Memorable Title

Tommy Jones*

1 Introduction

Human language is one of the most information rich sources of data that exists. Language is literally the medium humans use to communicate information to each other. And in an increasingly digitally connected world, the amount of text available for analyis has exploded. Improvements in computing power and algorithmic advances has driven staggering progress in machine translation, automatic summarization, information extraction and more.

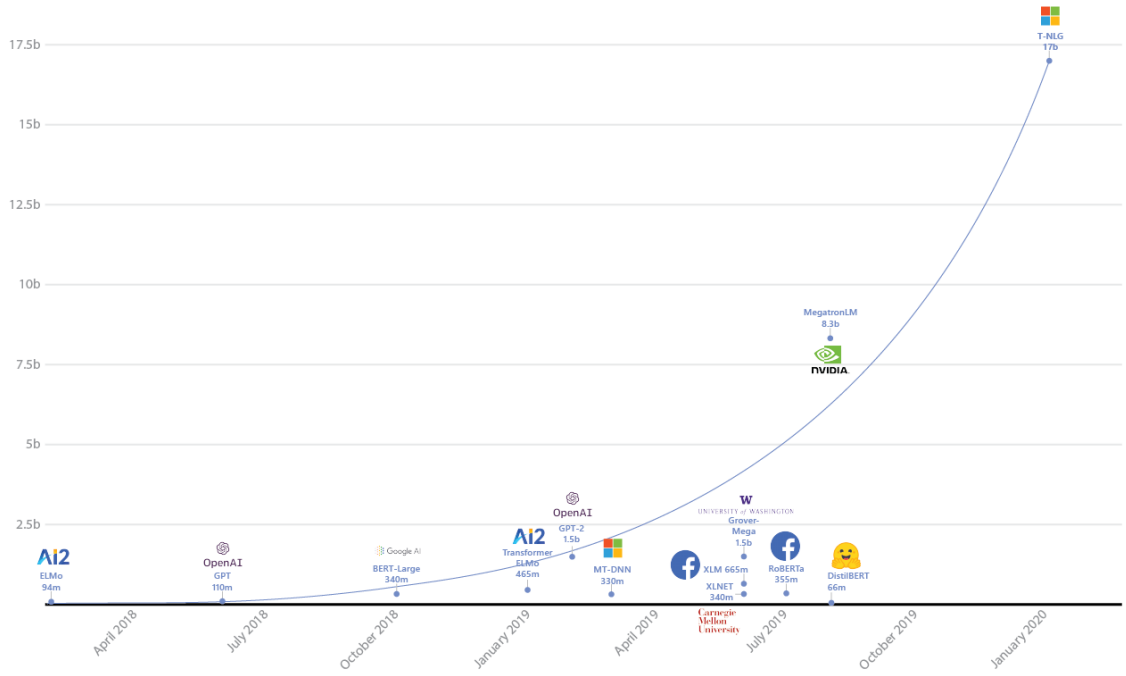
Current state of the art natural language processing (NLP) models belong to a class of deep neural networks called "transformers". Famous examples of tranformers include ERNiE (cite), BERT (cite), XLNet (cite), GPT-2 (cite), GPT-3 (cite), and more. Transformers operate on a transfer learning paradigm called "pre-train then fine tune". In the pre-training step, a "base" model is trained on an unsupervised or self-supervised (e.g. predict the next word) task with a very large volume of textual data. The fine tuning step involves replacing the output layer of the base model with a task specific output layer. Then a much smaller set of task-specific training data is used to update the weights of interior layers and lern new weights for the output layer.

This approach has obvious advantages. In terms of raw accuracy for benchmark task-specific objectives, transformers reign supreme. It seems, also, that one can get acceptable results with fewer labeled examples when starting with a base model than using traditional end-to-end models. (cite)

Yet in spite of these advances, machine learning for NLP remains a largely ad-hoc field. Save a handful of empirical laws, there is little statistical theory guiding the modeling of textual data. What theory does exist generally does not inform specification or use of statistical or machine learning models of text. Instead, the field has relied on icreasingly complex models, requiring tremendous computational power, to drive these advances.

