14.
Data Do Not Speak for Themselves: Interpretation and Model Selection in Unsupervised Automated Text Analysis

Juho Pääkkönen

1. Introduction

Methods of automated text analysis have become widespread in the emerging field of digital humanities. A large part of the newly available sources of big data consist of massive collections of text. Examples range from digitized literary corpora (Hoover et al.) to collections of political texts (Grimmer and Stewart) and social media communication data (Schwartz and Ungar). Automated text analysis holds the promise of unleashing the potential of such datasets, by enabling their systematic quantitative analysis in a cost-efficient manner (cf. King).

Among humanities scholars, automated methods have met with both enthusiasm and skepticism. In the context of composition and rhetoric, the hope is that automated methods would help implement with large datasets what Richard Haswell has called RAD research—or, replicable, aggregable, and data-supported research (Lang and Baehr 174-176). According to Haswell, RAD is "inquiry that is explicitly enough systematized in sampling, execution, and analysis to be replicated; exactly enough circumscribed to be extended; and factually enough supported to be verified" (201). Scholars in composition and rhetoric have adapted these principles, for instance, by formulating research designs which facilitate replication and comparison with future studies (Licastro), and establishing public databases with accessible tools for data-driven exploration (Miller et al.) Through extending the level of rigor requisite of RAD research to working with large datasets, Lang and Baehr proclaim that automated methods could

help give research in composition and rhetoric a firmer empirical grounding, lending the fields with increased credibility (175-176).

Hoffman and Waisanen note that, despite their usefulness, applications of automated text analysis have been rare in rhetoric (178-179). A possible reason for this might be that computational analysis is regarded as being "too objective or deterministic" (Clement) to be applicable in research fundamentally oriented towards interpretation (for similar observations see the introduction to this volume by Miller and Licastro). This attitude has been argued against in the context of literary studies by Tanya Clement and Amy Earhart, who emphasize that automated methods are never objective nor neutral, being implemented as computational tools that are the products of human design. Consequently, rather than dismissing automated methods as incompatible with interpretive research, Earhart argues that we should instead turn our attention to the "human element" involved in their design and use (Earhart).

In this chapter, I scrutinize the use of a particular kind of method in interpretive research, namely, text analysis based on unsupervised machine learning. Unsupervised learning is regarded as particularly useful for text analysis, because it is thought to enable the automatic discovery of meaningful patterns in vast collections of text, with minimal input on the part of the researcher (e.g. Mohr and Bogdanov 546). As such, unsupervised learning is claimed to provide an objective method for exploring large text corpora, which does not rely on imposing *a priori* interpretive schemes on the analyzed texts (Lee and Martin). My aim is to criticize this account, by arguing that it downplays the role of subjective interpretation in how unsupervised learning is often used to model text data. As an alternative view, I propose that unsupervised text analysis should be treated as crucially guided by interpretation, even when the process seems to be based on data-driven detection of patterns in text collections.

I begin in Section 2 by introducing the account of unsupervised text analysis as objective exploration. Section 3 then looks at an example of unsupervised text analysis method—topic modeling—which has recently become popular in the digital humanities (cf. Meeks and Weingart). Basing on this example, in Section 4 I argue that subjective judgments of result *interpretability* have a crucial role in guiding the topic modeling process, as the method is often used in the digital humanities. Yet topic modeling tends to be construed as providing factual information about meaning patterns in text data (e.g. Jockers, Jockers and Mimno). Section 4 discusses this disparity, and provides suggestions of how the results of topic modeling should be conceived from the point of view of responsible research. Section 5 concludes.

2. Unsupervised Text Analysis and Meaningful Pattern Detection

A central methodological problem in big data analytics is the detection of *meaningful patterns*. As data become more voluminous, the number of possible relationships between variables in the data simultaneously grows. Consequently, distinguishing superfluous relationships from patterns of real added value becomes increasingly difficult (Calude and Longo). This problem of identifying "small" but meaningful patterns in massive datasets is what Luciano Floridi has referred to as the "real, epistemological problem with big data" (436).

