Title:

Topic Model as Hermeneutic

Authors:
(in alphabetical order)

Sayan Bhattacharyya[1]
bhattach@umich.edu
School of Information,
University of Michigan, Ann Arbor

Travis R. Brown
trbrown@umd.edu
Maryland Institute of Technology in the Humanities,
University of Maryland, College Park

Clay Templeton
clayt@umd.edu
College of Information,
University of Maryland, College Park

Mailing address for communication (note: email strongly prefererred):

Maryland Institute of Technology in the Humanities (MITH),
University of Maryland
B0131 McKeldin Library
College Park, MD 20742

[1] Work done while at the Maryland Institute of Technology in the Humanities, University of Maryland, College Park

Abstract

A tradition exists in the machine learning and knowledge representation communities of reflection about the epistemological status of techniques in these fields (Brachman 1985, Frank et al. 2005). We follow in this tradition to consider probabilistic topic modeling, a relatively recent statistically-based machine learning technique (Steyvers & Griffiths 2007) that has achieved growing use in the analysis of large text corpora for automated discovery of "topics" (thematic patterns and categories) in the text, and has been adopted as a technique for "distant reading" (Moretti 2005, Nelson 2011) in the digital humanities. Topic modeling is an inductive inference mechanism (Landauer & Dumais 1997) that can take us far beyond the surface data of the text (Tenenbaum 2011). We present preliminary results from our ongoing investigations in topic modeling applied to a corpus of historical newspaper articles. We use our results to inquire what the epistemological status of topic modeling (and similar probability-based inductive approaches) may be, in terms of their being "semantically meaningful" (Chang et al., 2009) and their constituting an interpretive method or a "digital hermeneutic" (Paterson 2011). Interpretation of texts using quantitative techniques is sometimes subject to accusations of positivism and reductionism (Robinson 2009). We argue that topic modeling and related approaches may not be subject to this criticism. We suggest that Hans-Georg Gadamer's idea of a hermeneutic that moves beyond the mere adoption of the viewpoint of the text itself as well as the instrumental use of the text to recover supposed authorial intentions or affirm the reader's own subjectivity (Gadamer 1989) is a useful way to think about these approaches.

References:

Brachman, Ronald J. 'On the Epistemological Status of Semantic Networks'. In: *Readings in Knowledge Representation.* (Eds.) Brachman, R.J. and H.J. Levesque. Los Altos, CA: Kaufmann (1985), pp. 191-215.

Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish & David M. Blei. 'Reading Tea Leaves: How Humans Interpret Topic Models.' In *Proceedings of the Neural Information Processing Systems Conference* (NIPS). 2009.

Frank, Ulrich, Flaminio Squazzoni & Klaus G. Troitzsch. 'Epistemological Perspectives on Simulation: An Introduction'. In *Epistemological Aspects of Computer Simulation in the Social Sciences: Proceedings of the Second International Workshop, EPOS 2006, Brescia, Italy, October 5-6, 2006*. Berlin; Heidelberg : Springer Berlin Heidelberg. Lecture Notes in Computer Science. 2009. Vol. 5466, pp. 1-11.

Gadamer, Hans-Georg. *Truth and Method*. (Translation of *Wahrheit und Methode* (1960). Trans. Weinsheimer, Joel and Donald G. Marshall. 2nd rev. ed.) New York : Crossroad. 1989.

Landauer, Thomas K. & Susan T. Dumais. 'A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge'. *Psychological Review*, 1997, pp. 211-240.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History.* London; New York: Verso. 2005.

Nelson, Robert K. 'Of Monsters, Men — And Topic Modeling'. Opinionator. *The New York Times*. May 29, 2011. http://opinionator.blogs.nytimes.com/2011/05/29/of-monsters-men-and-topic-modeling/ . Retrieved July 18, 2011.

Paterson, Duncan. 'Digital Hermeneutics'. In Workshop on Digital Humanities and the Mechanics of Knowledge Production, Association for Asian Studies (AAS) and International Convention of Asia Scholars (ICAS) Joint Conference, Honolulu, 2011.

Robinson, Peter. 'What Text Really is Not, and Why Editors Have to Learn to Swim'. *Literary and Linguistic Computing*. 2009. Vol. 24, No. 1, pp. 41-52.

Steyvers, Mark & Thomas L. Griffiths, 'Probabilistic Topic Models'. In *Latent Semantic Analysis: A Road to Meaning.* In (Eds.) Landauer, T., D. McNamara, S. Dennis, and W. Kintsch. Mahwak, NJ: Lawrence Erlbaum. 2007.

Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths & Noah D. Goodman. "How to Grow a Mind: Statistics, Structure, and Abstraction". *Science*, 11 March 2011. Vol. 331, No. 6022. pp. 1279-1285.