Topic Modeling and the Sociology of Literature

Andrew Goldstone
Rutgers University, New Brunswick
andrewgoldstone.com

October 14, 2014
Penn Digital Humanities Forum

agenda

1. Why topic-model?
2. 2.1 How do you make it work?
   2.2 What's going on?
3. What can you do with a model?

Download these slides: andrewgoldstone.com/penn2014

let’s be reductive

let's be reductive

Even with the assistance of computers, one major difficulty of content analysis is that there is too much information in texts. Their richness and detail preclude analysis without some form of data reduction. The key to content analysis, and indeed to all modes of inquiry, is choosing a strategy for information loss that yields substantively interesting and theoretically useful generalizations while reducing the amount of information addressed by the analyst.

Robert Philip Weber, *Basic Content Analysis* (Beverly Hills, CA: Sage, 1985), 40

“the limitations are apparent”

Sociologists ordinarily analyze texts in one of three ways. Some scholars simply read texts and produce virtuoso interpretations based on insights their readings produce. The limitations of this approach for generating reproducible results are apparent.

Paul DiMaggio, Manish Nag, and David Blei, “Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding,” *Poetics* 41, no. 6 (December 2013): 577

post-Marxist pre-DH

The analytical phase proper consists mainly in constructing categories (containing a series of terms or instances...) and working with these categories. In this way, for example, one can compare the presence of categories in different texts from the same corpus or different corpora; examine the instances or representatives that embody the category in different texts; make a list of the qualities attributed to an instance, come to know the terms most often associated with a category.

post-Marxist pre-DH

The analytical phase proper consists mainly in constructing categories (containing a series of terms or instances…) and working with these categories. In this way, for example, one can compare the presence of categories in different texts from the same corpus or different corpora; examine the instances or representatives that embody the category in different texts; make a list of the qualities attributed to an instance, come to know the terms most often associated with a category.

1960s		1990s	
ENTREPRISE@	1,330	ENTREPRISE@	1,404
CADRE@	986	travail	507
SUBORDONNÉS@	797	organisation	451
DIRIGEANTS@	724	RÉSEAU@	450
…			

Luc Boltanski and Eve Chiapello, *The New Spirit of Capitalism*, trans. Gregory Elliott (1999; London: Verso, 2005), 546, 548

a modeling process

a modeling process

1. Obtain digitized texts
2. Featurize texts into “data”
3. Model the data
4. Explore the model: what is valid? what is interesting?
5. Use the model in an argument: explanatory analysis (?)

a modeling process

1. Obtain digitized texts
2. Featurize texts into “data”
3. Model the data
4. Explore the model: what is valid? what is interesting?
5. Use the model in an argument: explanatory analysis (?)

Andrew Goldstone and Ted Underwood, “The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us,” *New Literary History* 45, no. 3 (Summer 2014): forthcoming
http://rci.rutgers.edu/~ag978/quiet/#/about

obtaining texts

Data: not raw (1)

dfr.jstor.org

```
WORDCOUNTS,WEIGHT
the,766
of,482
and,305
in,259
to,224
a,195
new,101
```

data: not raw (2)

2012

```
10.2307/25501736,10.2307/25501736 ,Fantasies of the New Class: The New Criticism_ Harvard Sociology_ and the Idea of the University ,Stephen Schryer ,PMLA ,122 ,3 ,2007-05-01T00:00:00Z ,pp. 663-678 ,Modern Language Association ,fla , ,
```

2014

```
10.2307/25501736 10.2307/25501736 Fantasies of the New Class: The New Criticism, Harvard Sociology, and the Idea of the University Stephen Schryer PMLA 122 3 2007-05-01T00:00:00Z pp. 663-678 Modern Language Association fla This essay examines the professionalization of United States literary studies and sociology between the 1930s and 1950s ...
```

constituting the corpus

constituting the corpus

name	start	end
PMLA	1889	2007
Modern Philology	1903	2013
The Modern Language Review	1905	2013
The Review of English Studies	1925	2012
ELH	1934	2013
New Literary History	1969	2012
Critical Inquiry	1974	2013

21367 total articles.

