Thematic Hierarchies for Knowledge Discovery in Text

Akshay Peshave, Tim Oates

Dept. of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
{peshave1, oates}@umbc.edu

U.S. SEMANTIC TECHNOLOGIES SYMPOSIUM

March 2018
Ohio, USA

Introduction

Automated Knowledge Discovery [1] is an active area of research that seeks to address the need for extensive knowledge acquisition and elicitation, curation and archival for large quantities of text. A generic, flexible and extendable text analytics framework is based on robust theme detection methods. A novel method is described here to extract thematic hierarchies using the Latent Dirichlet Allocation (LDA) [2] topic models, noun-phrase extraction and phrase filtering heuristics. Further, a visual representation of theme dynamics, the "Document Thematic Map (DTmap)", is created to enable text segmentation [3, 4] using the theme-mix.

Thematic Phrases Extraction

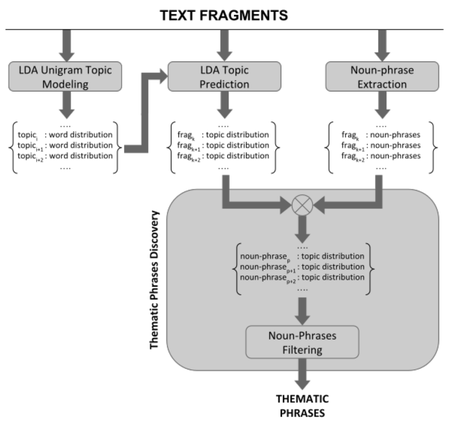

Figure 1: Thematic Phrases Discovery

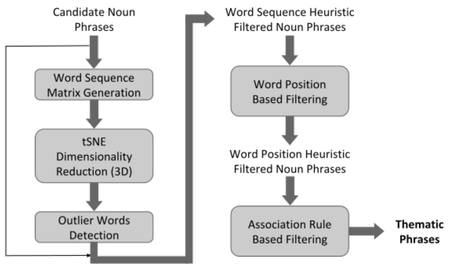

Figure 2: Noun-phrases Filtering

Noun-phrases Filtering Heuristics

Figure 3: Word Sequence Matrix Generation

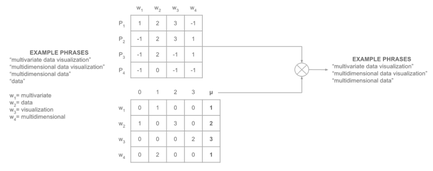

Figure 4: Word Position Based Phrase Filtering

Extracted Thematic Phrases: Examples

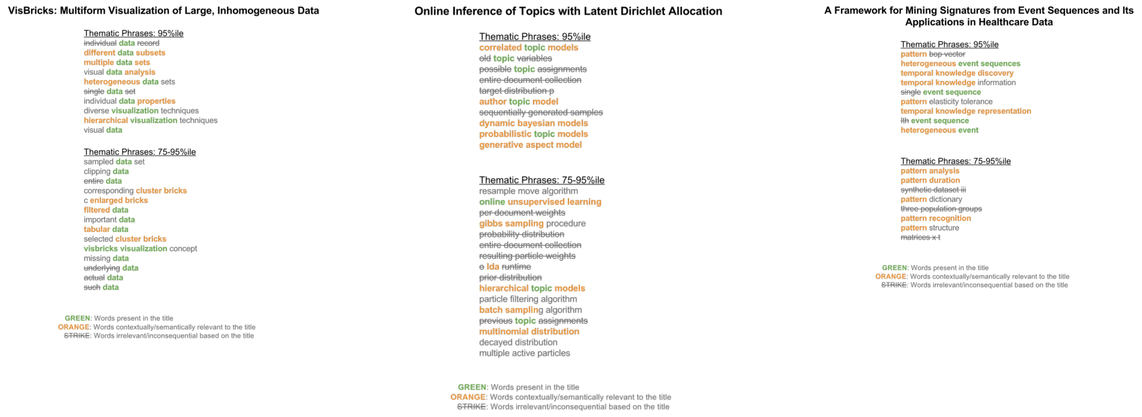

Figure 5: Six Handpicked Thematic Phrases

Document Thematic Maps

A DTmap is an image of a document consisting of a series of NxN pixels blocks. Each block represents a sentence in the document. Block colorings represent the corresponding sentence's thematic mix. DTmaps may be used for tasks such as text segmentation and visual analysis of theme dynamics in text.

Figure 6: All Thematic Phrases

Thematic Coverage and Precision

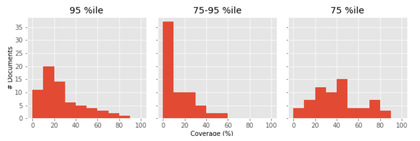

Figure 7: Abstract Word Coverage

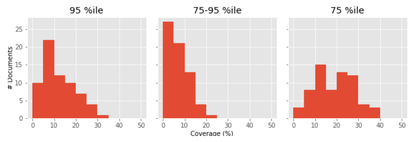

Figure 8: Abstract Noun-phrase Coverage

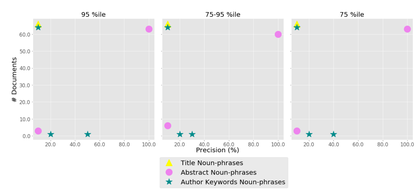

Figure 9: Thematic Phrases Precision

Future Work

Preliminary evaluation of the method shows good and precise coverage of the thematic basis of input text. Further evaluation of the quality of thematic phrases is planned using:

- ROUGE score, by providing the phrases as input to text summarization methods [5].
- Comparison of DTmaps with other topic model based text segmentation methods.
- Qualitative assessment of thematic phrases using subject matter expert surveys.

Construction of a knowledge graph comprising the thematic hierarchy using detected themes and text segments detected in corresponding DTmaps for large text corpora is ongoing. WORDNET integration is planned to allow for semantic similarity measures for thematic phrase candidates.

References

[1] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," *International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 112–117, 1995.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993–1022, 2003.

[3] M. Riedl and C. Biemann, "Text segmentation with topic models," *Journal for Language Technology and Computational Linguistics*, vol. 27, no. 1, pp. 47–69, 2012.

[4] J. F. Canny, T. L. Rattenbury, and C. Sciences, "A Dynamic Topic Model for Document Segmentation," *Electrical Engineering*, 2006.

[5] A. Nenkova and K. McKeown, "Automatic Summarization," *Foundations and Trends® in Information Retrieval*, vol. 5, no. 3, pp. 235–422, 2011.