Outline for Senior Thesis

Ziv Epstein
ziv.epstein@pomona.edu

October 4, 2016

1 Introduction to Topic Modeling

This section provides a overview of topic modelling as a field.

2 Non Negative Matrix Factorization

2.1 How NNMF solves Topic Modeling

We begin with the two foundation papers introducing the NNMF concept and standard algorithms. Lee and Seung [6] were the first to introduce this idea, and to propose that topic modelling could be thought of as a matrix factorization problem. Ho [4] expands on this notion by elaborating on optimization schemas and corresponding algorithms that we will be taking advantage of.

2.2 Hierarchical Topic Modelling

We then consider the work of Griffiths and Tenenbaum [3], who extened the notion of topic models to a hierarchical domain. We aim to replicate this structure but using an NNMF implementation instead of Latent Dirlichet Allocation (LDA).

3 Algorithms

We then discuss methods for using NNMF to generate a hierarchical topic model.

3.1 Single Linkage Graph Construction

Our method to do is construct a distance matrix from the topic repersentations generated from NNMF. With this graph, we will employ community detection algorithms [2] to build the hierarchy. In particular, there already exist methods for generating a hierarchical community structure within complex networks [5] that we will take advange of to build our topic model.

3.2 Deep Semi NMF

We then consider the algorithms and notions in the domain of Deep Semi NMF as a model for Hierarchical NMF [1, 7]

4 Visualization

In this brief section, I will motivate the visualization of hierarchical topic modeling.

5 Hierarchical Topic Model

We will then construct our own model for learning a hierarchy of topics within the model itself. This will be a blend of the Single Linkage graph construction model and the Deep Semi NMF.

6 Results

We will then apply the method of our Hierarchical Topic Model to several datasets, and compare our results with previous work.

6.1 Synthetic Data

We will generate synthetic data with hierarchical topics and verify that our algorithm succesfully extracts them

6.2 Standard Data

We will run our algorithm on data cannonically associated with this task. For our purposes, the 20 News Group Data set will probably be sufficient.

6.3 Afghan Data

A collaborator at NYU Abu-Dhabi has hand-curated a dataset of Afghani magazines, which is notable for sociologists. We will run our algorithm on this data set and see what happens

7 Discussion/Conlusion

Here we will conclude by situating this work within the field, compare its results with similar methods and propose future work.

References

[1] Jennifer Flenner and Blake Hunter. A deep nonnegative matrix factorization.

[2] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[3] DMBTL Griffiths and MIJJB Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16:17, 2004.

[4] Ngoc-Diep Ho. *Nonnegative matrix factorization algorithms and applications*. PhD thesis, ÉCOLE POLYTECHNIQUE, 2008.

[5] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[6] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[7] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Bjoern Schuller. A deep semi-nmf model for learning hidden representations. In *ICML*, pages 1692–1700, 2014.