the most probable topic of a document given unigram count features of only the top 15 terms of each topic. We also use measures of information entropy of documents across topics and mutual information between topics and documents to understand the details of how document-topic distributions are affected by these changes.

## 3 Results

We conduct experiments using 10 trials per treatment on a variety of corpora across our different works, including ArXiv papers,[2] New York Times articles (Sandhaus, 2008), Reuters newswires (Graff, 1995), biographies from IMDb,[3] reviews from the Yelp Dataset Challenge,[4] and US State of the Union addresses.

**Document duplication** Using synthetic repetition of data, we find that as documents are repeated, topic models begin to devote topics exclusively to the repeated documents. Repeated documents show very low topical entropy and high likelihood. However, text without these repetitions is largely unaffected: repeated text is quickly fit well to one or a few topics, leaving the rest of the model unaffected, except for the implicit loss of modeling power caused by "losing" one or more topics. We find that topic models can accommodate occasional duplicates and fit topics to a repeated string across many documents, but that this is more difficult if the repeated text has similar language to the content of interest. In our experiments, effects of duplication were minimal until duplicate documents became a substantial proportion of the corpus, whether one document repeated over a thousand times or 1% of the corpus repeated four times.

**Stopword removal** Except for the dozen or so most frequent words, removing stopwords has no

[2]Retrieved from ArXiv (http://www.arxiv.org).
[3]Courtesy of IMDb (http://www.imdb.com).
[4]Retrieved from Yelp (http://www.yelp.com/dataset_challenge)

substantial effect on model likelihood, topic coherence, or classification accuracy. Mutual information between the document ID and topic assignment of each token reveals that removing stopwords does not affect the specificity of non-stopword topic assignments to their document distributions. We find that removing determiners, conjunctions, and prepositions can improve model fit and quality, but that further removal has little effect on inference for non-stopwords and thus can wait until after model training.

**Stemming** Once likelihood is normalized for vocabulary reduction, stronger stemming methods (such as the Porter stemmer (Porter, 1980)) perform significantly worse in improving likelihood than not stemming or weak morphological treatments. Topic coherence is also not improved between pre-stemming and post-stemming topic models. VI shows that, instead of forcing topic models to be more consistent, stronger stemming treatments produce less consistent topic assignments. At least for English, morphological conflation treatments such as stemmers and lemmatizers can worsen topic model quality on a variety of measures, while LDA turns out to be quite good at combining morphological variants by itself.

**Recommendations** We conclude the following:

1. Document duplication can alter model inference, but requires substantial quantities of repetition and can be sequestered into individual topics.

2. Aside from extremely frequent stopwords, removal of stopwords does little to impact the inference of topics on non-stopwords.

3. Topic model inference often places words sharing morphological roots in the same topics, making morphological conflation such as stemming redundant and potentially damaging to the resulting model.

Our results substantially simplify the work of practitioners. Many burdensome tasks turn out to have little effect, such as stemming and corpus-specific stoplists. As a result, the methodology of post-processing a corpus instead of pre-processing can allow practitioners the option to test out pre-processing options on one trained model to decide on a treatment best suited for their application. The results also suggest that when possible, it is good for practitioners to pre-process more lightly to avoid discarding useful word information.

## 4 Acknowledgments

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. We would like to thank our reviewers for their helpful and insightful comments.

## References

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* pages 31–40.

Jordan Boyd-Graber, David Mimno, David Newman, Edoardo M Airoldi, David Blei, and Elena A Erosheva. 2014. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications* pages 3–34.

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

Mark Dredze, Hanna M. Wallach, Danny Puller, and Fernando Pereira. 2008. Generating summary keywords for emails using topics. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, IUI '08, pages 199–206. https://doi.org/10.1145/1378773.1378800.

Sean Gerrish and David M Blei. 2012. How they vote: Issue-adjusted models of legislative behavior. In *Advances in Neural Information Processing Systems*. pages 2753–2761.

Andrew Goldstone and Ted Underwood. 2012. What can topic models of pmla teach us about the history of literary scholarship. *Journal of Digital Humanities* 2(1):39–48.

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

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

Marina Meilă. 2003. Comparing clusterings by the variation of information. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, Springer Berlin Heidelberg, volume 2777 of *Lecture Notes in Computer Science*, pages 173–187. https://doi.org/10.1007/978-3-540-45167-9_14.

David Mimno. 2011. Reconstructing Pompeian households. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. AUAI Press, pages 506–513.

David Mimno. 2012. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage (JOCCH)* 5(1):3.

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*. Association for Computational Linguistics, pages 262–272.

David J Newman and Sharon Block. 2006. Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology* 57(6):753–767.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program* 14(3):130–137.

Lisa Rhody. 2012. Topic modeling and figurative language. *Journal of Digital Humanities* 2(1):19–35.

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

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. *EACL 2017* page 432.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics* 4:287–300.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, New York, NY, USA, ICML '09, pages 1105–1112.