In the context of text analysis, sophisticated methods based on machine learning have been developed for pattern detection. The process of automated text analysis with machine learning can be roughly described as follows. The data to be analyzed consists of a set of *documents*, such as digitized student writings (Burstein et al.), newspaper articles (Block), literature (Jockers, Tangherlini and Leonard), and so on. In order to automatically analyze the data, the documents are first *preprocessed* to be amenable for computational analysis. A computer is then programmed to look at *features* in the document contents, such as words, word

collocations, or sentences. These features are used to determine a *classification* of the documents into different categories, based on observed similarities and dissimilarities between the documents. For instance, newspaper articles can be classified according to political orientation, by examining their use of political vocabulary. To take another example, student essays can be partitioned according to thematic content, by examining which words frequently occur together in them. Finally, the results of automated analysis will always have to be *evaluated* and *interpreted*, and the assigned interpretations *validated*.

In this chapter, my focus will be on the evaluation and interpretation of results. While I cannot comment on other crucial parts of the process, such as preprocessing and validation, I refer the reader to Grimmer and Stewart, who provide a comprehensive discussion on all parts of automated text analysis.

Broadly speaking, automated text analysis based on machine learning can be divided to *supervised* and *unsupervised* methods. Both kinds of methods seek to classify documents into categories, based on features observed in them. However, the two approaches differ in how they accomplish this goal, the most important difference here pertaining to how interpretation on part of the researcher enters the process.

In supervised learning, documents are classified according to a classification scheme, which has been pre-specified by the researcher. The machine learning algorithm is first given a relatively small set of documents, which have been manually labeled according to the pre-specified classes. Using this *training data* set, the algorithm records how features in the documents are related to their manually assigned class labels. The result of this process is called a *model* of the data. After having been created using the smaller training data set, the model can be used to classify documents in the full, unlabeled dataset by observing their features.

By contrast, in unsupervised learning, document classification is not based on a scheme pre-specified by the researcher, but is rather determined *inductively*, by examining similarities and differences between features occurring in the documents. For instance, documents can be classified into the same category because their vocabulary is very similar, or because some groups of words frequently occur together in them. Unsupervised learning, too, yields a model of the data. However, the model is constructed by directly examining features in the full data set, using no a priori information about the relationship between features and the document classes to be determined. To distinguish between supervised and unsupervised classification, the latter is often referred to as *clustering*.

Because of their inductive working, unsupervised methods are regarded as particularly useful for detecting meaningful patterns in text data. The idea here is that *a priori* interpretive schemes limit the range of information retrieved to the pre-specified categories, which might not suffice to capture all potentially interesting patterns present in the data (e.g. Tangherlini and Leonard 728-729, Jockers 120-124). Furthermore, at worst, a priori classification schemes can force documents into fixed categories, thus distorting the discovered patterns and information gleaned from them (Biernacki). Unsupervised learning, the argument goes, is able to bypass these problems by avoiding *a priori* assumptions and grounding analysis directly on features actually present in the data (Lee and Martin). As such, unsupervised methods promise to analyze data by *revealing patterns* rather than *imposing interpretation* on data, enabling the exploration of massive text corpora in a seemingly more objective, data-driven fashion (cf. Block and Newman 83).

