Available online at www.sciencedirect.com
SciVerse ScienceDirect
ELSEVIER
Poetics xxx (2013) xxx–xxx
POETICS
www.elsevier.com/locate/poetic

Differentiating language usage through topic models

Daniel A. McFarland [a,*], Daniel Ramage [b], Jason Chuang [c], Jeffrey Heer [c], Christopher D. Manning [d], Daniel Jurafsky [e]

[a] *Graduate School of Education, Stanford University, 485 Lausen Mall, Stanford, CA 94305, USA*
[b] *Google, Inc., USA*
[c] *Computer Science & Engineering Department, University of Washington, 185 Stevens Way, Seattle, WA 98195, USA*
[d] *Department of Linguistics & Computer Science Department, 353 Serra Mall, Stanford, CA 94305, USA*
[e] *Department of Linguistics, Stanford University, Building 460, Stanford, CA 94305, USA*

Abstract

Sociologists wishing to employ topic models in their research need a helpful guide that describes the variety of topic modeling procedures, their issues, and various means of resolving them so as to convincingly answer sociological questions. We present this overview by recounting a series of our prior collaborative projects that have employed and developed various forms of topic models to understand language differentiation in academe. With each project, we encountered a variety of model-specific issues concerning the validity of topics and their suitability to our data and research questions. We developed a variety of novel visualization techniques to make sense of topic-solutions and used a variety of techniques to validate our results. In addition, we created a variety of new topic modeling techniques and procedures suitable to different kinds of data and research questions.

Keywords: Topic models; Language differentiation; Text; Domains; Culture; Sociology

1. Introduction

During the last decade, we have witnessed an explosion of freely available digitized material, a significant amount of which is text. In many instances, we can see the written communication of an entire community over time. In others instances, we have large-scale corpora that are

* Corresponding author.
E-mail addresses: mcfarland@stanford.edu (D.A. McFarland), dramage@gmail.com (D. Ramage), jcchuang@cs.stanford.edu (J. Chuang), jheer@cs.stanford.edu (J. Heer), manning@stanford.edu (C.D. Manning), jurafsky@stanford.edu (D. Jurafsky).

0304-422X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.poetic.2013.06.004

repositories of a population’s knowledge and communication (e.g., ISI Web of Knowledge). Affixed to this text is an assortment of meta-information, or labels – such as the names of persons, groups, addresses, and so on. In short, today’s sociologist is confronting an ever-expanding treasure trove of communicated text that people categorize, express in association with various social groups, and that is changing over time.

Multiple subfields of sociology have long been interested in language usage and how it can be differentiated. In the sociology of culture, language usage is referenced in characterizations of different subcultures, idiocultures (Fine, 1987) and genres of writing (Griswold, 2008). These are often identified through slang, special terms, and how symbolic codes are used. In social movements, scholars have identified forms of rhetoric and interaction frames as normative styles of language (McLean, 1998; Snow et al., 1986; Vicari, 2010). Here, certain arguments and key terms are used to redirect social situations. In the sociology of knowledge and the sociology of science, texts have been carefully studied so as to distinguish epistemic cultures (Knorr-Cetina, 1999), elements of argumentation (Latour, 1987), paradigms (Kuhn, 1970), thought styles, and thought collectives (Fleck, 1979 [1935]), to name but a few. In each of these instances, we have references to distinctive languages and styles of argumentation and reasoning. For lack of a better term, one can regard these differentiated styles and subsets of language usage as ‘‘domains’’ (White, 1992).

In general, sociologists identify subsets and styles of language usage through counts of certain words or qualitative inspection of meta-level usages of language – such as in frame, narrative and genre analysis or the characterization of inferential methods (Franzosi, 2010; Knorr-Cetina, 1999; Vicari, 2010). In many regards, sociological analysis of language usage has not dramatically changed over the last thirty years. We still use demanding coding procedures or the same content analytic techniques to differentiate patterns of language and symbolic usage that were initially developed in the 1980s (Weber, 1985). Some work focuses on relations between coded categories, but direct assessment of raw language and its internal relations has been slow in coming (for review, see Mohr, 1998; Kirchner and Mohr, 2010).

By contrast, linguistics and computer science have merged efforts and advanced the empirical study of language in their shared field of computational linguistics and natural language processing (Jurafsky and Martin, 2009; Manning and Schütze, 1999). This computationally based field has developed a suite of novel, linguistically informed methods capable of identifying patterns of language usage in large bodies of text and communication. These include statistical language models (Goodman, 2001) that use multi-word sequences (n-grams) as a linguistic context and are a part of modern state-of-the-art machine translation (Koehn et al., 2007) and speech recognition systems (Rabiner and Juang, 1993). Statistical language models are good at assigning likelihoods to word sequences, but they are less useful for exploratory analysis of a document collection. Other models explicitly account for the linguistic structure in text – such as probabilistic context-free grammar (PCFGs [Marcus et al., 1993]), head-driven phrase structure grammar (HPSGs [Pollard and Sag, 1994]), lexical functional grammar (LFGs [Dalrymple, 2001]), and dependency grammars (Kübler et al., 2009) – and they have found roles in natural language understanding tasks (such as Dagan et al., 2006; Lin and Pantel, 2001) that require finer-grained analysis of the content of individual sentences or of the relationships between words. These models are less useful for studying high-level relationships between documents. Supervised document classification is another class of techniques that can be used to automatically assign one of an existing set of labels to new documents, where some external mechanism, such as human feedback, provides information on the correct classification for documents (Lewis et al., 2004). Finally, unsupervised clustering techniques can place related documents together on the basis of the words they possess without using external label information (Strehl et al., 2000).

A method suited to the study of high-level relationships between documents is a class of probabilistic techniques called "topic models" (such as Latent Dirichlet Allocation [Blei et al., 2003]), which identify distinct "bags" of words co-occurring in documents. These models allow us to study large scale statistical trends of word use across a document collection over time. The labeled topic models we will describe in this paper can be seen as a hybrid of the classification and clustering approaches described above – a form of semi-supervised document classification specifically adapted to the task of understanding a document collection.

In this paper, our goal is to relate how topic modeling can be effectively used and adapted by sociologists in their research.[1] To provide such an account, we draw upon our previously published work focused on large academic corpora. Our prior work attempted to answer a series of research questions:

> What does a population talk about? What topics? Have their topics changed over time? (Hall et al., 2008)
>
> What terms do our categories reference? Have our categories changed over time? (Johri et al., 2011; Ramage et al., 2009a,b)
>
> Do groups have their own language? Does their language change over time? (Ramage et al., 2011)
>
> Do groups transfer their language, and how? Do some fields colonize others? Do others grow isolated? (Anderson et al., 2012; Ramage et al., 2011)

What one immediately notices about these questions is that they are all descriptive: they reflect the fact that we first want to understand how language usage is differentiated. After we reliably identify those patterns, we intend to develop models of prediction and causation, asking what factors generate patterns of language differentiation. However, our initial questions are more modest and descriptive in nature.