featurization

- *bag of words* representation: standard but not inevitable (unless you only have access to the bags…)
- "document": bibliographic item, or larger, or smaller?
- feature classes (*types*): tokenizing, standardizing, stemming, lemmatizing
- pruning: stop lists, infrequent types

there's no app for that

```r
# fv is a vector of filenames
counts <- vector("list",length(fv))
n_types <- integer(length(fv))
for(i in seq_along(fv)) {
    counts[[i]] <- read.csv(fv[i],strip.white=T,header=T,
        as.is=T,colClasses=c("character","integer"))
    n_types[i] <- nrow(counts[[i]])
}
wordtype <- do.call(c,lapply(counts,"[[","WORDCOUNTS"))
wordweight <- do.call(c,lapply(counts,"[[","WEIGHT"))
data.frame(id=rep(filename_id(fv),times=n_types),
           WORDCOUNTS=wordtype, WEIGHT=wordweight,
           stringsAsFactors=F)

# etc. etc. etc. etc. etc. etc.
```

model: how to write an article

1. Fix a length: 5000 words

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
    2.1 *the late 19th century*, 40% or 2000 words

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
   2.1 *the late 19th century*, 40% or 2000 words
   2.2 *power/subjectivity*, 40% or 2000 words

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
   - 2.1 *the late 19th century*, 40% or 2000 words
   - 2.2 *power/subjectivity*, 40% or 2000 words
   - 2.3 *social class*, 20% or 1000 words

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
   2.1 *the late 19th century*, 40% or 2000 words
   2.2 *power/subjectivity*, 40% or 2000 words
   2.3 *social class*, 20% or 1000 words
3. Randomly choose words from each topic

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
   2.1 *the late 19th century*, 40% or 2000 words
   2.2 *power/subjectivity*, 40% or 2000 words
   2.3 *social class*, 20% or 1000 words
3. Randomly choose words from each topic
   3.1 *late 19th*: *wilde*, 20; *james*, 15…

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
   2.1 *the late 19th century*, 40% or 2000 words
   2.2 *power/subjectivity*, 40% or 2000 words
   2.3 *social class*, 20% or 1000 words
3. Randomly choose words from each topic
   3.1 *late 19th*: *wilde*, 20; *james*, 15…
   3.2 *power/subjectivity*: *own*, 15; *power*, 10; *subject*, 8; *discourse*, 7…

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
    2.1 *the late 19th century*, 40% or 2000 words
    2.2 *power/subjectivity*, 40% or 2000 words
    2.3 *social class*, 20% or 1000 words
3. Randomly choose words from each topic
    3.1 *late 19th*: *wilde*, 20; *james*, 15…
    3.2 *power/subjectivity*: *own*, 15; *power*, 10; *subject*, 8; *discourse*, 7…
4. Leave words in random order

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
    2.1 *the late 19th century*, 40% or 2000 words
    2.2 *power/subjectivity*, 40% or 2000 words
    2.3 *social class*, 20% or 1000 words
3. Randomly choose words from each topic
    3.1 *late 19th*: *wilde*, 20; *james*, 15…
    3.2 *power/subjectivity*: *own*, 15; *power*, 10; *subject*, 8; *discourse*, 7…
4. Leave words in random order
5. Publication and fame

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
   2.1 *the late 19th century*, 40% or 2000 words
   2.2 *power/subjectivity*, 40% or 2000 words
   2.3 *social class*, 20% or 1000 words
3. Randomly choose words from each topic
   3.1 *late 19th*: *wilde*, 20; *james*, 15…
   3.2 *power/subjectivity*: *own*, 15; *power*, 10; *subject*, 8; *discourse*, 7…
4. Leave words in random order
5. Publication and fame

model: how to write an article

1. Fix a length: 5000 words
2. Randomly choose topic proportions
   2.1 *the late 19th century*, 40% or 2000 words
   2.2 *power/subjectivity*, 40% or 2000 words
   2.3 *social class*, 20% or 1000 words
3. Randomly choose words from each topic
   3.1 *late 19th*: *wilde*, 20; *james*, 15…
   3.2 *power/subjectivity*: *own*, 15; *power*, 10; *subject*, 8; *discourse*, 7…
4. Leave words in random order
5. Publication and fame