The above account does not imply that unsupervised text analysis would be able to do away with interpretation altogether. Rather, the relevant difference between supervised and

unsupervised analysis has to do with *when* interpretation comes to play in the process. Whereas in supervised learning, analysis begins with the development of an interpretive scheme, which is then used to code the data, in unsupervised learning the process is inverted. As Mohr and Bogdanov put it, in unsupervised analysis one first counts features in the documents, and only then begins to interpret the results. According to them,

> One counts, and then one begins to interpret...interpretation is still required, but from the perspective of the actual modeling of the data, the more subjective moment of the procedure has been shifted over to the post-modeling phase of the analysis. (Mohr and Bogdanov 560)

Neither does unsupervised learning make deep understanding of the analyzed corpus unimportant (Mohr and Bogdanov 560). However, Mohr and Bogdanov argue that unsupervised text analysis constitutes a "fundamental shift in the locus of methodological subjectivity" (Mohr and Bogdanov 561), which frees the researcher from having to guess what kinds of meaning patterns are to be expected as results of the analysis (Mohr and Bogdanov 561-564). By applying unsupervised methods, the researcher can instead "count" features in the texts to reveal potentially interesting patterns of meaning. Then, interpretations can be grounded on what is revealed in the data, rather than *a priori* hypothesizing. On this view, modeling data comes first, and interpretation second.

In what follows, I will argue that this view runs the risk of downplaying the role of interpretation in many cases of unsupervised text analysis. In humanistic research, judgments of result interpretability often have a crucial role in guiding the unsupervised modeling process. Overlooking these judgments can lead to portraying the results of unsupervised learning as factual and objective, although they depend on interpretation. In the next two sections, I will argue that judgments of interpretability, when at play, should be recognized as a part of how text

data are modeled. To discuss these issues in more detail, it is best to start by looking at an example of an unsupervised text analysis method.

3. Example: Topic Modeling

Topic modeling is an unsupervised method for discovering hidden thematic structure in large collections of text (Blei, "Probabilistic Topic Models"). Originally developed for use in information retrieval to facilitate the browsing of large collections of unstructured text, during the past decade topic models have become to be increasingly applied in digital humanities, for instance, to track thematic changes in newspaper articles (Block); identify prevalent topics in contemporary poetry (Rhody); distinguish between rhetorical and topical language in scientific writing (Séaghda and Teufel); and to track thematic changes in 19th century literature (Jockers, Jockers and Mimno).

The most often used variant of topic modeling in the digital humanities is known as latent Dirichlet Allocation (Blei, "Topic Modeling and Digital Humanities"). This method works by taking a set of documents as input and assuming that they are each made up of a fixed number of shared topics, which each document exhibits to a varying degree. The documents are then analyzed statistically to find groups of frequently co-occurring words. These groups of words constitute the topics, which are returned as a result of the analysis, along with information about the extent to which they are exhibited in each document (Blei, "Probabilistic Topic Models" 78-82).

Topic modeling requires that the user specifies the number of topics to be discovered in the documents. Aside from this decision and the choice of two other *hyperparameters* controlling the probability distributions used for modeling the texts (see e.g. Tangherlini and Leonard 732),

the method is unsupervised in that it requires no pre-specified information concerning content of the topics to be discovered.

To give an example of what topic modeling results look like, I fit an LDA model with five topics to the AssociatedPress dataset—an example dataset provided as part of *topicmodels*, a package for topic modeling in the R language (Grün and Hornik). Table 14.1 depicts the 15 most probable words occurring in each topic returned by the model.

Table 14.1. Results from LDA with five topics on the AssociatedPress dataset.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
government	percent	police	i	people
soviet	million	court	bush	two
united	year	state	president	officials
states	new	two	new	air
president	billion	case	house	city
party	company	law	going	miles
military	last	office	time	area
union	market	years	think	three
two	prices	federal	dukakis	fire
south	stock	attorney	campaign	center
people	york	drug	years	spokesman
official	business	mrs	get	health
war	oil	school	first	day
news	trade	told	dont	reported
east	late	judge	bill	report

These kinds of results are useful for exploring the thematic content of text documents, because sets of frequently co-occurring words easily lend themselves for interpretation in terms of shared semantic concepts (Blei, “Topic Modeling and Digital Humanities”). Topic modeling

results are often presented as word lists, such as the ones above, containing the most probable words in each topic. Such representation of the results is then examined to give interpretations to the generated topics. For instance, the first topic in Table 14.1 could be labeled "USA-Soviet relationship." Alternatively, the same topic could be labeled "Cold War." The assignment of topic labels is based on subjective evaluation, and thus one should always be open to debate any given interpretation. To facilitate interpretation of topic modeling results, useful visualization tools have been developed specifically for humanistic research (Klein et al.)

