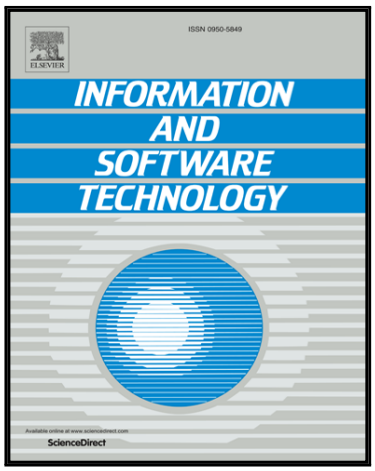

What is Wrong with Topic Modeling? (and How to Fix it Using Search-based Software Engineering)

Amritanshu Agrawal, Wei Fu, Tim Menzies

PII: S0950-5849(17)30086-1
DOI: 10.1016/j.infsof.2018.02.005
Reference: INFSOF 5959

To appear in: *Information and Software Technology*

Received date: 31 January 2017
Revised date: 8 November 2017
Accepted date: 19 February 2018

Please cite this article as: Amritanshu Agrawal, Wei Fu, Tim Menzies, What is Wrong with Topic Modeling? (and How to Fix it Using Search-based Software Engineering), *Information and Software Technology* (2018), doi: 10.1016/j.infsof.2018.02.005

What is Wrong with Topic Modeling?
(and How to Fix it Using Search-based Software Engineering)

Amritanshu Agrawal*, Wei Fu*, Tim Menzies*

Department of Computer Science, North Carolina State University, Raleigh, NC, USA

Abstract

Context: Topic modeling finds human-readable structures in unstructured textual data. A widely used topic modeling technique is Latent Dirichlet allocation. When running on different datasets, LDA suffers from "order effects", i.e., different topics are generated if the order of training data is shuffled. Such order effects introduce a systematic error for any study. This error can relate to misleading results; specifically, inaccurate topic descriptions and a reduction in the efficacy of text mining classification results.
Objective: To provide a method in which distributions generated by LDA are more stable and can be used for further analysis.
Method: We use LDADE, a search-based software engineering tool which uses Differential Evolution (DE) to tune the LDA's parameters. LDADE is evaluated on data from a programmer information exchange site (Stackoverflow), title and abstract text of thousands of Software Engineering (SE) papers, and software defect reports from NASA. Results were collected across different implementations of LDA (Python+Scikit-Learn, Scala+Spark) across Linux platform and for different kinds of LDAs (VEM, Gibbs sampling). Results were scored via topic stability and text mining classification accuracy.
Results: In all treatments: (i) standard LDA exhibits very large topic instability; (ii) LDADE's tunings dramatically reduce cluster instability; (iii) LDADE also leads to improved performances for supervised as well as unsupervised learning.
Conclusion: Due to topic instability, using standard LDA with its "off-the-shelf" settings should now be depreciated. Also, in future, we should require SE papers that use LDA to test and (if needed) mitigate LDA topic instability. Finally, LDADE is a candidate technology for effectively and efficiently reducing that instability.

Keywords: Topic modeling, Stability, LDA, tuning, differential evolution.

1. Introduction

The current great challenge in software analytics is understanding unstructured data. As shown in Figure 1, most of the planet's 1600 Exabytes of data does not appear in structured sources (databases, etc) [1]. Mostly the data is of *unstructured* form, often in free text, and found in word processing files, slide presentations, comments, etc.

Such unstructured data does not have a pre-defined data model and is typically text-heavy. Finding insights among unstructured text is difficult unless we can search, characterize, and classify the textual data in a meaningful way. One of the common techniques for finding related topics within unstructured text (an area called topic modeling) is Latent Dirichlet allocation (LDA) [2].

This paper explores systematic errors in LDA analysis. LDA is a non-deterministic algorithm since its internal weights are updated via a stochastic sampling process (described later in this paper). We show in this paper that this non-determinism means that the topics generated by LDA on SE data are subject to order effects, i.e., different input orderings can lead to different topics. Such instability can:

- Confuse users when they see different topics each time the algorithm is re-run.
- Reduce the efficacy of text mining classifiers that rely on LDA to generate their input training data.

To fix this problem, we propose LDADE: a combination of

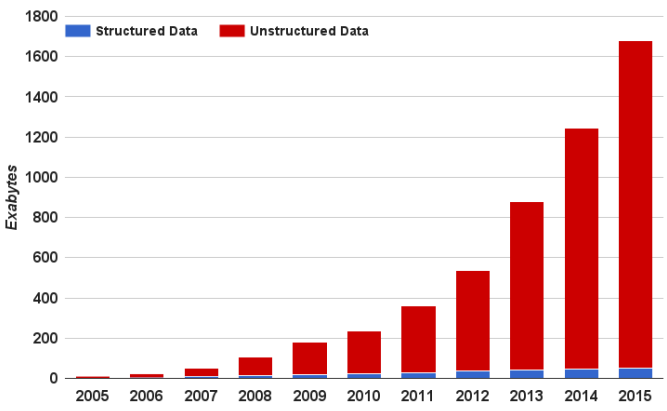

Figure 1: Data Growth 2005-2015. From [1].

*Corresponding author: Tel: +1-919-637-0412 (Amritanshu)
Email addresses: aagrawa8@ncsu.edu (Amritanshu Agrawal), wfu@ncsu.edu (Wei Fu), tim.menzies@gmail.com (Tim Menzies)
URL: https://amritag.wixsite.com/amrit (Amritanshu Agrawal)