**MONTAGUE — Exploring**

<1 figure>

**Exploring Large Datasets with Topic Model Visualizations**

Montague, J., Simpson, J., Rockwell, G., Ruecker, S. and Brown, S.

How can topic model visualizations aid in the exploration of large datasets? The standard method for visualizing topic models involves a two-stage process that begins with the use of a tool specifically designed to construct topic models on a specified corpus. MALLET (McCallum, 2002), a command line tool that implements the Latent Dirichlet Allocation (LDA) algorithm described by Blei et al. (2003), provides output in the form of text files that can be shared directly with other tools and seems to be the default tool of choice for the first stage. Regardless of the topic modeller used, its output is then passed to a separate tool for the second stage in the production of the visualization. These separate tools are either general visualization tools or topic model specific visualization tools. The use of general visualization tools is attractive because it typically requires a lower technical competency and allows the user to leverage skills they already have, but general tools provide fewer targeted affordances for exploration. Still, despite the existence of at least a dozen tools that have been used to visualize topic models, it remains the case that none of them are particularly well suited to exploring large datasets comprising thousands or millions of entries in a corpus distributed over a wide range of time and with a correspondingly large set of candidate topics produced by topic modelling.

This paper provides a solution to this problem by describing a tool that has been designed specifically to address this problem and that is now entering production. Before describing the tool and providing a short demonstration of its functionality, we summarize the current landscape of topic model visualizations and provide recommendations based on various use-cases.

**Topic Modelling Visualizations**

The visualization methods used across these tools can be fruitfully categorized using a set of modified categories drawn from work done by Katifori et al. (2007) on ontology visualizations:

• **Traditional charts** (Bar, line, pie, scatterplot, etc.): D-VITA (Gunnemann et al., 2013); Tableau (Sharkey and Ansari, 2014); ParallelTopics (Dou et al., 2011); NIPS (Iwata et al., 2008); MetaToMATo (Snyder et al., 2013).

• **Network graphs**: Gephi (Chen et al., 2013); Topicnet (Gretarsson et al., 2012).

• **Zoomable tools**: an unnamed tool by Chaney and Blei (2012).

• **2D matrix**: Serendip (Alexander et al., 2014); Termite (Chuang et al., 2012).

The variety of visualization techniques being pursued are evidence of the difficulties related to displaying relevant information from topic modelling, as each visualization method has been developed or chosen with a particular application in mind. The tools that particularly focus on exploring the underlying data and revealing connections are Serendip and MetaToMATo, neither of which is a graph-based visualization, and both of which become overwhelmed by large datasets. Gephi is a general-purpose graph-based visualization tool that has been used to visualize large datasets (Munster, Jockers), but the images it produces are static and it does not easily lend itself to sharing information derived from topic models. Drawing on lessons from a review of all these tools, but Serendip, MetaToMATo, and Gephi in particular, the tool we report on constructing is targeted at exploration using graph-based visualizations of topic models on large corpuses.

**Topic Modelling Philosophy Journals**

Motivation for this project has been brought about in part by work done over the past year with a corpus of philosophy journals acquired from JSTOR (Simpson et al., 2014). While investigating this corpus we made use of chart-based visualizations with R after preprocessing with MALLET. While there was much to be learned from the changing prevalence of particular topics over time, there were also a number of things that the visualizations were not able to tell us. In particular, our first visualizations made it difficult to easily see connections between the different topics, to see relations between the different journals, and to have this information present itself with increasing detail as required for the investigation. This challenge arose in large part because the dataset covered 27,536 journal articles from 10 different journals published between 1876 and 2008. Datasets of this size and diversity are becoming increasingly common, and so we moved to pursue a tool capable of visualizing the results of topic modelling. Not finding a suitable one for the exploration of large datasets, we set about designing a new one.

**Visualizations in 3D**

MacEachren et al. (1994) suggest that all visualizations are meant to principally perform one of the following tasks: clearly present previously discovered information, analyze a particular dataset, combine multiple datasets, or explore data for knowledge discovery. It is the last category on which our visualization research team has been particularly focused: visualizations that allow the user to not simply understand or analyze a dataset, but to explore it in hopes of uncovering new information.

As Lauren Frederica-Klein says in her 2013 digital humanities start-up grant application for the Interactive TOpic Model and MEtadata Visualization (TOME) project*,* regarding Wise et al.’s Galaxies visualization, ‘By forcing all thematic differences into a single two-dimensional presentation, information is inevitably lost’. In an effort to minimize this information loss, we are now creating an exploration tool for the purpose of experiencing a topic-modeled corpus in 3D space, using the free, open-source JavaScript framework *Famo.us*.

The *Famo.us* framework is designed to help ‘create smooth, complex UIs for any screen’, and will provide us some substantial benefits, first and foremost being that its CSS-like styling code is relatively straightforward to learn and implement. Additionally, using physics-based animations, three orientation axes, and opacity control, it is capable of rendering a touch-screen manipulable, three-dimensional environment, the interactivity afforded by which is something we believe will help greatly foster user exploration. As explained by Card et al. (1999), ‘This additional [3rd] dimension projects from the viewpoint toward infinity, creating a large visible workspace’, a decidedly beneficial quality when visually exploring a large dataset.

Chuang et al. (2012) describe the illegible results of experimenting with topic modelling visualizations where text is displayed directly in the visualization matrix. To reduce the volume of unwanted information, our design implements Healey’s (1996) notion of ‘knowledge discovery’, allowing the user to *filter* unwanted data, and what Petre and Green (1992) describe as ‘secondary notation; the use of layout and perceptual cues which are not formally part of the notation (elements such as adjacency, clustering, whitespace, labeling and so on)’. Like Healey, we believe this will help ease any display or perception bottleneck and afford the user a unique opportunity to discover and explore unknown trends and relationships in the data. The user will customize the visibility of elements of the dataset and secondary notation by adjusting a series of controls to make alterations to the sensitivity of, for instance, clustering algorithms, labeling, or the visibility of edges in a network.

