Memorable Title

Tommy Jones[*]

Abstract

This is the abstract.
It consists of two paragraphs.

1 Introduction

Probabilistic topic models are widely used latent variable models of language. Popularized in 2002 by latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) many related models have been developed, for example (Blei and Lafferty 2007), (Roberts et al. 2013), (Nguyen et al. 2015), and more. These models share common characteristics and estimate the probability of topics within contexts and tokens within topics.[1] Even today LDA remains one of the most popular topic models, and one of the simplest.

Topic models have been applied to a variety of tasks. These tasks include information retreival (Wei 2007), analysis of historical texts (Newman and Block 2006), machine translation and related tasks (Vulic, Smet, and Moens 2011), and more. In recent years, the machine learning community has focused more on deep architectures typified by text embeddings (Grohe 2020) and pre-train then fine tune transformers, for example (Henderson 2020). Yet probabilistic topic models have remained popular analytical methods in fields such as computational social science (Roberts, Stewart, and Airoldi 2016) and the digital humanities (Erlin 2017).[2]

In spite of their sustained popularity, probabilistic topic models remain challenging to use. Some of these challenges are conceptual. Probabilistic topic models have user-set tuning parameters, called "hyperparameters" in the machine learning literature, whose optimal settings are not obvious. Moreover, because probabilistic topic models estimate parameters for a process that is *not* how people write. Because of this, there is no ground truth against which to compare models for a sense of "correctness" that researchers can use to develop modeling strategies and metrics to guard against pathological misspecification.

In some cases, challenges are more practical. Software implementing probabilistic topic models can be challenging to use and offer limited functionality. In particular, those that employ probabilistic topic models in industry often have a need to update models based on new or updated data. To date, there has been little research on transfer learning for probabilistic topic models. No off-the-shelf software implements such a paradigm. The result is that applied practicioners face an unpleasant tradeoff. Either models go stale or they must be re-trained from scratch. In the former case, innacuracies creep in over time. In the latter case, topics are re-initialized at random, breaking continuity with the old model.

What is more, transfer learning is becoming paramount to modern machine learning for natural language. The last few years have seen an explosion of "transformer" models which rely on a paradigm of training a "base" model on an unsupervised or semi-supervised task. These base models tend to use as much language data as possible. Then the base model is transfered to a smaller dataset on a narrow supervised task. The result has been an impressive increase in performance on many standard NLP benchmarks. No such paradigm exists for probabilistic topic models.

[*] George Mason University Dept. of Computational and Data Sciences, tjones42@gmu.edu

[1] Technically, probabilistic topic models estimate the probability that any token was sampled from a topic given the context and the probability of sampling each specific token given the topic, respectively.

[2] As we will see in Section 3, probabilistic topic models can share similar conceptual frameworks with newer methods.

In an attempt to address these shortcomings, I propose three research studies, each building on the last. In each, I will focus on Latent Dirichlet Allocation (LDA) for its simplicity and popularity. LDA is closely related to other probabilistic topic models. This enables a natural extension of this research to other probabilistic topic models.

The first study relates some empirical laws of language to LDA as a generative process. This enables a principled method for conducting simulation studies of LDA. Simulation studies are a natural means for imposing a sense of "correctness" in studying statistical models (Morris, White, and Crowther 2019). Afer linking LDA to empirical laws of language, this first study will use a combination of simulations and analytical derivations to address hyperparameter settings for LDA. The objective is not to develop methods for finding the "correct" model on real data, as no such model exists. Rather it is to set up guardrails to avoid models that are pathologically misspecified where an obviously better model does exist.

The second study develops methods for transfer learning in LDA. This enables the applied practicioner to update models with new data, preservig continuity with previously-trained models. It also takes a first step extending LDA towards the state of the art "pre-train then fine tune" paradigm currently popular in natural language processing.

The final study introduces `tidylda`, a software package for the R programming language (R Core Team 2013). `tidylda` integrates into a wider programming paradigm in the R language known as "tidy" programming. It also implements several novel methods for and related to LDA, including transfer learning.