(a not so arbitrary example)

modeling parameters

modeling parameters

```
library(mallet)
trainer <- MalletLDA(n_topics,alpha_sum,b)
trainer$model$setNumThreads(threads)
trainer$model$setRandomSeed(seed)
trainer$loadDocuments(instances)
trainer$setAlphaOptimization(n_hyper_iters,n_burn_in)
trainer$train(n_iters)
trainer$maximize(n_max_iters)
```

modeling parameters

```
library(mallet)
trainer <- MalletLDA(n_topics,alpha_sum,b)
trainer$model$setNumThreads(threads)
trainer$model$setRandomSeed(seed)
trainer$loadDocuments(instances)
trainer$setAlphaOptimization(n_hyper_iters,n_burn_in)
trainer$train(n_iters)
trainer$maximize(n_max_iters)
```

Some help with this: github.com/agoldst/dfrtopics

tabula rasa?

An important, general digital humanities goal…might be called tabula rasa interpretation—the initiation of interpretation through the hypothesis-free discovery of phenomena….However, tabula rasa interpretation puts in question [the aspiration] to get from numbers to humanistic meaning.

Alan Liu, "The Meaning of the Digital Humanities," *PMLA* 128, no. 2 (March 2013): 414

model outputs (1)

model outputs (I)

```
0.17606 see even own both rather view role
0.12924 other different process experience individual two bot
0.00777 beowulf old english ic pe mid swa
0.04118 law legal justice rights right laws case
0.01694 voltaire rousseau mme french corneille plus diderot
0.03112 shakespeare play hamlet king scene plays lear
0.10974 words voice speech own like know way
0.02935 derrida other always question text even time
0.02637 new public city urban american space world
```

model outputs (2)

- each individual feature (word) of each document is assigned to an estimated-most-likely topic ("final sampling state")

  Virginia Woolf$_{62}$ once wrote$_{50}$ that putting$_{43}$ a serious argument$_{7}$ into a review$_{17}$ is like cramming a large$_{50}$ parcel$_{29}$ into the pocket$_{43}$ of a good$_{50}$ coat$_{43}$

model outputs (2)

- each individual feature (word) of each document is assigned to an estimated-most-likely topic ("final sampling state")

  Virginia Woolf$_{62}$ once wrote$_{50}$ that putting$_{43}$ a serious argument$_{7}$ into a review$_{17}$ is like cramming a large$_{50}$ parcel$_{29}$ into the pocket$_{43}$ of a good$_{50}$ coat$_{43}$

  truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$

model outputs (2)

- each individual feature (word) of each document is assigned to an estimated-most-likely topic ("final sampling state")

  Virginia Woolf$_{62}$ once wrote$_{50}$ that putting$_{43}$ a serious argument$_{7}$ into a review$_{17}$ is like cramming a large$_{50}$ parcel$_{29}$ into the pocket$_{43}$ of a good$_{50}$ coat$_{43}$

  truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$ truth$_{109}$

whence:

- a $k \times V$ matrix of the probability of each feature in each topic
- a $k \times N$ matrix of proportions of topics in each of N documents

lies, damn lies, and topics (1)

We refer to the latent multinomial variables in the LDA model as topics, so as to exploit text-oriented intuitions, but we make no epistemological claims regarding these latent variables beyond their utility in representing probability distributions on sets of words.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research* 3 (March 2003): 996n1

lies, damn lies, and topics (2)

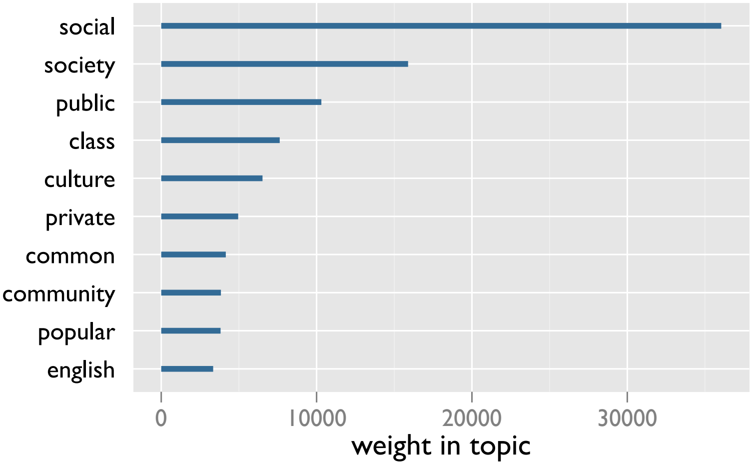

Figure: A thematic topic

lies, damn lies, and topics (3)

