Master Thesis project: Evaluating visualization techniques for topic models

November 14, 2016

Principal: Måns Magnusson (mans.magnusson@liu.se)

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 Purpose

The purpose of this project is to evaluate different reweighing schemes to visualize topics for a given corpus (or multiple corpora). The main focus will be evaluations using terms or multiple terms that summarize a given topic.

Different evaluation metrics and methods of topic models 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.