Contact: Måns Magnusson, STIMA. email: mans.magnusson@liu.se

1 Rare words in topic models

1.1 Introduction

Topic models has become a more and more of popular approach to unsupervised latent modeling of documents in the last ten years. In most application preprocessing of the corpus is done *ad hoc* to handle common problems with that occur in natural language. Stemming of tokens, removal of stop words and removal of rare words are all commonly done, but there has been little effort to study the effects of these approaches on the final models. Lately some studies has shown that the commonly assumed "truths" regarding preprocessing in the research community can be questioned. Such as the large negative effects of stop words on the topics, the need and benefit of stemming tokens and the general effect of corpora preprocessing decisions. [4, 2]

1.2 Purpose

This project will look into another preprocessing that is commonly done before estimating in topic models, the effect of removing rare words. Commonly this is done using either a rare word limit or using tf-idf weighting of terms. The project will study the effect of removing rare words using different methods for different known corpora and will evaluate the the effect on models by removal using multiple approaches to topic model evaluation. Some theoretical work on the effect of conditional on non-rare words may also be done if this is of interest.

References

[1] Dhrumil Mehta Al Johri, Eui-Hong (Sam) Han. Domain specific newsbots, live automated reporting systems involving natural language communication. 2016.

[2] Matthew James Denny and Arthur Spirling. Assessing the consequences of text preprocessing decisions. *Available at SSRN 2849145*, 2016.

[3] Måns Magnusson, Jens Finnäs, and Leonard Wallentin. Finding the news lead in the data haystack: Automated local data journalism using crime data. 2016.

[4] Alexandra Schofield and David Mimno. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300, 2016.

[5] Eirik Stavelin. Computational journalism. when journalism meets programming. 2014.