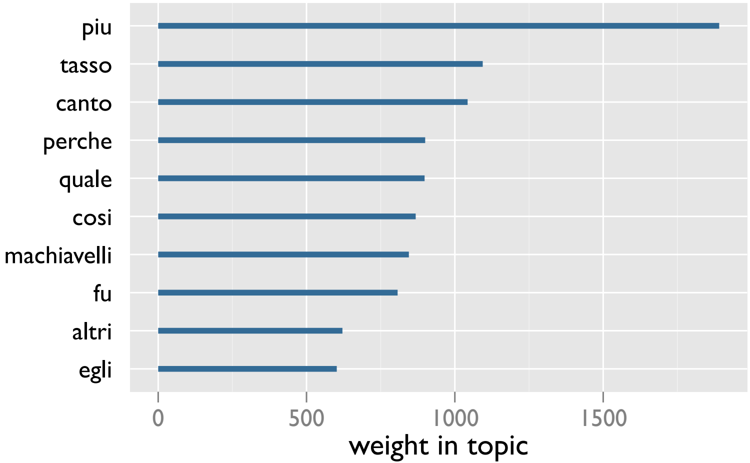

Figure: A “foreign” language topic

lies, damn lies, and topics (4)

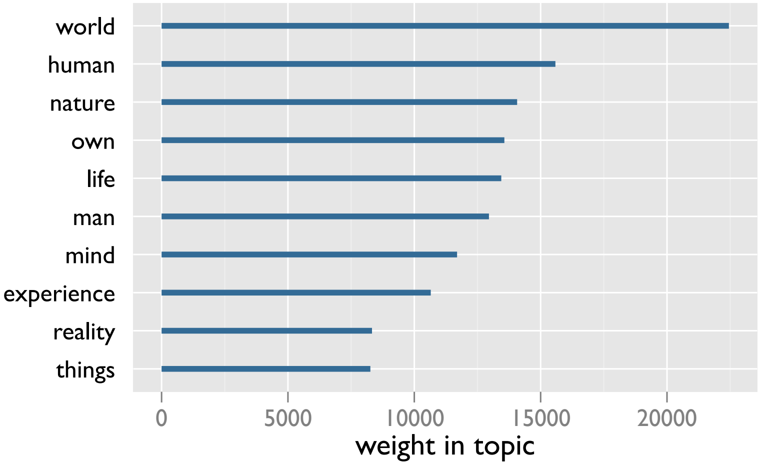

Figure: A broadly discursive topic

lies, damn lies, and topics (5)

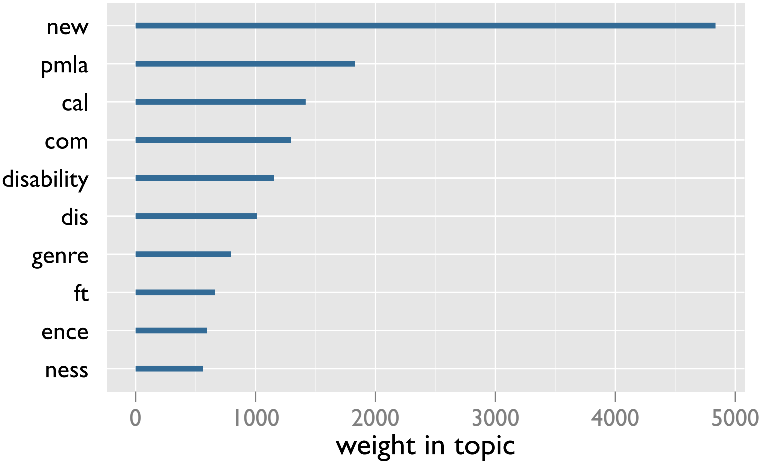

Figure: A garbage topic

iterative exploration

- agoldst.github.io/dfr-browser
- *Quiet Transformations*: rci.rutgers.edu/~ag978/quiet/