Figure 1 depicts the number of parameters in several famous examples of transformers. These deep neural network models are not only complex, they are expensive to train. GPT-3 (not pictured), the latest and greatest member of this class has approxmiately 17.5 billion parameters and cost an estimated $12 million to train. (cite)

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



https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

In addition to increasing complexity and cost, there is some evidence that the marginal returns to research in several subfields of machine learning are declining. (cite) When progress in one direction begins to slow, it may be time to push in another direction. Perhaps it is time to revisit statisitcal theory?

I propose reexamining a model that has become less popular in machine learning circles, Latent Dirichlet Allocation (LDA). Why? With the above comments in mind, LDA has some desireable properties. It models a data generating process which may be linked to the empirical laws of language. This property makes LDA, and related models, candidates for helping to develop a more robust statistical theory for modeling language. Akin to what we have for linear regression, statistical theory helps guide modeling decisions. This often results in models that are accurate, parsimonious, and interpretable. Modern NLP models acheive only the first. And while LDA may be less popular at the cutting edge of machine learning, it and its variants are still popular in fields such as computational social science (Roberts, Stewart, and Airoldi 2016) and the digital humanities (Erlin 2017). Finally, I believe that I have developed method of transfer learning for LDA, allowing it to be used in a pre-train then fine tune paradigm similar to that which is employed in transformer models.

To complete the requirements of my dissertation, I propose making three contributions. The first is a theoretical study of the LDA data generating process (LDA-DGP). The second is a study exploring transfer learning for LDA. The third is a software package for the R language that draws on my research and a framework known as the "tidyverse" (cite) to make a principled, flexible, performant, and user-friendly interface for training and using LDA models.

The remainder of this proposal is organized as follows: Section 2 reviews the foundations of LDA. Section 3 outlines my proposed study of the LDA-DGP. Section 4 outlines my proposed study of transfer learning for LDA. Section 5 introduces *tidylda* an in-develpment R package for LDA. Finally section 6 offers a timeline for completing the proposed dissertation.

2 Background: Latent Dirichlet Allocation

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

Latent Dirichlet Allocation (LDA) is a Bayesian model of language. It models an idealized stochastic process for how words get on the page. Instead of writing full, syntatictically-coherent, sentences, the author samples a topic from a multinomial distribution and then given the topic samples a word. The process for a single draw of the n-th word for the d-th document, $w_{d,n}$, is

1. Sample a topic $z_{d,n} \sim Multinomial_K(1, \boldsymbol{\theta}_d)$
2. Given topic $z_{d,n}$, sample a word $w_{d,n} \sim Multinomial_V(1, \boldsymbol{\phi}_{z_{d,n}})$

The variable $z_{d,n}$ is latent. The author repeats this process Nd times until the document is "complete".

For a corpus of D documents, V unique tokens, and K latent topics, the goal is to estimate two matrices: $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$. The d-th row of $\boldsymbol{\Theta}$ comprises $\boldsymbol{\theta}_d$, above. And the k-th row of $\boldsymbol{\Phi}$ comprises $\boldsymbol{\phi}_k$. LDA estimates these parameters by placing Dirichlet priors on $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$.

- $\boldsymbol{\theta}_d \sim Dirichlet_K(\boldsymbol{\alpha})$
- $\boldsymbol{\phi}_k \sim Dirichlet_V(\boldsymbol{\beta})$

The LDA model can be loosely represented as a Bayesian network. In the LDA literature, this diagram is called a "plate" diagram, as it has boxes (plates) representing different levels of the model. A plate diagram for LDA is below.

3 Study 1: Examining the LDA-DGP

LDA models a process that generates language data. For ease of discussion, call this the "LDA-DGP". The LDA-DGP, described in more detail below, is clearly not how people write. Yet the degree to which the LDA-DGP can—or cannot—generate data that share the same statistical properties of human language has received little study. If the LDA-DGP can generate data that reasonably approximates human language data, then studying a collection of synthetic corpora drawn from the LDA-DGP can inform systematic strategies and rules for selecting hyperparameters and diagnosting model misspecification.