Topic modeling enables researchers to examine and interpret word co-occurrence patterns in large text corpora, and moreover to do so without having to pre-specify what themes they are looking for. While topic interpretations are based on subjective evaluation, the topics themselves are generated by examining word co-occurrence patterns in the documents. However, as we already saw above, the modeling process itself also involves decisions on the part of the researcher, for instance selecting the number of topics to be generated. It is precisely these decisions, I will argue in the next section, through which interpretation crucially influences the topic modeling process.

4. Model Selection in Topic Modeling

Choosing the appropriate number of topics is a much-discussed issue both in the computer science literature as well as in digital humanities (e.g. Greene, Blei, “Probabilistic Topic Models”, Jockers). The reason for this is that the decision has direct influence on results produced by topic modeling. As such, the issue is an instance of a broader problem in statistical modeling known as the problem of *model selection*. Any dataset exhibits a multiplicity of diverging patterns, which can be captured by modeling the data differently (McAllister). The

problem then is to come up with a criterion to decide, which patterns in the data should be modeled, and which left out of focus (McAllister).

To demonstrate how the contents of topics vary with the number of topics, Table 14.2 presents the top 15 words in topics generated by a run of LDA on the AssociatedPress dataset, with the number of topics set to ten. As can be seen by comparing these results with those depicted in Table 14.1, some parts of the topics generated by the different runs stay relatively similar. For instance, topic 1 in the 5-topic model has similar content with topic 1 in the 10-topic model. However, some of the words in these topics also appear in high ranks in the topic 6 of the 10-topic model. Nevertheless, neither pair shares exactly matching content. Which set of topics should we take to better represent the thematic structure in the data?

Table 14.2. Results from LDA with ten topics on the AssociatedPress dataset.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
united	new	bush	percent	police	government	court	i	million	air
states	health	house	market	people	party	case	years	company	two
soviet	state	president	year	two	soviet	attorney	dont	billion	first
west	people	dukakis	prices	killed	president	state	like	year	time
military	city	campaign	new	city	political	judge	school	new	officials
east	water	committee	million	army	south	federal	get	workers	news
war	high	i	higher	three	people	law	people	money	three
american	center	bill	rose	fire	years	office	just	business	service
foreign	california	new	oil	died	national	drug	going	last	last
talks	national	senate	stock	man	news	charges	family	corp	days
president	study	congress	dollar	reported	members	trial	life	pay	spokesman
world	san	white	late	four	union	prison	think	plan	space
countries	medical	budget	price	authorities	leader	district	know	inc	wednesday
officials	number	republican	trading	home	gorbachev	department	say	federal	second
aid	officials	democratic	cents	night	rights	investigation	says	co	flight

Computer scientists have developed various quantitative metrics for comparing different topic models against each other. One popular approach measures how well each model is able to predict the words occurring in a set of documents that belong to the analyzed corpus, but were not used for training the models in question (e.g. Wallach et al.) Another approach is to evaluate the stability of the generated topics upon small changes in the dataset (Greene et al.) However, the aim in all quantitative model selection techniques is to single out one model as the best among all candidates (McAllister 884-885).

There are a number of problems with this approach. First, different quantitative metrics often indicate different models as the best candidate, and comparisons between them can be

challenging (McAllister). Second, the results from an experimental study by (Chang et al.) suggest that predictive accuracy in topic modeling can be negatively correlated with how easy it is to give modeling results a coherent semantic interpretation.