Example: interpreting *social work form*
rci.rutgers.edu/~ag978/quiet/#/topic/58

terms in context

16 criticism work critical theory art critics critic nature method view
18 man moral good nature men human virtue reason world order
30 myth garden golden venus tree color flowers green ritual nature
38 nature natural man world human new ideas theory idea universe
82 life world own man human experience nature both becomes vision
93 world human nature own life man mind experience reality things
106 wordsworth keats nature poet romantic ode mind see poetry prelude

defects of the virtues

The top few words in a topic only give a small sense of the thousands of the words that constitute the whole probability distribution.

Benjamin M. Schmidt, “Words Alone: Dismantling Topic Models in the Humanities,” *Journal of Digital Humanities* (Winter 2012)

moving target

article year	top topic 16 words
1890	attempt method art opposition esthetic
1900	work subject proper principles art
1910	criticism nature critics ideas work
1920	unity art work ideas method
1930	criticism theory work method critical
1940	criticism critics work theory critical
1950	criticism work critical method critics
1960	work criticism art critical critics
1970	criticism theory view work art
1980	criticism critical work theory critics
1990	criticism work critics critical critic
2000	critical work criticism critics theory
2010	work art theory criticism critics

Table: Top words assigned to Topic 16 *criticism work critical theory*

virtues of the defects

virtues of the defects

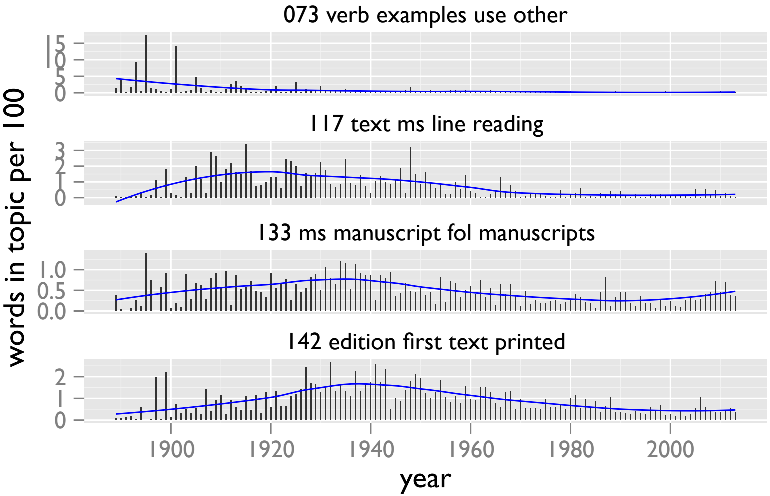

Figure: Philology and textual-studies topics

rise and rise

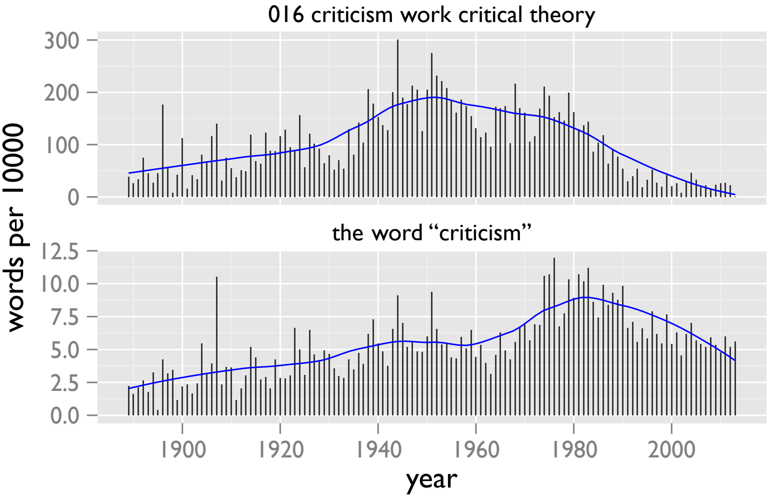

Figure: Criticism as topic and key word

“criticism” and theory

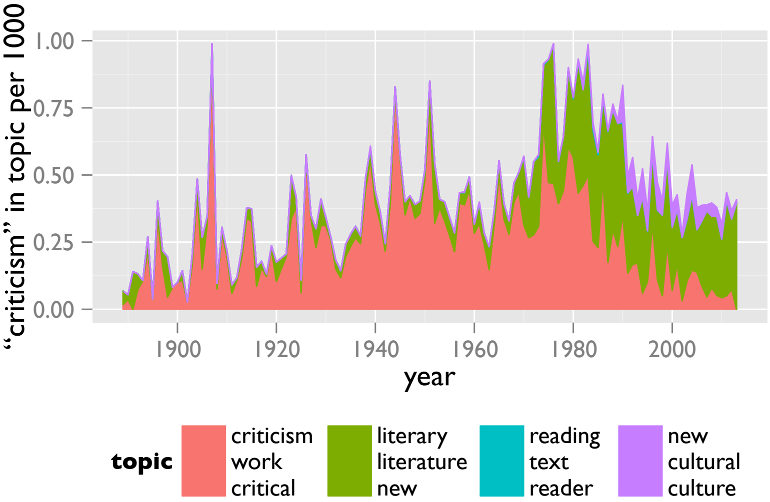

