

## STATISTICAL ANALYSIS AND DATA MINING

Original Article

## Conducting sparse feature selection on arbitrarily long phrases in text corpora with a focus on interpretability

Luke Miratrix✉, Robin Ackerman

First published: 18 July 2016

<https://doi.org/10.1002/sam.11323> About  Access

PDF Tools



Share

## Abstract

We propose a general framework for topic - specific summarization of large text corpora, and illustrate how it can be used for analysis in two quite different contexts: an Occupational Safety and Health Administration (OSHA) database of fatality and catastrophe reports (to facilitate surveillance for patterns in circumstances leading to injury or death), and legal decisions on workers' compensation claims (to explore relevant case law). Our summarization framework, built on sparse classification methods, is a compromise between simple word frequency - based methods currently in wide use, and more heavyweight, model - intensive methods such as latent Dirichlet allocation (LDA). For a particular topic of interest (e.g., mental health disability, or carbon monoxide exposure), we regress a labeling of documents onto the high - dimensional counts of all the other words and phrases in the documents. The resulting small set of phrases found as predictive are then harvested as the summary. Using a branch - and - bound approach, this method can incorporate phrases of arbitrary length, which allows for potentially rich summarization. We discuss how focus on the purpose of the summaries can inform choices of tuning parameters and model constraints. We evaluate this tool by comparing the computational time and summary statistics of the resulting word lists to three other methods in the literature. We also present a new R package, **textreg**. Overall, we argue that sparse methods have much to offer in text analysis and is a branch of research that should be considered further in this context. © 2016 Wiley Periodicals, Inc. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2016

About Wiley Online Library



Help &amp; Support



Opportunities



Connect with Wiley