For these reasons, many authors have argued that in research aiming to understand text corpora by exploring patterns of meaning, the primary model selection criterion should be *topic interpretability* (e.g. Jockers, Grimmer and Stewart), loosely defined by Jockers as the "ease with which a human being can look at and identify" a topic (128). Following this line of thinking, Mohr and Bogdanov argue that the aim in topic modeling should be to select among different possible models "so that a best fit can be found between the number of topics and an overall level of interpretability" (560).

Now, my problem here is not with interpretability as a model selection criterion *per se*. On the contrary, I think it is only commonsensical to require that modeling results have semantic coherence in order for them to be useful in humanistic research. What I wish to argue against is the use of interpretability as a model selection criterion, *while* presenting topic modeling as an objective method for discovering meaning patterns in text data. If model selection is based on interpretability, then we cannot say that in unsupervised analysis, modeling data—or "counting" features—comes first, and interpretation only happens later. Rather, *interpretation is a part of how the features are counted*. Deciding to use LDA topic modeling already constitutes a decision regarding how to represent texts. As such, topic modeling is a computational tool which enables one to distant read (Moretti) collections of texts, using representations provided by the tool (cf. Earhart, Clement, Ramsay). Typically, in topic modeling this happens through a workflow where the modeling results are used to guide subsequent close reading of the documents, in order to arrive at an interpretation of what the topics represent (see also the chapter by Chen in this

volume for a discussion of distant and close reading big data). In this sense, heavy-duty interpretive work in topic modeling indeed does seem to take place only in the post-modeling phase of the analysis. However, when the representations used for close reading are selected on the basis of interpretability judgments, we should recognize that they themselves are the products of an interpretative process. In these cases, interpretation determines how the text are represented, as much as the statistical regularities in data, on which the models latch.

Therefore, I argue that a responsible construal of unsupervised modeling based on interpretability should portray it as interpretation of texts using an automated tool for representation, rather than as a search for the best fit among objective representations of texts. It follows that an evaluation of topic modeling should strive as far as possible to subject its underlying interpretability judgments to collective debate. Here I agree with Earhart, who argues that we should “critique machine-aided interpretation in the same way that we critique interpretation through traditional means”.

To see that this issue has some weight to it, let us compare how quantitative metrics and interpretability differ as model selection criteria. Whereas quantitative techniques aim to select one model as the best, interpretability allows for multiple models to be equally good representations of a text corpus. Interpretable results are often produced with many different values as the number of topics (Greene et al. 498). The AssociatedPress word lists depicted above demonstrate this point. At face value, both models seem to generate topics which lend themselves for interpretation. Such diverging results are often described as corresponding to different "granularity" or "resolution" with which the structure of the data can be represented (e.g. Jockers 128, Greene et al. 498). Underlying this idea is the realist assumption that patterns of meaning at different levels can simultaneously coexist in a large corpus, and that formal

methods such as topic modeling can be used for measuring them (Ignatow, Mohr). This assumption might well be true. It is plausible to assume that complex communicative artifacts such as texts contain different layers of meaning, and that word use patterns captured by topic modeling can give us information about their structure. However, my point is that if model selection is based on topic interpretability, then we have no clear-cut way of distinguishing between two models, the results of which seem to lend themselves equally well to interpretation. In this case, there is no best fit, but instead multiple well-fitting, but different models.

The upshot is that in such situations, model selection turns on subjective evaluations of the researcher, regarding which model best addresses her research questions, while matching her background knowledge of the modeled corpus and the phenomena studied. Singling one model as the best by interpretability is bound to be difficult by such criteria, given that in topic modeling there are often hundreds of different candidate models. Thus, model selection by interpretability has a certain degree of arbitrariness to it, even when the researcher is very well informed by her subject matter. This arbitrariness becomes problematic, if possible divergences in the results of different interpretable models are not recognized and discussed. My claim is that this is precisely what we run the risk of doing, if we downplay the role of interpretation in model selection. To substantiate this claim, and to formulate some suggestions with respect to how I think modeling evaluation should proceed, let us shortly look at an example of topic modeling.

