Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2025:1**
http://www.math.su.se

Full stop: a statistical analysis on the effect of stop words on the appreciation of classical literature

Morris Lundberg Allerholm*

February 2025

Abstract

This thesis explores the application of linear regression in computational text analysis. The subject of the analysis is the appreciation of literary classics and the primary details of interest are the frequencies of the most commonly used words in English: stop words. To develop the models, both the Lasso and best subset selection are implemented. Their evaluation reveals that there are connections between usage of stop words and general appreciation of classics. Some common words stand out as being especially important for the public's appreciation. The resulting models are also shown to have predictive ability.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: morrislual@gmail.com. Supervisor: Johannes Heiny and Xuechun Hu.