Figure: “Criticism” across topics

reading

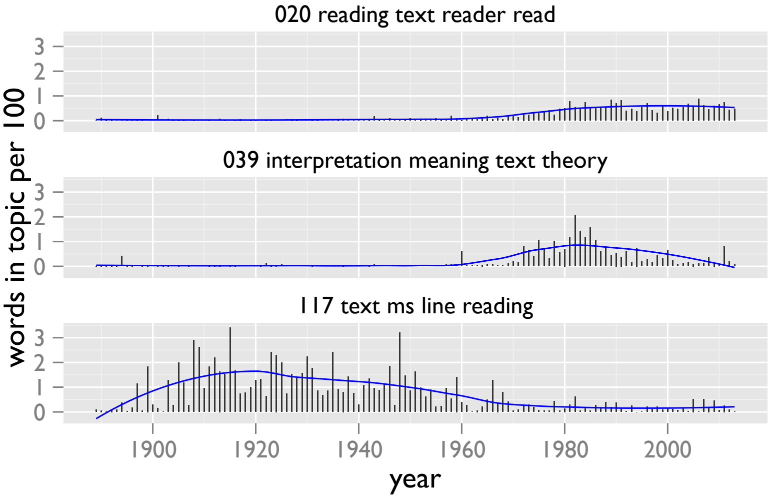

Figure: Reading and interpretation as topics

recent developments

143 new cultural culture theory
015 history historical new modern
058 social work form own
138 social society public class
069 world european national colonial
019 see new media information
025 political politics state revolution
077 human moral own world
048 human science social scientific
036 economic money value labor
004 law legal justice rights
102 feeling emotional moral pleasure
108 violence trial crime memory

Browser visualization: topics sorted by time of peak
rci.rutgers.edu/~ag978/quiet/#/model/list/year/down

polemic: no returns

further: discussions

- David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (March 2003): 993–1022
- David M. Blei, "Probabilistic Topic Models," *Communications of the ACM* 55, no. 4 (April 2012): 77–84
- David Mimno, "Computational Historiography," *Journal on Computing and Cultural Heritage* 5, no. 1 (April 2012): article 3
- John Mohr and Petko Bogdanov, eds., "Topic Models and the Cultural Sciences," special issue, *Poetics* 41, no. 6 (December 2013)
- Scott Weingart and Elijah Meeks, eds., "Topic Modeling," special issue, *Journal of Digital Humanities* 2, no. 1 (2012)
- Justin Grimmer and Brandon M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21, no. 3 (Summer 2013): 267–297

further: software

- "MALLET: Machine Learning for Language Toolkit," http://mallet.cs.umass.edu
- Blei group software http://www.cs.princeton.edu/~blei/topicmodeling.html
- David Mimno, jsLDA, http://mimno.infosci.cornell.edu/jsLDA/
- visualizations: see http://agoldst.github.io/dfr-browser/#the-polished-options
- next on my Xmas list: the structural topic model http://cran.r-project.org/web/packages/stm/