In a study of thematic changes in 19th century literature, Jockers chose to use LDA topic modeling to extract 500 topics from a corpus of 3,346 books included in the Stanford Literary Lab's digital collection (135). He interpreted the results by examining wordcloud representations of the most frequent words in each extracted topic, yielding topics with labels such as "Affection and Happiness" and "Enemies" (Jockers 126, 136-139). He then examined how these topics,

among others, vary in the modeled corpus according to author gender and nationality, and found that the former is to a greater extent represented by female authors, while authors writing on the latter topic tend to be male (Jockers 136-139). In a companion study, Jockers and Mimno go on to test the statistical significance of these findings, reporting as one of their results that author gender is a reliable predictor of thematic content.

In these studies, the authors selected one topic model—with 500 topics—among many different candidates, and used it to generate factual claims about a document corpus. Moreover, the results are based on the judgment of the authors that the model with 500 topics is a good representation of the corpus. To be sure, Jockers does report checking his interpretations of the 500 topics with colleagues (135), and is very explicit about many choices he made in the modeling process. However, despite this carefulness, we are not provided with information regarding the different possible models of the same corpus, or the interpretability of their respective topics. What basis do we have for thinking that some other model, with different topic distributions, would not have provided an equally good, or even a better representation?

Suppose now that there is another model of the same corpus, which is also well interpretable, but which diverges from the 500-topic model with respect to how the different topics are distributed across the documents. Suppose further that, upon testing Jockers and Mimno's hypotheses with this model, we find that gender no longer predicts thematic content. This is of course only speculation, but the point is that as long as we are in the dark about potential different models for the analyzed corpus, we do not know whether claims made on the basis of the one singled out are reliable.

Jockers does provide one possible solution to this problem, by arguing that judgments of topic interpretability are unlikely to vary between persons. He claims that the disagreement

between different model evaluators will "almost always" (130) concern the accuracy of particular labels assigned to topics, and not whether the topics represent semantically coherent themes. In a similar vein, Mohr and Bogdanov seem to think that there exists an "actual number of topics in the corpus", which can be identified by a "well-informed observer (a subject-area specialist) who understands the discursive context of the corpus." (560) Were this the case, we could trust interpretability judgments not to vary too much between different evaluators, so as to provide a robust enough criterion for model selection.

My argument against this claim is that we simply do not know whether it is true or not. I have not come across extensive experimental research investigating the extent to which individual judgments of topic interpretability vary (see, however Nelimarkka). Neither have I seen studies on the possible effects of modelers' preconceptions on interpretability judgments. In any case, whether interpretability judgments stay similar across multiple evaluators is an issue to be investigated empirically and, as of present, we seem to be lacking sufficient evidence for deciding on the matter.

Meanwhile, I suggest that presentations of topic modeling results based on interpretability should strive to facilitate collective scrutiny of model selection as far as possible. Ideally, this would involve a discussion concerning the stability of results with respect to changes in the number of topics. Modelers should specify, which models within the range of generated candidates were evaluated for interpretability, and whether they were considered to be good representations of the corpus. Here, the criteria used for assessing topic interpretability should be stated as explicitly as possible (see, for instance, Mimno et al. for a useful discussion of topic quality assessment criteria). Furthermore, information should be provided whether the obtained results (e.g. statistical significance of observed patterns) hold for other plausible

models. Thus, readers would gain a sense of the range of different possible representations of the corpus, and access to information about how they were assessed. Whenever judgments of interpretability are used as a criterion for model selection, I maintain that following the above suggestions would help assess topic modeling results in terms of what they essentially are—that is, as machine-aided interpretations (Ramsay).

