Hierarchical Topic Modelling
Mathematics Senior Thesis

Ziv Epstein
ziv.epstein@pomona.edu

November 11, 2016

Contents

1 Introduction to Topic Modeling

Here we introduce the notion of topic modeling, cite previous work [6] [4] and talk about its relation to society. As you can see from first draft, this is mostly done already.

2 Non Negative Matrix Factorization

2.1 How NNMF solves Topic Modeling

Talk about the linear algrebraic intuition behind NNMF and how it solves topic modeling. This is also already done.

2.2 Hierarchical Topic Modellng

Introduce the core idea of the thesis, which is extending the ideas of NMF to the hierarchical domain. Give some examples, in the vein of [3]

3 Algorithms

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

3.1 Single Linkage Graph Construction

This is the algorithm we invited. My first draft already fully explores this.

3.2 Semi-Supervised NNMF

Here, I will fully describe the ideas and implementation details of SSNMF. I will talk about the intuition behind it and how it introduces notions of SVMs into NNMF.

3.3 Deep Semi NMF

Here I will give a brief overview of the notion of neural networks as a generalization of nested matrix factorization. I will describe backpropogation as an optimization scheme and compare it with other methods. I will this describe two works that use look at deep NMF, one of which is motivated by these neural networks, the other of which is motivated by more traditional techniques.

4 Visualization

Here I will motivate the visualization of this data, provide some heuristics, survey the field of hierarchical data visualization and propose my alterations.

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/Conclusion

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] Hyekyoung Lee, Jiho Yoo, and Seungjin Choi. Semi-supervised nonnegative matrix factorization. *IEEE Signal Processing Letters*, 17(1):4–7, 2010.

[8] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

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