Since any single visualization is fundamentally limited in its ability to communicate very complex relationships, like Andrew Goldstone’s DfR browser (2013), and Snyder et al.’s Metadata and Topic Model Analysis Toolkit (MetaToMATo) (2013), our concept merges multiple visualizations into what Snyder et al. term ‘a single faceted browsing paradigm for exploration and analysis of document collections’. Word clouds generated via topic modelling, histograms, line graphs, and network diagramming are all experienced via a zoomable user interface (ZUI). Snyder et al. (2013) explain that ‘effectively exploring and analyzing large text corpora requires visualizations that provide a high level summary’. Figure 1shows how the viewer will be able to move from that high-level summary, ‘flying’ deeper and deeper into the data, simultaneously experiencing multiple visualizations from unique vantage points, and, as espoused by Chuang et al. (2012), ‘drilling down’ into additional layers of information as desired.



Figure 1.As the user navigates deeper into the dataset, more and different relationships are revealed.

**References**

**Alexander, E., Kohlmann, J., Valenza, R., Witmore, M. and Gleicher, M.** (2014). Serendip: Topic Model-Driven Visual Exploration of Text Corpora, https://graphics.cs.wisc.edu/Papers/2014/AKVWG14/Preprint.pdf.

**Blei, D. M., Ng, A. Y. and Jordan, M. I.** (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* **3**: 993–1022.

**Card, S. K., Mackinlay, J. D. and Shneiderman, B.** (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco.

**Chaney, A. J.-B. and Blei, D. M.** (2012). Visualizing Topic Models. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, 4–7 June 2012, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/download/4645%26lt%3B/5021.

**Chen, A. T., Sheble, L. and Eichler, G.** (2013). Topic Modeling and Network Visualization to Explore Patient Experiences. *Visual Analytics in Health Care*, Washington, DC, 16 November 2013, http://ruby.ils.unc.edu/~atchen/pubs/Chen\_Sheble\_Eichler\_VAHC2013.pdf.

**Chuang, J., Manning, C. D. and Heer, J.** (2012). Termite: Visualization Techniques for Assessing Textual Topic Models. Capri Islands, Italy, http://vis.stanford.edu/files/2012TermiteAVI.pdf.

**Dou, W., Wang, X., Kraft, T. and Ribarsky, W**. (2011). Identifying Topical Trends in Social Media with Topic Modeling. University of North Carolina, Charlotte. http://vialab.science.uoit.ca/textvis2011/papers/textvis%202011-dou.pdf.

**Frederica-Klein, L.** (2013). TOpic Model and MEtadata Visualization Project (TOME), digital humanities startup grant application.

**Goldstone, A.** (2013). drfbrowser: Take a MALLET to Literary History, http://agoldst.github.io/dfrbrowser/web.

**Gretarsson, B., O’Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D. and Smyth, P.** (2012). Topicnets: Visual Analysis of Large Text Corpora with Topic Modeling. *ACM Transactions on Intelligent Systems and Technology* (TIST), **3**(2): 23.

**Günnemann, N., Derntl, M., Klamma, R. and Jarke, M.** (2013). An Interactive System for Visual Analytics of Dynamic Topic Models. *Datenbank-Spektrum,* **13**(3): 213–23.

**Healey, C.** (1996). Effective Visualization of Large Multidimensional Datasets. Department of Computer Science Thesis, University of British Columbia.

**Iwata, T., Yamada, T. and Ueda, N.** (2008). Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, 24–27 August 2008, pp. 363–71, http://dl.acm.org/citation.cfm?id=1401937.

**Jiao, Z. L., Liu, Q., Li, Y.-F., Marriott, K. and Wybrow, M.** (2013). Visualization of Large Ontologies with Landmarks. In *Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications*, Barcelona, pp. 461–70, http://www.csse.monash.edu.au/~mwybrow/papers/jiaoivapp2013.pdf.

**Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C. and Giannopoulou, E.** (2007). Ontology Visualization Methods Survey. *ACM Computing Surveys,* **39**(4) (2 November): 10–es, doi:10.1145/1287620.1287621.

**MacEachren, A. M., Bishop, I., Dykes, J., Dorling, D. and Gatrell, A.** (1994). Introduction to Advances in Visualizing Spatial Data. In *Visualization in Geographical Information Systems.* London: John Wiley & Sons, pp. 51–59.

**McCallum, A. K.** (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

**Petre, M. and Green, T. R. G.** (1992). Learning to Read Graphics: Some Evidence That ‘Seeing’ an Information Display Is an Acquired Skill. *Journal of Visual Languages and Computing,* **4**: 55–70.

**Sharkey, M. and Ansari, M.** (2014). Deconstruct and Reconstruct: Using Topic Modeling on an Analytics Corpus. In *LAK Workshops*, http://ceurws. org/Vol1137/ lakdatachallenge2014\_submission\_1.pdf.

**Simpson, J., Rockwell, G. and Sinclair, S.** (2014). The Epidemiology of Ideas. *Canadian Society for Digital Humanities Annual Conference*, Brock University, May 2014.

**Snyder, J., Knowles, R., Dredze, M., Gormley, M. and Wolfe, T.** (2013). Topic Models and Metadata for Visualizing Text Corpora. *Proceedings of the NAACL HLT 2013 Demonstration Session*, Atlanta, 10–12 June 2013, pp. 5–9.