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.

Yet in spite of these advances, machine learning for natural language processing 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

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

1.1 Latent Dirichlet Allocation

2 Study 1: Examining the LDA-DGP

3 Study 2: Transfer Learning for LDA

4 Software: {tidylda}, an R Package

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

5 References