In pursuing these questions, we encountered a variety of methodological issues, leading us to adopt and develop different topic modeling techniques (e.g., unsupervised, supervised, and mixed). In order to identify processes of change, we developed novel means of representing topic-flows, and domain-interrelations, over time. As we applied these forms of topic modeling, we confronted a number of issues concerning the validity and nature of the topics we identified. Were we after unrecognized latent topics, or ones that members recognize? What recourse did we have to verifying or confirming topics on some level? In addressing these questions, we had to adapt existing topic modeling techniques and supplement them with data visualization (Chuang et al., 2009, 2012a,b,c) and validation efforts (Ramage et al., 2009a,b).

In what follows, this paper will present various classes of topic models, and provide guidelines on how these models best solve problems in various domains. In particular, we describe the series of efforts we took and how we have come to interpret their utility for sociological research. We will mostly cite previous published work, and we will draw on a corpus of dissertation abstracts from 240 research-oriented universities in the United States, filed in the period of 1980–2010 (ProQuest). This corpus entails well over 1 million abstracts and their accompanying meta-information (date, names, etc.). In some illustrations, we will focus on dissertations associated with subject labels we believed were related to anthropology (e.g., "culture," "anthropology," "archeology," etc.). That corpus reflects the sort of sampling sociologists may do to more

[1] Computational linguistics affords a variety of other techniques that could advance sociological analyses of meaning-making (for a review, see O'Connor et al., 2010).

Please cite this article in press as: McFarland, D.A., et al., Differentiating language usage through topic models. Poetics (2013), http://dx.doi.org/10.1016/j.poetic.2013.06.004

carefully inspect a particular knowledge domain or language community like anthropology. Nevertheless, this data and analyses of it are presented merely as illustrations of the sort of research we performed more fully in previously published work.

Next, we introduce various forms of topic models, how we employed and interpreted them, and how we tried to improve upon them to ask and answer a series of different research questions about knowledge domains.

2. Types of models

When it comes to differentiating subsets of language-usage, sociologists are mostly concerned with related terms that are used in the same fashion within representative texts.[2] Topic models, or more specifically, *latent Dirichlet allocation* (LDA – Blei et al., 2003) models, help identify these sets of similar terms. LDA is a probabilistic model of language that identifies sets of words, or ''bags of words,'' that co-occur across documents. LDA is called a ''topic model'' because the identified sets of words tend to reflect underlying topics that, in combination, characterize every document in a corpus (Blei et al., 2003; Blei and Lafferty, 2006; Buntine and Jakulin, 2004; Griffiths and Steyvers, 2004; McCallum et al., 2004; Rosen-Zvi et al., 2004).

Mathematically, a *topic* is a specialized probability distribution over words. And a *topic model* specifies a probabilistic procedure by which documents can be generated. As such, each document is modeled as a mixture of multinomial distributions over words in different proportions. In more formal terms – given as input a desired number of topics K and a set of documents containing words from a word vocabulary V – LDA models infer K topics each a multinomial distribution over words V. Simultaneously, the models recover the per document mixture of topics that best describes each document. For example, a document about using lasers to measure biological activity might be modeled as a mixture of words from a ''physics'' topic and a ''biology'' topic, each consisting of its own characteristic distribution over words.

There exist several open source implementations for topic models. Our group publishes one such implementation, the Stanford Topic Modeling Toolkit (http://nlp.stanford.edu/software/tmt/), which implements the methods described in this article and works with document collections as stored in a single comma-separated value (CSV) spreadsheet. Labels for each document can optionally be stored in one or more columns, and they are usually delimited by a space character when multiple labels are stored in a single cell. The text of the document or its abstract (usually on the order of a few hundred words to few thousand words) is stored in one or more other cells. Sample scripts are provided for training a model on the text, computing topic distributions for each document, and writing those distributions in CSV format for further analysis in a practitioner's tool of choice, such as R or Excel. Other high quality implementation of many related topic models exist, such as MALLET (http://mallet.cs.umass.edu/), which is maintained by researchers at University of Massachusetts Amherst.

[2] Distinct from this are methods for the analysis of token terms and bigrams that represent an advance over current sociological applications of content analysis that count word frequencies (Weber, 1985). Perhaps the most common is a numerical statistic for how important a word is within a document and corpus: term frequency – inverse document frequency, or TF-IDF (see Salton and McGill, 1983). The TF-IDF value increases proportionally to the number of times a word appears in the document, but high frequency words in the corpus are weighted less than rare ones. This helps control for the fact that some words are generally more common than others and, therefore, not a basis of distinction. While TF-IDF effectively scores the similarity of documents using the same token (rare) words, it cannot assign a high score to the shared use of related terms (e.g., ''heat'' and ''thermodynamics''). This is why it is ill suited to the identification of what we term ''language-domains.''

2.1. Unsupervised topic models

Both unsupervised and supervised topic models have been applied to examine language in social media (Ramage et al., 2010), medical literature (Newman et al., 2006), and academic publications (Erosheva et al., 2004). LDA and other unsupervised topic model variants are the most common. These models are called ‘‘unsupervised’’ because they do not incorporate manual notation into the learning procedure of topics. LDA may learn topics that are hard to interpret and the model lacks an explicit interface for fine-tuning the generated topics to suit an end-use application. The number of topics that the model discovers is either left as a free parameter (in the case of LDA) or tuned via hyper-parameters (in the case of the hierarchical Dirichlet processes; Teh et al., 2006).

Common to all unsupervised topic models is the idea that language is organized by latent dimensions that actors may not even be aware of. When applied to everyday speech, basic (unsupervised) topic models usually identify areas of discussion – like driving and stop signs, and distinguish that from, say, dating. As such, basic topic models are generally not well suited to identifying language-communities or network-domains (White, 1992).[3] Later in this article, we will relate the specific instances where topic modeling can be guided (or supervised) so as to identify sets of words most commonly associated with social groups.

2.1.1. Validation of topics

One of the challenges of unsupervised topic modeling is the identification of the number of latent topics characterizing a corpus. How many topics are there in a corpus – 100 or 1000, 100 or 102? In many regards, topic modeling has the same challenges as factor analysis or cluster analysis – naively selecting too few or too many topics may lead to substantively different interpretations of results. Topics must be carefully validated before interpretative conclusions can be drawn.

In our work, we came across a variety of feasible approaches to validating the set of topics identified. One simple means of assessing topic models is to look at the qualities of each topic and discern whether they are reasonable. One measurable quality of a topic concerns its relevance to the corpus or *load*.

