Frontier Identification: Topic Modeling and SNR

Large amounts of digitalized text, referred to here as *documents*, are becoming increasingly available to the average Internet user via online news articles, social media outlets, question-answer forums, and more. Most of these users are hoping to perform one of two actions when presented with such a large amount of information: browse or search. Historically, the ability to browse or search relied heavily on the use of document metadata. For example, searching for a book or comparing related books in a library database requires knowing titles, genres, authors, or a manually defined set of keywords. Metadata can be hard to come by and often requires a lot of human involvement either from the user performing the action (deciding which keywords to use) or the creators of the search/browse algorithm (labeling each document and/or generating the metadata). With constant improvements in technology and a trend towards increased micro-blogging, we now have access to the full-text of documents meaning it may be possible to reduce our dependance on metadata. Eventually, sifting through massive amounts of documents will rely solely on the full-text of the documents themselves and will occur with such speed and efficiency that new documents will be search-/browsable shortly after being uploaded.

Topic modeling addresses the desire to improve user experience in the large corpora described above. The generative model Latent Dirichlet Allocation (LDA) is currently the most widely used topic modeling algorithm. First introducted by [1], LDA identifies latent topics within discrete datasets by considering each topic to be a multinomial distribution over the words in the vocabulary. In turn, each document is composed of topics sampled from a Dirichlet distribution over the topic-space. LDA was successful at addressing issues that previous models had with overfitting, unrealistic increases in parameter counts, and infeasible runtimes [1].

Following its arrival, LDA was applied to many situations both with and without success. For more than a decade, there was only a casual understanding of which situations were well-suited to LDA and which were not. Researchers made ad-hoc changes to the model to suit the properties of their corpus. They gave minimal consideration to how these changes would generalize to other types of data. For example, researchers soon discovered that LDA was ill-suited to documents that were too short (such as tweets) and documents that were too long, with too many topics (such as books). [2] addresses the issue of documents being too long and containing too many topics with DCM-LDA, a model which they tested on 1.4 billion words and more than 12,000 topics. However, their claims for generalization apply only to *increases* in the size of the corpus, the length of documents, and the number of topics. They do not address changes in the other direction. This is left to others such as the modified LDA models seen in [4, 5] which utilize separate forms of aggregation to craft larger documents out of shorter documents. Like [2], they only test and evalute their new method on the specific type of document it was designed to address.

Tang et al. [3] provided the first deliberate study of LDA parameters and their limitations as well as the specific properties that make a dataset well-suited for LDA. Their experiments specifically focused on measuring the contraction rate of topics around the true topic-space center as the number of documents and size of each document increases to infinity. Results showed the importance of choosing the correct number of topics and setting the Dirichlet priors to large or small values depending on the number of topics associated with each document.

Despite a near-complete understanding of LDA's limiting factors after [3], resarch is still missing the mark as publications continue to show modifications being fine-tuned to specific datasets [5]. Consistent, generalized methods for how to construct better documents and evaluate resulting models are lacking.

It is well-understood that co-occurrence of words is the underlying phenomenon that defines a "topic" [2-5]. Future work should utilize this knowledge when deciding *how* to generate appropriately sized documents during pre-processing. In addition, if modifications to LDA are going to continue, we need a consistent method for evaluating model performance. Currently, classification tasks and comparison against ground-truth topic labels are the only two methods used. These are risky methods because ground-truth is not always available and generating labels for datasets is not always feasible.

Reflecting again on the concerns of users (browsing and searching) there is perhaps an overlooked phenomenon driving their behaviour: an unconscious goal or signal. In both browsing and searching, the user either knows what it is they are looking for beforehand or learns what they were searching for after coming across it while browsing. An SNR approach to topic modeling could therefore drastically improve both of these actions. Understanding what makes some topics strong signals (or interesting) and other topics weak signals (or boring) is the next logical step in analyzing these different extensions on to LDA.

Most publications already include some sort of quality measure for individual topics and distance measures for comparing topics or topic-spaces. All use topic coherence to establish a quality for topics. Zhao uses human subjects to score topic quality (an insightful method although it does not scale well). Jensen-Shannon Distance is used by [2, 4, 5] to compare topics within the same topic-space and by [3] to compare topics from different topic-spaces. The only missing step is to take these similar approaches and conduct one formal method for evaluating derived topic qualities.

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3* (Jan), 993-1022.

[2] Mimno, D., & McCallum, A. (2007, June). Organizing the OCA: learning faceted subjects from a library of digital books. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 376-385). ACM.

[3] Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014, January). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning* (pp. 190-198).

[4] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011, April). Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval* (pp. 338-349). Springer, Berlin, Heidelberg.

[5] Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016, August). Topic Modeling of Short Texts: A Pseudo-Document View. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2105-2114). ACM.