The remainder of this document is organized as follows:

- Section 2 gives a brief history of embedding models for text, a broader class of models encompassing probabilistic topic models.
- Section 3 re-states the formulation for LDA, compares it to related models, and discusses training algorithms for LDA.
- Section 4 explores current approaches for evaluating and studying probabilistic topic models, with a focus on LDA.
- Section 5 gives an overview of simulation studies in statistics broadly and how they have been applied to probabilistic topic models.
- Section 6 reviews some empirical laws of language that a synthetic data set of language must honor to be considered a valid simulation of natural language.
- Section 7 outlines the proposed dissertation studies

2 A Brief History of Embedding Models for Text

3 Latent Dirichlet Allocation (and Friends)

4 Proposed Studies

4.1 Simulation Studies for LDA

4.1.1 Background

Using simulated data to study LDA is not new [cite]. However, the degree to which it can approximate the true statistical properties of human language has not been closely examined.

4.1.1.1 Evaluation Methods for Probabilistic Topic Models

4.1.1.2 Creating Ground Truth with Simulation Studies

4.1.1.3 Emperical Laws of Language

4.1.2 Approach

This study has three goals:

1. To simulate data using LDA's generative process that is statisically as close to human language as possible,
2. To quantify the degree to which simulated data does and does not statistically approximate human-generated language data, and
3. To study simulated corpora to inform modeling decisions, such as hyperparameter selection.

One must have a baseline of real data for comparison. Then one can measure the accuracy of aggregate statistical patterns between the simulated and actual data. Individual patterns &mdash e.g. counts of single tokens, counts of each tokens within a document, correlation between any two specific documents, etc. &mdash cannot be captured. Simulations provide population-level data patterns, but not meaningful representations of individual data points.

The above compells two needs for this study. The first need is a sample of corpora that demonstrate the ability of simulations to generalize broadly, not work for just one or a few languages or contexts. This does not need to represent "language" as a whole. Rather it must be diverse enough to demonstrate that simulated data can capture patterns across languages and contexts. The second need is for a set of representative population-level statistics sufficent to describe a corpus. The degree to which simulations can represent the sufficient statistics of real corpora is key to this study, particularly the first two goals above.

4.1.3 Expected Results

words

4.2 Transfer Learning for LDA

4.3 Introducing `tidylda` for R

words words words

5 Expected Timeline

References

Blei, David M., and John D. Lafferty. 2007. "A correlated topic model of Science." *The Annals of Applied Statistics* 1 (1): 17–35. https://doi.org/10.1214/07-aoas114.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3.

Erlin, Matt. 2017. "Topic Modeling, Epistemology, and the English and German Novel." *Journal of Cultural Analytics*. https://doi.org/10.22148/16.014.

Grohe, Martin. 2020. "word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data." *arXiv*.

Henderson, James. 2020. "The Unstoppable Rise of Computational Linguistics in Deep Learning." *arXiv*.

Morris, Tim P., Ian R. White, and Michael J. Crowther. 2019. "Using simulation studies to evaluate statistical methods." *Statistics in Medicine* 38 (11): 2074–2102. https://doi.org/10.1002/sim.8086.

Newman, David J., 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): 753–67. https://doi.org/10.1002/asi.20342.

Nguyen, Thang, Jordan Boyd-Graber, Jeffrey Lund, Kevin Seppi, and Eric Ringger. 2015. "Is Your Anchor Going Up or Down? Fast and Accurate Supervised Topic Models." In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the Acl*, 746–55. https://doi.org/10.3115/v1/n15-1076.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Roberts, Margaret E, Brandon M Stewart, and Edoardo M Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111 (515): 988–1003. https://doi.org/10.1080/01621459.2016.1141684.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." In.

Vulic, Ivan, Wim De Smet, and Marie-Francine Moens. 2011. "Identifying Word Translations from Comparable Corpora Using Latent Topic Models." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Wei, Xing. 2007. "Topic Models in Information Retrieval." MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE.