Research Questions: Topic Modeling and SNR

What effect does modifying the input corpus have on the performance of topic modeling?

The answer to this question requires generating a fixed approach for topic modeling. This system will have the most optimal settings for the most generic topic modeling algorithm. Then, a set of performance metrics need to be defined to measure the systems success at completing the task (topic modeling). To ensure that experiments are varied, a set of corpus metrics needs to be defined so that the input to the system can be sure to cover a wide range of possible values. This will allow the most complete test for causal effects of corpus properties on model performance.

First, a fixed approach for topic modeling. There has been extensive work to develop and revise various approaches to this problem [1-7]. The literature has mostly agreed that Latent Dirichlet Allocation (LDA), see *Figure 1*, is the best of these approaches [1,2]. In its most generic form, LDA requires several input variables before it can be fit to a specific corpus. To fix the model, it is necessary to fix: the number of topics (K), the prior distribution of topics within documents (Alpha), and the prior of words within topics (Beta). Previous work has shown that the model performs best with asymmetric Alpha and symmetric Beta [5]. In [4], it was shown that the posterior contraction rate of topic distributions is independent of the number of topics. Therefore, it is only necessary to choose K to be sufficiently high such that the model can effectively separate any topics present in the corpus.

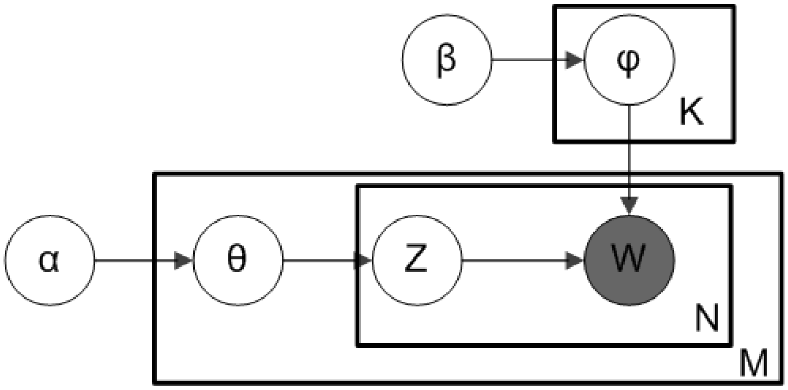

Figure 1. The Latent Dirichlet Allocation model.

This work differs from that of past work in that it fixes the topic modeling approach and seeks to define or observe causal changes in its performance under modification of the input corpus. Past work jumps directly from observing poor performance of the model on some corpora to modifying the model ad-hoc to accommodate the specific corpora. This project seeks to bridge the gap and explain the phenomenon of *why* some corpora result in poor model. In particular, the answer to this research question will involve observation and characterization of causal relationships between different corpus properties and resulting performance of the fixed model. Unlike past work, this project treats the corpus as the independent variable and observes model performance as the dependent variable.

For metrics to measure the performance of the model, the literature has widely agreed that Jensen-Shannon distance and Coherence are sufficient [3,6,7]. In addition to applying these metrics, the forefront of technology would be to prove that they satisfy a set of one or more basic axioms to ensure their usefulness as measures of performance. This requires answering the question: what axioms ensure a good performance metric?

From here, to identify causal relationships between corpus properties and model performance, it is necessary to define corpus properties that can be modified. Past work has hinted at causal relationships between performance of the model and both number of documents as well as average document length [6]. To push beyond the frontier of this knowledge, the next task is to determine threshold values for these properties on the base LDA model and also to ask, what other potential corpus properties might affect model performance?

To stay at the forefront of technology as well as science, the properties of the corpus need to be presented in a way that ties back to real-world scenarios. In previous topic-modeling research, real-world datasets were used ad-hoc and primarily referred to by their source (e.g. Twitter, Wikipedia, NYT) instead of their characteristics (e.g. multinomial, long-text, coherent, short-text, etc.). With minor exceptions, corpus characteristics such as document length were only mentioned as a means to explain poor performance and justify ad-hoc changes to improve performance (such as combining short documents before fitting, seen in [6]). As mentioned in the Frontier report, the forefront of technology is an increasing availability of digitized text. With increased availability comes increased variability. Therefore, to best answer the question posed above while remaining at the forefront of technology, it is necessary to also answer the question: Where can the phenomena (i.e. corpus properties) being tested in this project be found in real-world data?

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

[2] Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1608.

[3] 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.

[4] 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).

[5] Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in neural information processing systems* (pp. 1973-1981).

[6] 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.

[7] 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.