5. Conclusion

Unsupervised methods of text analysis hold the promise of enabling the systematic and cost-effective analysis of massive textual datasets. Moreover, these methods have the advantage that their results do not depend on pre-specified categorization schemes, allowing researchers to explore patterns in data without limiting their inquiries with *a priori* conceptualizations. In particular, unsupervised methods have been argued to shift the locus of subjective interpretation in text analysis to a phase after modeling (Mohr and Bogdanov 560).

In this short chapter, I have given reasons for thinking that accounts of unsupervised text analysis should put more emphasis on the role of interpretation. Basing on the example of topic modeling, I argued that whenever texts are modeled basing on subjective judgments of interpretability, model selection should be conceived of as an interpretive process rather than a search for the objectively best representation of texts. Consequently, I proposed that modelers should take steps to make their interpretive judgments explicit and open to debate.

While I used topic modeling as an example to make my argument, I believe that the main point will hold more generally for cases of unsupervised text analysis where model selection depends on interpretability. However, I contend that such claims should be evaluated on a case-by-case basis.

Acknowledgements

This research was supported by the Finnish Funding Agency for Technology and Innovation Tekes, the Finnish Foundation for Economic Education, and the KONE Foundation (project: “Algorithmic Systems, Power and Interaction”).

Works Cited

Biernacki, Richard. “Humanist Interpretation Versus Coding Text Samples.” *Qualitative Sociology*, vol. 37, 2014.

Blei, David. “Probabilistic Topic Models.” *Communications of the ACM*, vol. 55, no. 4, 2012.

Blei, David. “Topic Modeling and Digital Humanities.” *Journal of Digital Humanities*, vol. 2, no. 1, 2012.

Block, Sharon. “Doing More with Digitization. An Introduction to Topic Modeling of Early American Sources.” *Common-place: The Interactive Journal of Early American Life*, vol. 6, no. 2, 2006.

Block, Sharon, and David Newman. “What, Where, When, and Sometimes Why: Data Mining Two Decades of Women’s History Abstracts.” *Journal of Women’s History*, vol. 23, no. 1, 2011.

Burstein, Jill, et al. “Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays.” *IEEE Intelligent Systems*, vol. 18, no. 1, 2003.

Calude, Christian, and Giuseppe Longo. “The Deluge of Spurious Correlations in Big Data.” *Foundations of Science*, vol. 22, no. 3, 2017.

Chang, Jonathan, et al. “Reading Tea Leaves: How Humans Interpret Topic Models.” *Advances in Neural Information Processing Systems 22*, Curran Associates, 2009.

Clement, Tanya. “Text Analysis, Data Mining, and Visualizations in Literary Scholarship.” *Literary Studies in the Digital Age,* edited by K. Price and R. Siemens, MLACommons, 2013, https://dlsanthology.mla.hcommons.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/.

Earhart, Amy. “Data and the Fragmented Text: Tools, Visualization, and Datamining or is Bigger Better?” *Traces of the Old, Uses of the New: The Emergence of Digital Literary Studies*, A. Earhart, Michigan Publishing, 2015, http://dx.doi.org/10.3998/etlc.13455322.0001.001.

Floridi, Luciano. “Big Data and Their Epistemological Challenge.” *Philosophy & Technology*, vol 25, no. 4, 2012.

Greene, Derek, et al. “How Many Topics? Stability analysis for Topic Models.” *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014 Proceedings, Part I,* edited by T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, Springer, 2014.

Grimmer, Justin, and Brandon Stewart. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis*, vol. 21, no. 3, 2013.

Grün, Bettina, and Kurt Hornik. “topicmodels: An R Package for Fitting Topic Models.” *Journal of Statistical Software*, vol. 40, no. 13, 2011.

Haswell, Richard. “NCTE/CCCC’s Recent War on Scholarship.” *Written Communication*, vol. 22, no. 2, 2005.