For example, when we apply a basic topic model to our corpus of anthropology-related dissertations, we identified a solution with 40 topics, some of which dramatically differed in their salience to the corpus, and in their relevance. Table 1 shows two of these topics and lists the words most associated with the latent dimension, as well as the keywords and subject categories students applied to their dissertations when submitting to ProQuest (what can be called ‘‘meta-data’’ or ‘‘labels’’). The counts are the number of times the word occurs in the topic (salience), and the total words reflect the extent to which the topic loaded on, or was relevant to, the corpus. Words and topics with little load are considered to be noise. The first topic is an expected one, and concerns qualitative methods. It has a relatively high load compared to other topics in the model. The second topic is unexpected and is considered noise. It shows that our selection of dissertations referencing keywords and subjects using ‘‘culture’’ may have actually meant lab cultures relevant to an entirely different field. Simple inspection shows that not only is the topic

[3] To identify language communities, linguists often rely on different features of language reflective of dialect and style – such as phonetic, prosodic, or syntactic variables. Those features are not typically represented in current topic models. However, there is a way to identify language-communities (or network-domains), and it relies on particular corpora and labeled variants of LDA that use community-provided labels. We will describe this approach below.

Table 1
Words in two topics of the anthropology corpus.

Qual methods (01)		Omitted topic (24)	
Words	# Words	Words	# Words
interview	7378.48	cell	2202.98
particip	7115.60	cultur	1634.60
experi	4628.45	product	603.28
observ	3892.81	protein	520.59
includ	2409.87	growth	433.98
collect	2310.36	increas	431.87
conduct	2219.95	rate	400.68
inform	1908.60	express	377.88
qualit	1889.27	concentr	375.91
method	1873.30	gene	317.28
ethnograph	1865.33	tissu	302.66
find	1604.32	acid	289.83
activ	1487.01	plant	271.57
live	1457.80	activ	269.97
theme	1451.91	vitro	266.99
person	1371.66	produc	264.95
question	1094.05	line	264.48
in-depth	1052.63	level	264.48
understand	1031.45	respons	257.38
explor	967.12	condit	235.42
Total (load)	147099.63	Total (load)	66615.48
Top subjects	# Documents	# Documents	# Documents
cultural anthro	627.57	chemical eng	83.72
sociology	172.17	cellular bio	79.08
minority & ethnic	107.93	molecular bio	46.14
Top keywords			
culture	71.15	cell culture	60.25
identity	30.96	tissue culture	20.38
women	29.87	apoptosis	10.93

qualitatively incorrect, but the load is far lower in this corpus. When very few documents and words load on a topic, it is likely to be rare, insignificant, and likely noise. Too many poorly loading topics suggest the model as a whole has not been well identified. In this specific instance, it also suggests that there was a problem during data selection, upstream of model training.

Another measurable quality of a topic concerns its coherence or focus. Here, the standard measure is that of *entropy*, a measure of information content. The coherence and organization of a document's words is indicated by the entropy of its posterior topic distribution. Prior work finds that the entropy of the estimated topic distribution on a true document is lower than that of a fake and randomly generated document (Misra et al., 2008). Hence, high topic entropy typically arises in documents that are confusable among many different topics. The same measure of entropy can be applied to the word distribution in a topic. A topic has less coherence when there is no core word-set that stands out. Such topics, again, may be cases for exclusion and revision when examining topic model output.

Another means of assessing the number of topics entails *perplexity analysis* and, a standard machine learning approach of training a model on a portion of the data before testing it on held-out data (for review of these approaches, see Newman et al., 2009; Wallach et al., 2009). Topic

Please cite this article in press as: McFarland, D.A., et al., Differentiating language usage through topic models. Poetics (2013), http://dx.doi.org/10.1016/j.poetic.2013.06.004

model perplexity is scored by first splitting each document in half. The per-document topic distribution is estimated on the words in the first half of the document, and this distribution is used to compute an average of how surprised the model is by the words in the second half of the document. The perplexity score is the exponentiation of the cross-entropy of the second half of the document under the model parameters inferred for the first half, and it can be interpreted as the number of equally probable word choices that would result in a similarly surprising second half of the document. Lower numbers indicate a surer model, but better perplexity scores do not always produce more interpretable topics. Nonetheless, perplexity scores can be used as stable measures for picking among alternatives, for lack of a better option. In general, perplexity is reduced as the number of topics increases, much like an infinite regress. A good rule of thumb is to pick a number of topics that produces reasonable output and after the perplexity score has begun to flatten out.

Remiss in this approach is a reality check, an expert, or what many might regard as a *ground truth*. Quantitative approaches assess how well clusters of words predict remaining documents. Some of these identified clusters and topics can be ones that no one recognizes or that are data errors (such as the fact that some concern petri dish cultures instead of human cultures). In addition, they can be topics defined at a unit of coarseness that does not correspond with practitioners' sense of the domain.

In some of our work, we developed greater trust in topic results by using experts of a knowledge domain to assess the quality of identified topics (Chuang et al., 2012b; Hall et al., 2008; Ramage et al., 2009a,b).[4] In so doing, we assumed knowledge domains are socially constructed and a matter of shared subjective perception (Hacking, 1999). For example, in our analysis of the field of computational linguistics using the ACL Anthology Network corpus (Radev et al., 2009), we used data visualization to make model results quickly interpretable (see Chuang et al., 2012a) and we called upon the expertise of the two leading authors of the field's primary textbooks (Jurafsky and Martin, 2009; Manning and Schütze, 1999).[5] Their perception was regarded to be a "ground truth" and topics were accepted and rejected on the grounds of whether the two experts recognized and agreed that the identified word-sets constituted research areas. Experts also assigned names to topics, which enabled further analysis. In the computational linguistics domain, for example, the experts labeled an induced topic containing words like "string, sequence, transformation, left, right, match, symbol, pair" as Automata Theory, and labeled a topic containing frequent words like "here, there, rather, might, fact" as a noise topic that was then removed from consideration.[6]

As such, we relied on data visualization, domain-knowledge about that academic world's categorization, and agreement across expert views. This resulted in the identification of 72 different research topics that arose in the Association of Computational Linguistics over a 30-year period (Anderson et al., 2012).

[4] Others have identified the number of topics via post hoc judgments (the researcher uses judgement). In many instances, this results in a large assortment of dropped topics and some residual concern about the quality of the researcher's judgment. The reliance on expert opinion is considered more valid (Hall et al., 2008). And where multiple experts exist, reliance on their points of agreement is considered even more valid yet (Anderson et al., 2012).

[5] As textbook writers, these authors have had to organize the full set of ACL topics and represent experts who have a broad coverage of the entire field. In contrast, other types of experts might be too concentrated or "buried" in their own fields of specialty and very good at verifying specific subfields but not good for the task of identifying the broader intellectual organization of the field.

[6] In all our work on topic models, even more common words (the highest frequency English words "the," "a," "of," and so on) are always removed from the texts before our topic models are induced.

Table 2
Four recognized topics in anthropology dissertations.

Social structures (17)		Physical anthropology (26)		Archeology (02)		Identity studies (34)	
Words	# Words	Words	# Words	Words	# Words	Words	# Words
commun	18550.71	popul	1706.61	site	2170.42	ident	10643.51
member	2827.34	human	1653.48	archaeolog	1658.92	practic	5770.56
organ	2825.08	variat	1486.05	period	1206.46	discours	5530.85
structur	2571.59	genet	1342.73	region	1166.21	cultur	5206.53
network	2519.23	size	1165.51	evid	968.93	nation	4881.50
interact	2481.27	bone	1117.50	pattern	877.60	construct	4017.05
relationship	2465.34	sampl	1110.39	late	870.23	global	2962.51
group	2459.84	morpholog	1084.33	popul	832.45	wai	2745.27
individu	2382.06	differ	1056.88	settlement	725.97	local	2627.24
relat	2041.83	pattern	936.35	materi	711.29	polit	2547.49
societi	1732.18	primat	841.03	suggest	677.95	examin	2353.10
kinship	1217.70	speci	784.02	vallei	662.49	negoti	2226.53
form	1195.73	function	682.28	earli	643.40	power	2018.67
activ	1174.04	compar	649.31	indic	632.59	ethnograph	1986.65
exchang	1119.24	measur	635.61	prehistor	622.92	explor	1973.67
role	1049.98	modern	616.66	remain	551.88	argu	1959.14
share	1027.18	rel	587.84	middl	505.58	subject	1810.53
associ	1000.87	bodi	577.62	area	495.92	relat	1740.51
maintain	904.08	indic	574.24	ancient	434.89	project	1740.16
institut	878.40	evolut	574.12	reconstruct	404.62	space	1705.02
Total (load)	139452.03	Total (load)	115108.50	Total (load)	85276.60	Total (load)	236260.35
Top subjects	# Documents		# Documents		# Documents		# Documents
cultural anthro	695.24	physical anthro	611.98	archeology	314.24	cultural anthro	1100.75
sociology	139.24	anatomy & phys	128.55	cultural anthro	279.05	sociology	244.84
minority & ethnic	91.84	genetics	74.71	physical anthro	184.86	womens studies	197.98
Top keywords							
culture	54.69	primates	36.10	prehistoric	19.64	identity	123.88
community	30.49	evolution	29.87	peru	19.11	culture	122.23
identity	27.98	functional morph	13.62	culture	16.94	gender	75.89

Once latent topics are trusted by a variety of means, sociologists can begin to study how they vary over time. In so doing, one can identify the ebb and flow of different language-domains or research-areas within a field. Using the anthropology dissertation corpus, we can illustrate how some of the most recognized topics change. In particular, we identify a topic concerning the modern era topic of social structure; then topics specific to the anthropological subfields of physical anthropology and archeology; and finally the topic of identity which has become more central to the anthropology discipline over time. Table 2 shows the words and subject-theme labels affixed to these topics.

If we measure the load of topics for each year (as sums of those word-sets, or number and percent of documents using those words), then we can begin to plot changes. Fig. 1 shows how the word load and percentage of documents referencing these word-arrays changes over time.

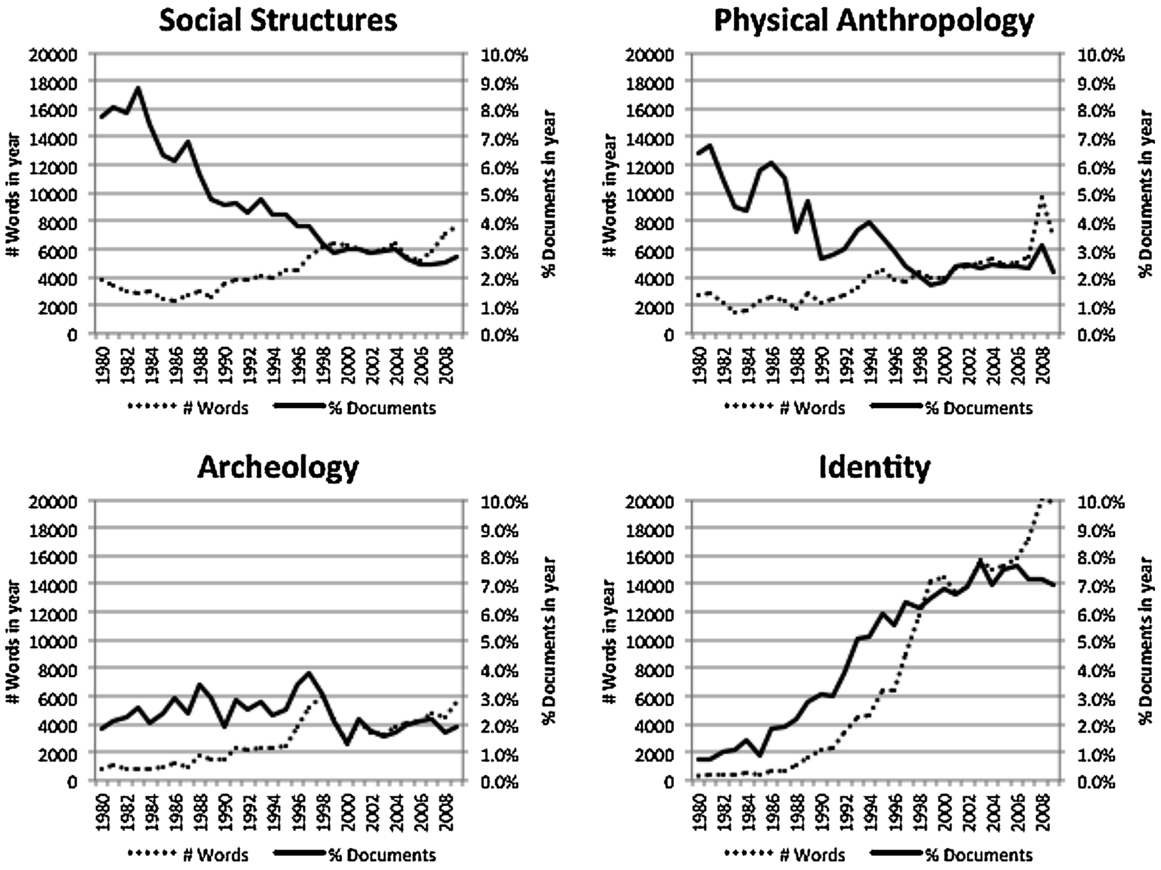

Fig. 1. Topic dynamics in anthropology (dissertation abstracts).

Notably, most topic word loadings increase (solid line), and this is a function of the digital corpus becoming larger over time as more and more dissertations get digitized and filed. As such, it is not a very suitable metric for illustrating change. The percentage of documents referencing these topics shows how the salience of a topic to a corpus changes over time (dotted line). If we look to each figure, we see that modern topics concerning social structure (Nadel, 1957) have declined. Other similar modernist topics, like ones concerning ritual, have also declined. Similarly, physical anthropology's standing in the field of anthropology seems to have diminished. Archeology has had a minor, but relatively stable presence in anthropology-related research (other topics like linguistic anthropology are even smaller). And finally, certain topics like those concerning identity (as well as others concerning media studies) have grown tremendously in the last 30 years. Hence, from identifying reasonable topics in a corpus and plotting their salience to a corpus over time, one can demonstrate field-level shifts in the sorts of language-domains being used.

These shifts can then be related to other social events inside or outside the field – for example, showing similarities and differences in the topics studied by male and female researchers (Vogel and Jurafsky, 2012) or studying the role of government funding in the progression of a particular field (Anderson et al., 2012).

2.2. *Supervised topic models*

Recently, scholars have proposed several modifications to LDA so as to incorporate supervision and afforded labeling/categorization schemes. One common form of supervision posits that a single label (of metadata) is generated from each document's empirical topic mixture distribution, such as in Supervised LDA (Blei and McAuliffe, 2007). While such variations allow human-provided labels (or even corpus-provided) to affect the learned topics, they do not learn direct correspondences between labels and words. For that, we can turn to the event models of supervised text classifiers – such as *naïve Bayes* (McCallum and Nigam, 1998), which assumes that each label directly corresponds to a multinomial distribution over words. As such, the association of each label to a weighted bag of words can be learned by the model.

In many ways, these label-specific word distributions act as what sociologists might regard as category-domains (''catdoms''), or words correspondent with category-designations. The identification of category-domains can be useful in many instances. For example, scholars might want to know the scholarly language associated with certain nations (e.g., the features of Chinese political science papers versus Israeli ones, so as to see the effects of states on studies of political systems), subject categories (e.g., the language features distinguishing computational biology from bio-statistics), or even authors and years. One may even want to know how the words associated with a category change over time so as to understand how the meaning of ''liberal'' today differs from what it meant in 1970.

When it comes to identifying knowledge-domains, supervised models may fall short because many of our labels and categories do not designate a recognized social group. For example, papers may be associated with an outlet name (newspaper or journal), an author, and even a year. However, in some instances documents are labeled by group names and perceived network-positions. In the case of scholarly papers, authors frequently assign their papers to a subject category reflecting a subfield – e.g., biophysics, computational linguistics, etc. These labels or categories are ostensive definitions of a thought community or domain. They are especially helpful in identifying sub-languages associated with perceived communal categories – so a category-domain believed to correspond with a thought community (or ''netdom,'' see White, 1992).

One complication for applications of naïve Bayes-like supervised text models is that documents are often affixed with multiple labels. Hence, an article can have multiple authors, and a single interdisciplinary dissertation can span thought communities and belong to several. For example, the text could be categorized as relevant to structural biology, biophysics and systems biology. Simple supervised models do not allow for multiple labels, and even if we extend the model to allow multiple labels, we do not know which words in the document represent which label(s). These considerations led to the development of *labeled LDA* (L-LDA). L-LDA models every document as a bag of words with a bag of labels, assigning a document's words to its labels rather than to a latent semantic space (Ramage et al., 2009a,b). In this manner, labeled LDA is a generative model for multi-labeled corpora that marries the multi-label supervision common to modern text datasets with the word-assignment ambiguity resolution of the LDA family of models. L-LDA therefore provides a means to identify category-domains even when there are ambiguous and multiple category-designations affixed to the documents.

Labeled LDA enables us to identify the terms our categories reference. And by exploring the changing load of word sets over time, we learn several things. First, we learn what a category semantically entails. For example, using labeled LDA, we studied an entire field's corpus and sought to learn the language signature of each author. In this way, we could then return to a multi-authored paper and discern who likely contributes most of the content (Johri et al., 2011). Second, we learn how a category-linked bag of words increases and declines in relevance. For example, we applied labeled LDA to an entire field's corpus in the effort to learn what each nation's topic signature was in that research field and how its relevance changed over time (Gomez and McFarland, 2012). The nation-specific topics had a highly variable load on the corpus, suggesting some nations were defining the field more than others and that some were rising in relevance (e.g., China). Third, we have a means of identifying whether the specific words associated with a category change over time. For example, when we looked more closely at the words in nation-topics, we saw the most salient terms were nation specific ones (e.g., referring to physical landmarks and nation-specific events), and then later, they became more regional in nature.

By utilizing corpus-provided labels, labeled LDA laid bare further complications that caused us to revisit and improve the underlying text model's applicability for our analyses. In many corpora, the set of labels assigned to documents can change over time (new category schemas emerge), and many labels are either too coarse or too fine-grained. For example, in ProQuest, dissertations are affixed with subject categories, and those for sociology have fewer available subject designations than does physics. Are the languages associated with each subject comparable if the labels correspond to different units of analysis? It is likely that fields are not categorized at the same level of definition, and this may be an error. Nonetheless, supervised topic models like L-LDA will only be as accurate as the labels afforded them and affix one topic per label. Facing these issues, we developed *partially latent Dirichlet allocation*, or P-LDA, so as to acquire comparable topic-definitions. P-LDA is a topic model that is a mixture of LDA and L-LDA – it is analogous to L-LDA except that it allows more than one latent topic per label (Ramage, 2011). As such, one can input subject categories as labels, and define a number of sub-clusters of language used within each label (i.e., subfields). We term this *parametric P-LDA*.

A non-parametric or *unsupervised extension of P-LDA* allows the number of topics within categories to vary. There, one can use perplexity analysis to decide an optimal number of topics arising within each label, and as such, use a mixed approach to validation. However, it is computationally expensive to run such models and infer a cutoff level for the number of topics within them (Ramage, 2011). This makes it an infeasible procedure for studying large corpora.[7] An implementable alternative is a to run P-LDA where the number of topics within subfield categories is predefined at a fixed number that captures most of the language dimensions characterizing texts within subject categories. Since a higher number of topics results in small, meaningless dimensions, we found that most subjects were well represented by 8–12 topics with a few being noise, but not overwhelmingly so (Ramage, 2012).

In some instances, multiple, equally valid category schemes pertain to a corpus. In our studies, we found that authors affixed dissertations in ProQuest with an average of 1.6 of the 262 available subject categories. Some fields are associated with more labels than others, suggesting the category scheme may not be perfect. The National Research Council provides an equally valid, alternative categorization scheme – containing 52 coarser categories that exhibit less variation in size and scope across fields. Which is the better categorization scheme? In many regards, they are equally valid. Rather than view this as a problem, we see it as an opportunity to improve the validity of our results. We view the two schemas as if they are two informed experts with slightly different opinions. Where the two schemas align, or where the two (or more) experts agree, we see greater evidence of consensus. To accomplish this, we run P-LDA for each categorization schema at various numbers of subtopics (1–20) per label and identify the level at which their topic solutions most correlate (~12 topics for ProQuest and 8 for NRC had a .98 correlation). This correspondence reveals the latent topics undergirding the various community-categorizations being applied by authors (Ramage, 2012).

Once we identify category-domains that we trust, we again repeat our focus on dynamics and change, but also begin to look at ways to demonstrate knowledge flows across papers, fields, and even faculty careers (see works by Nallapati, Ramage, Anderson). Here, the research begins to approach the concept of network-domains more fully, as network connections become pathways and conduits across which we follow language transfers. Daniel Ramage's (2011) work is perhaps the culmination of the aforementioned efforts. He uses the supervised form of PLDA described above and identifies topic solutions that agree across two distinct subject-category schemas.

One benefit of L-LDA and PLDA is that, after training the model to learn word distributions for each label, one can directly measure how much any given document draws on the language of each field. In this manner, one can assess whether documents failed to apply a label when they should have. For example, this article and journal is labeled as "sociology," but it clearly draws on topics specific to the field of computational linguistics. As such, we can use the topic-loadings to identify documents whose labeling may be incorrect. Or, seen another way, we can see this paper as an instance of a document from sociology borrowing the language of computational linguistics.

Framing cross-label topic usage as borrowing, we can create a matrix of language-borrowing across fields. We perform this analysis and find interdisciplinarity is a highly directional process of knowledge transfer, and that fields assume different roles in the pattern of transfers. In particular, we find that methodological (statistics, math), technological (computer science), and abstract (philosophy) topical areas are all borrowed by other fields and not vice versa. Hence, fields like biology draw on the topics of statistics and computer science more than the field of statistics and computer science draw on the topics of biology. And certain fields like classics and

[7] Recent algorithmic advances (Wang et al., 2011) hold promise for scaling non-parametric extensions of PLDA moving forward.

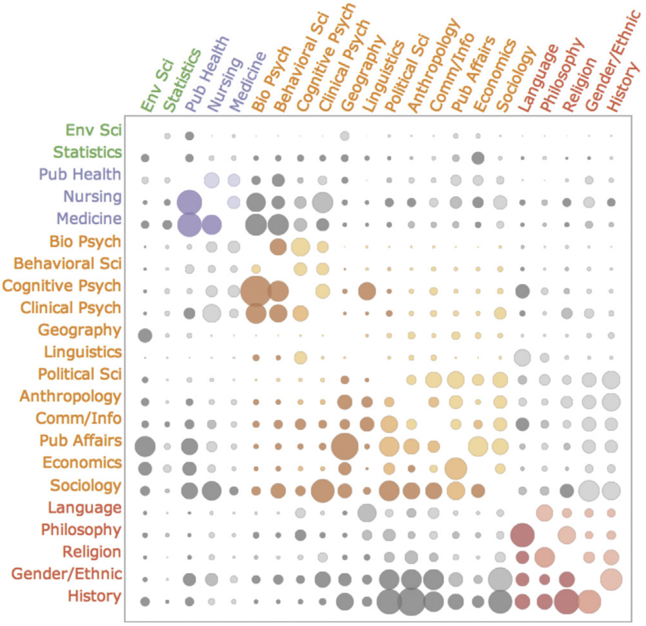

Fig. 2. Language borrowing across fields (2000–2010 dissertation abstracts).

animal studies seem to be unrelated to other fields, both diminishing in size and relevance over time.

Hence, when a corpus contains documents that are labeled with community membership designations (like academic fields and subjects), we can use applications of L-LDA and PLDA to identify patterns of knowledge flow. In this manner, we come close to not only identifying language-communities (or thought communities), but their dynamic interrelation through language transfer over time.[8] Readers can view our efforts to visualize these inter-field relations for all subjects in previously published work that analyzed over 1 million dissertation abstracts (Chuang et al., 2012c; Ramage, 2012). Fig. 2 presents a sub-matrix of language flows across fields, with a particular focus on the fields most related to sociology. The value of cell i,j represents the fraction of words in column j that are incorporated from row i. The area of the circle shows the degree of language sharing, and the darker circle indicates which i,j or j,i relation exports more (i,i diagonals, or internal language referencing, is not shown). Notably, the figure shows large circles for sociology, suggesting it exports and imports language.

In Fig. 3, we sum these flows in line plots for a small set of academic fields so as to illustrate the extent to which each draws on the language specific to their own subject (diagonal of Fig. 2) or to that of any external field (column of Fig. 2). The *y*-axis in each figure measures the fractional

[8] Our project also identified how topics flowed across citations (Nallapati and Manning, 2010; Nallapati et al., 2011), and how persons flowed across topics (Anderson et al., 2012). In the former study, Nallapati first identified topics, and then he followed citation relations that were within the same topic to develop a sort of intellectual history of a line of research. In the second study, Anderson first identified topics, and then he mapped out how faculty move across these topics during their career. Both lines of work portray careers – one is a topic's career across documents, and the other is a faculty's career across research areas.

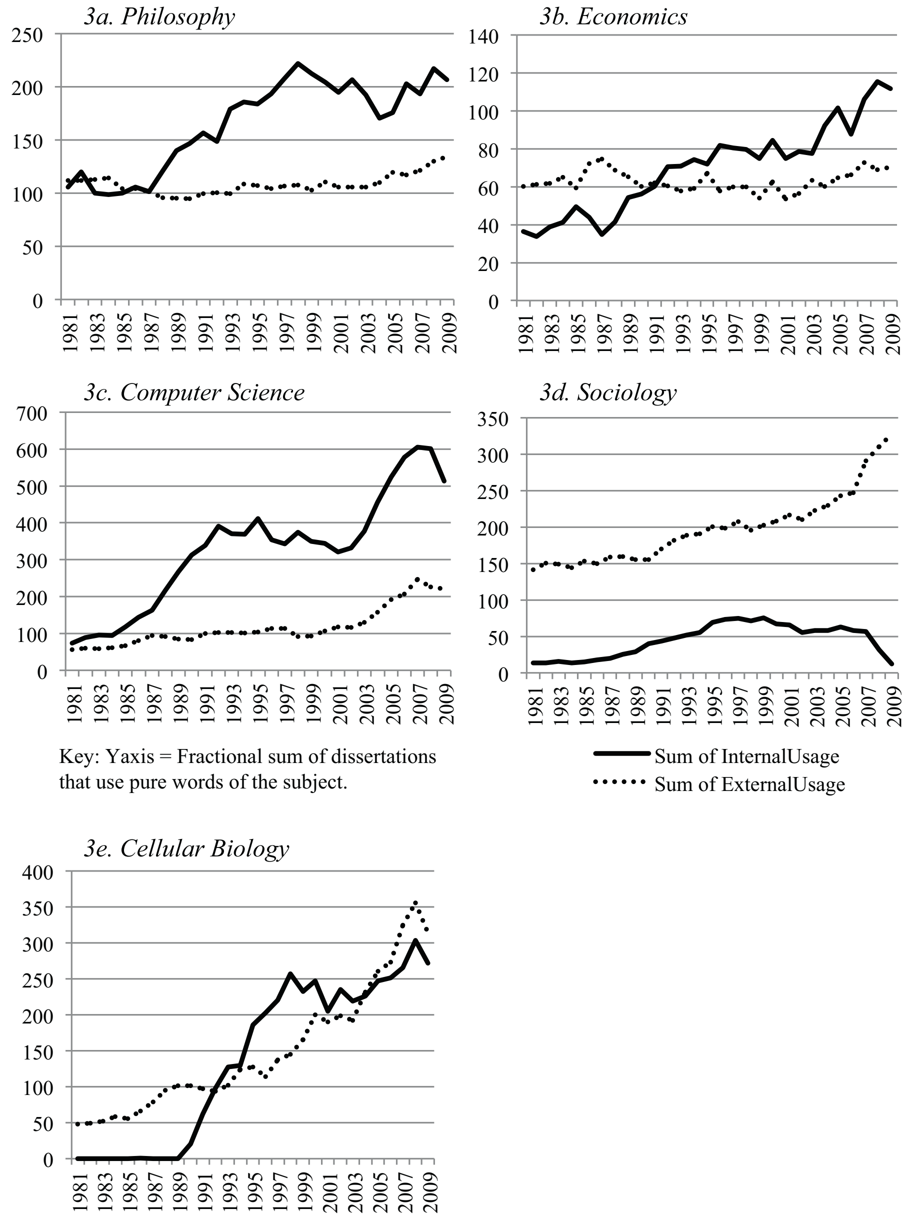

Fig. 3. Internal versus external word usage by field (dissertation abstracts).

sum of words in each field's dissertations that draw on the words specific (or unique) to their subject category or to that of external ones. The *x*-axis shows how these relations change by year. Fields that rely mostly on their own language are more paradigmatic and inward-focused. Fields that rely on external language borrow more heavily from other fields and are outward-focused.

In Table 3, we cross-classify these states and hypothesize the intellectual state of various fields depicted in our figures. Cell 1 is a field exploiting what it knows and solving the problems of its

Please cite this article in press as: McFarland, D.A., et al., Differentiating language usage through topic models. Poetics (2013), http://dx.doi.org/10.1016/j.poetic.2013.06.004

Table 3
Field states.

	Relies on internal language	
	High	Low
Relies on external language		
High	2. Innovating paradigm	3. Discovery area (explore mode)
Low	1. Insular paradigm (exploit mode)	4. Dying area

paradigm, much like how Kuhn (1970) described fields engaging in normal science. Cell 2 reflects a field developing a paradigm and innovating it. Cell 3 concerns a field that is exploring other knowledge, and seeking solutions elsewhere – a field forever in discovery mode without forming a clear paradigm of thought. And last, cell 4 concerns a field that is likely dying since it has diminished in both internal and external relevance.

Our figures suggest that philosophy, economics and computer science may be increasingly inwardly focused and draw on their own concepts (#1). Biophysics is a quickly expanding interdisciplinary field that seems to be drawing both on its own language and that of other fields, building a lexicon and expanding its foci of research into other topics (#2). And sociology is a discovery area (non-paradigmatic) where the dissertations draw more on the language of other fields than its own (#3).

Table 4
Summary table on types of topic modeling procedures.

Type of domain	Type of model	When used	Validation procedures
Latent domains (unsupervised)	Latent Dirichlet allocation (LDA)	To discover N number of latent topics that persons may or may not be aware of.	*Statistical properties*: Topic load, entropy, perplexity *Ground truth*: Visual inspection, expert confirmation, expert agreement.
Manifest domains (supervised)	Naïve bayes	When 1 label per document. Can identify topic or set of words specific to each label.	It is as accurate as the labels/categories afforded.
	Labeled-LDA	When >1 label per document. Can identify topic or set of words specific to each label, even when multiple labels are affixed to each document.	Category schema is assumed to be correct (akin to expert) – category confirms topic.
Manifest domains with latent subareas (mixed)	Parametric, partially latent Dirichlet allocation (PLDA)	When >1 label per document and labels are assumed to entail latent sub-topics (user decides # of topics within each label).	Category schema is assumed correct to a point, and reliance is on statistical properties and ground truth assessments thereafter.
	Non-parametric, partially latent Dirichlet allocation	When >1 label per document and labels are assumed to entail latent sub-topics (model discovers # of topics within each label).	Can use multiple category schema, and the optimal solution is one that agrees with both schemas.

3. Summary

This paper summarizes much of what our project has learned about topic modeling after having used it in various forms to study the differentiation of knowledge in academic corpora. In doing our research, we learned that topic models require careful thought and revision if they are to be successfully applied to social science research questions. Hence, we developed a variety of validation procedures, created new forms of supervised (and partially unsupervised) LDA, and found ways to make topic modeling suitable to the study of language-domains and their dynamic interrelation. Our efforts were far from exhaustive, but perhaps they will provide some guidance to future sociologists hoping to apply topic models to their own research questions.

Below is a summary table (Table 4) of the types of models we utilized and the validation procedures discussed above. For more specific descriptions of each type of model and how they were applied, please see the referenced material. This table merely relates the types of topic models one can use, the type of language domain they will reveal, when we thought it best to use it, and what sort of validation procedures can be put in place to develop greater trust in one's results.

Acknowledgements

We thank members of the Mimir project for comments. These data were collected by the Mimir Project conducted at Stanford University by Daniel A. McFarland, Dan Jurafsky, Chris Manning, and Woody Powell. This project has been generously funded by the Office of the President at Stanford University and by NSF Award #0835614.

References

Anderson, A., McFarland, D.A., Jurafsky, D., 2012. Towards a Computational History of the ACL: 1980–2008. In: ACL Workshop on Rediscovering 50 Years of Discoveries. Available at:http://www.stanford.edu/~jurafsky/anderson12.pdf.

Blei, D., McAuliffe, J., 2007. Supervised topic models. Neural Information Processing Systems 21, Available athttp://books.nips.cc/nips20.html.

Blei, D.M., Ng, A., Jordan, M., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models. In: Pohoreckhy, A., Bottou, L., Littman, M.L. (Eds.), Proceedings of the International Conference on Machine Learning. pp. 113–120.