[Describe LDA-DGP]

Paragraph on history of research: 2011 goldwater et al 2015 taylor law thing 2019 guy dissertation

The study I propose has three components, described below. The first component is analytical, linking the three statistical laws of language to the LDA-DGP. The second component will generate a set of synthetic corpora using the LDA-DGP. [use statistical design to define a sample space covering any corpus likely to be fit with LDA and generate synthetic data to cover it] The third component, will analyze these corpora and the models that generated them to develop strategies for picking hyperparameters and diagnosing pathological model misspecification when using LDA on real corpora.

[1] While there are distinct differences in the definitions of "word" and "token", for the purposes of this work I will use the two terms interchangibly for simplicity.

3.1 Linking LDA to Empirical Laws of Language

3.1.1 Thesis

If a data generating process purports to simulate the process generating human language, then the resulting data should display statistical properties consistent with empirical laws of language. The degree to which the LDA-DGP can or cannot capture these statistical properties is a measure of how well lessons derived from studying the LDA-DGP may extend to real-world corpora.

This portion of the study is largely analytical, as opposed to statistical. I will mathematically examine the LDA-DGP and attempt to link it to relevant statistical laws of language. If successful, this will form a principled link between LDA and natural language.

3.1.2 Empirical Laws of Language

The gross statistical properties of language are captured in a set of empirical laws. The most famous is Zipf's law [cite] which captures the relationship between a word's frequency in a corpus and its rank when words are ordered from most to least frequent. Zipf's law is not the only such law. Altman and Gerlach describe 9 such laws, depicted in Table [X]. [cite]

[TABLE X ABOUT HERE]

Of these laws, three are relevant to LDA. They are Zipf's, Heap's, and Taylor's laws. LDA itself is principally a "bag of words" model, where word order within a document does not matter.[2] Nor does LDA concern itself with subword components such as phonemes. For these reasons, the other 6 laws listed in Table [X] do not apply.

3.1.2.1 Zipf's Law

Zipf's law describes the relationship between a word's frequency and its frequency-rank as a power law. It is typically parameterized by

$$f_r \propto r^{-a} \tag{1}$$

where f_r is a word's frequency, r is its frequency-rank, and b is a free parameter to be fit from data. An alternative, statistical parameterization is

$$P(f_r|b) \propto f_r^{-b} \tag{2}$$

which states that the probability of observing a given word frequency is subject to a decreasing power law. The two parameterizations may be mapped to each other where $b = 1 + a^{-1}$. See Figure [X] (a) for a graphical depiction of Zipf's law.

Zipf's law has also been extended into the Zipf-Mandlebrodt law to provide a better fit for high frequency words. The Zipf-Mandlebrodt law is often parameterized by a discrete probability distribution.

$$P(f_r|N, q, s) \propto (f_r + q)^{-s} \tag{3}$$

[2]The bag of words assumption can be relaxed in LDA with certain preprocessing steps. For example, LDA may be used to find topics in a skip-gram term co-occurrence matrix. Skip-grams inherently capture some properties of word order. LDA itself needs only count data.

3.1.2.2 Heap's Law

Heap's law describes the relationship between the number of unique words in a corpus and the total number of words in a corpus. It is typically parameterized by

$$V \propto N^c \tag{4}$$

where V is the number of unique words (vocabulary size), N is the total number of words, and c is a free parameter to be fit from data. Typically c is between 0.4 and 0.6.

3.1.2.3 Taylor's Law

3.1.2.4 Relationship Between Laws

3.1.3 Approach and Preliminary Results

3.2 Using the LDA-DGP to Simulate Corpora

3.2.1 Thesis

3.2.2 Simulation Studies and Topic Models

3.2.3 Approach and Preliminary Results

3.3 Lessons for LDA Model Specification

3.3.1 Thesis

3.3.2 Folklore and Heuristics

3.3.3 Approach and Preliminary Results

4 Study 2: Transfer Learning for LDA

4.1 Thesis

4.2 Background

4.3 Approach and Preliminary Results

5 Software: *tidylda*, an R Package

5.1 Background

- What other LDA packages exist and why do we need a new one
  - User friendly, speed, principled defaults
- "Tidyverse" and "tidy text mining": theoretical framework

5.2 Current State of *tidylda*

5.3 Anticipated Contributions

6 Anticipated 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.

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.

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.