Hoffman, David, and Don Waisanen. “At the Digital Frontier of Rhetoric Studies: An Overview of Tools and Methods for Computer-Aided Textual Analysis.” *Rhetoric and the Digital Humanities*, edited by J. Ridolfo and W. Hart-Davidson, University of Chicago Press, 2015.

Hoover, David, et al. *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*. Routledge, 2014.

Ignatow, Gabe. “Theoretical Foundations for Digital Text Analysis.” *Journal for the Theory of Social Behaviour*, vol. 46, no. 1, 2015.

Jockers, Matthew. “Theme.” *Macroanalysis. Digital Methods & Literary History*, M. Jockers, University of Illinois Press, 2013.

Jockers, Matthew, and David Mimno. “Significant Themes in 19th-Century Literature.” *Poetics*, vol 41, 2013.

King, Gary. “Preface: Big Data Is Not About The Data!” *Computational Social Science: Discovery and Prediction*, edited by R. M. Alvarez, Cambridge University Press, 2016.

Klein, Lauren, et al. “Exploratory Thematic Analysis for Digitized Archival Collections.” *Digital Scholarship in the Humanities*, vol. 30, 2015.

Lang, Susan, and Craig Baehr. “Data Mining: A Hybrid Methodology for Complex and Dynamic Research.” *College Composition and Communication*, vol 64, no. 1, 2012.

Lee, Monica, and John Levi Martin. “Coding, Counting and Cultural Cartography.” *American Journal of Cultural Sociology*, vol 3., no. 1, 2014.

Licastro, Amanda. “The Problem of Multimodality: What Data-Driven Research Can Tell Us About Online Writing Practices.” *Communication Design Quarterly Review*, vol. 4, no. 4, 2016.

McAllister, James. “Model Selection and the Multiplicity of Patterns in Empirical Data.” *Philosophy of Science*, vol 74, no. 5, 2007.

Meeks, Elijah, and Scott Weingart. “The Digital Humanities Contribution to Topic Modeling.” *Journal of Digital Humanities*, vol. 2, no.1, 2012.

Miller, Benjamin, et al. “The Roots of an Academic Genealogy: Composing the Writing Studies Tree.” *Kairos: A Journal of Rhetoric, Technology and Pedagogy*, vol. 20, no. 2, 2016, http://kairos.technorhetoric.net/20.2/topoi/miller-et-al/index.html.

Mimno, David, et al. “Optimizing Semantic Coherence in Topic Models.” Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011.

Mohr, John. “Measuring Meaning Structures.” *Annual Review of Sociology*, vol. 24, 1998.

Mohr, John, and Petko Bogdanov. “Introduction – Topic Models: What They Are and Why They Matter.” *Poetics*, vol. 41, no. 6, 2013.

Moretti, Franco. *Distant Reading.* Verso, 2013.

Nelimarkka, Matti. “Unsupervised Models Considered Dangerous? Discussing the Practices of Social Scientists with Topic Models for Automated Text Analysis.” Unpublished manuscript presented in *European Consortium for Political Research General Conference*, Oslo, 2017.

Ramsay, Stephen. “In Praise of Pattern.” *TEXT Technology: The Journal of Computer Text Processing*, vol. 14, no. 2, 2005.

Rhody, Lisa. “Topic Modeling and Figurative Language.” *Journal of Digital Humanities*, vol. 20, no. 1, 2012.

Schwartz, Andrew, and Lyle Ungar. “Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods.” *The ANNALS of the American Academy of Political and Social Science*, vol. 659, no. 1, 2015.

Séaghda, Diarmuid, and Simone Teufel. “Unsupervised Learning of Rhetorical Structure with Un-Topic Models.” Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014.

Tangherlini, Timothy, and Peter Leonard. “Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research.” *Poetics*, vol. 41, no. 6, 2013.

Wallach, Hanna, et al. “Evaluation Methods for Topic Models.” Proceedings of the 26th International Conference on Machine Learning, 2009.