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

2 Topic summaries in topic models

2.1 Introduction

Topic models are an unsupervised approach to model textual corpora. The basic topic model define a topic as a dirichlet distribution over the vocabulary. This is often difficult to visualize and commonly the types with the highest probability in each topic are shown as a proxy for the topic.

This approach has many difficulties in that only relatively common words can be the top words in larger corpora. This approach also do not take the fact into account that a type can be more or less discriminative over the different number of topics. This fact has sparked a number of suggestions on how to present topics using different reweighing methods. Unfortunately no rigorous evaluation has been done to see which approach that ends up being the best way of visualizing a given topic.

2.2 Purpose

The purpose of this project is to evaluate different reweighing schemes for topic models to visualize topics for corpora. The main focus will be evaluations using terms or multiple terms. Different evaluation metrics will be used to evaluate the best approaches to visualize individual topics. Potential suggestions on new approaches by the students can also be done and evaluated.

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.