Buntine, W., Jakulin, A., 2004. Applying discrete PCA in data analysis. In: Chickering, M., Halpern, J. (Eds.), Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI'04). AUAI Press, Arlington, VA, pp. 59–66.

Chuang, J., Ramage, D., Heer, J., Manning, C.D., 2009. Stanford Dissertation Browser. , Available athttp://nlp.stanford.edu/projects/dissertations/.

Chuang, J., Manning, C.D., Heer, J., 2012a. Termite: visualization techniques for assessing textual topic models. In: Proceedings of the International Working Conference on Advanced Visual Interfaces. Available at:http://vis.stanford.edu/papers/termite.

Chuang, J., Ramage, D., Manning, C.D., Heer, J., 2012b. Interpretation and trust: designing model-driven visualizations for text analysis. In: Proceedings of the ACM Human Factors in Computing Systems (CHI). Available athttp://vis.stanford.edu/papers/designing-model-driven-vis.

Chuang, J., Ramage, D., McFarland, D.A., Manning, C.D., Heer, J., 2012c. Large-Scale Examination of Academic Publications Using Statistical Models. In: Advanced Visual Interfaces Workshop (AVI Workshop 2012). Available athttp://131.107.65.14/en-us/events/acva/chuang.pdf.

Dagan, I., Glickman, O., Magnini, B., 2006. The PASCAL recognising textual entailment challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (Eds.), Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment,. Springer, pp. 177–190.

Dalrymple, M., 2001. Lexical Functional Grammar. Academic Press, New York.

Erosheva, E., Fienberg, S., Lafferty, J., 2004. Mixed membership models of scientific publications. Proceedings of the National Academy of Sciences of the United States of America (PNAS) 101, 5220–5227.

Fine, G.A., 1987. With The Boys: Little League Baseball and Preadolescent Culture. University of Chicago Press, Chicago.

Fleck, L., 1979 (1935). Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago.

Franzosi, R., 2010. Quantitative Narrative Analysis. Sage, Beverly Hills, CA.

Gomez, C., McFarland, D., 2012. Language, Knowledge, and Power in the International System—A Linguistic Analysis of Published Political Science Research by Nation-state from 1991 to 2008. Interdisciplinary Workshop on Information and Decision in Social Networks, Massachusetts Institute of Technology.

Goodman, J.T., 2001. A bit of progress in language modeling. Computer Speech & Language 15 (4) 403–434.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America (PNAS) 1, 5228–5235.

Griswold, W., 2008. Regionalism and the Reading Class. University of Chicago Press, Chicago.

Hacking, I., 1999. The Social Construction of What? Harvard University Press, Cambridge, MA.

Hall, D., Jurafsky, D., Manning, C., 2008. Studying the history of ideas using topic models. In: Laputa, M., Ng, H.T. (Program Co-Chairs), Proceedings of the 2008 Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, pp. 363–371.

Johri, N., Ramage, D., McFarland, D.A., Jurafsky, D., 2011. A study of academic collaboration in computational linguistics with latent mixtures of authors. In: Zervanou, K., Lendavi, P. (Eds.), Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics, Stroudsburg, PA, pp. 124–132.

Jurafsky, D., Martin, J.H., 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd ed. Pearson Prentice Hall, Upper Saddle River, NJ.

Kirchner, C., Mohr, J.W., 2010. Meanings and relations: an introduction to the study of language, discourse and networks. Poetics 38, 555–566.

Knorr-Cetina, K., 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge, MA.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., 2007. Moses: open source toolkit for statistical machine translation. In: Ananiadou, S. (Program Chair), Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Stroudsburg, PA, pp. 177–180.

Kübler, S., McDonald, R., Nivre, J., 2009. Dependency Parsing. Vol. 2 of Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA.

Kuhn, T., 1970. The Structure of Scientific Revolutions, 2nd ed. University of Chicago Press, Chicago.

Latour, B., 1987. Science in Action. Harvard University Press, Cambridge, MA.

Lewis, D.D., Yang, Y., Rose, T.G., Dietterich, G., Li, F., Li, F., 2004. RCV1: a new benchmark collection for text categorization research. Journal of Machine Language Research (JMLR) 5, 361–397.

Lin, D., Pantel, P., 2001. DIRT-discovery of inference rules from text. In: Provost, F., Srikant, R. (Program Chairs), Proceedings of the Seventh ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, pp. 323–328.

Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

Marcus, M.P., Marcinkiewicz, M.A., Santorini, B., 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19 (2) 313–330.

McCallum, A., Corrada-Emmanuel, A., Wang, X., 2004. The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. Department of Computer Science, University of Massachusetts, Amherst, Technical Report.

McCallum, A., Nigam, K., 1998. A comparison of event models for naive Bayes text classification. In: AAI-98 Workshop on Learning for Text Categorization, Volume 7. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf.

McLean, P.D., 1998. A frame analysis of favor seeking in the renaissance: agency, networks, and political culture. American Journal of Sociology 104, 51–91.

Misra, H., Cappe, O., Yvon, F., 2008. Using LDA to detect semantically incoherent documents. In: Clark, A., Toutanova, K. (Conference Chairs), Proceedings of the Twelfth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Stroudsburg, PA, pp. 41–48.

Mohr, J.W., 1998. Measuring meaning structures. Annual Review of Sociology 24, 345–370.

Nadel, S.F., 1957. The Theory of Social Structure. Cohen and West, London.

Nallapati, R., Manning, C.D., 2010. TopicFlow Model: Unsupervised Learning of Topic Specific Influences of Hyperlinked Documents. In: Neural Information Processing Systems Workshop on Machine Learning for Social Computing. Available at:In: https://sites.google.com/site/rameshnallapati/research/publications-by-year.

Nallapati, R., McFarland, D., Manning, C.D., 2011. TopicFlow model: unsupervised learning of topic-specific influences of hyperlinked documents. Journal of Machine Learning Research (JMLR) Workshop and Conference Proceedings 15, 543–551 (AISTATS 2011).

Newman, D., Asuncion, A., Chemudugunta, C., Kumar, V., Smyth, P., Steyvers, M., 2006. Exploring Large Document Collections Using Statistical Topic Models. In: KDD-2006 Demo Session. Available at:In: http://www.ics.uci.edu/∼asuncion/pubs/KDD_06_DEMO.pdf.

Newman, D., Karimi, S., Cavedon, L., 2009. External evaluation of topic models. In: Kay, J., Thomas, P., Trotman, A. (Eds.), Proceedings of the 14th Australasian Document Computing Symposium. School of Information Technologies, University of Sydney, pp. 11–18.

O'Connor, B., Bamman, D., Smith, N.A., 2010. Computational Text Analysis for Social Science: Model Assumptions and Complexity. In: NIPS Workshop on Computational Social Science. Available at:In: http://people.cs.umass.edu/∼wallach/workshops/nips2011css/papers/OConnor.pdf.

Pollard, C.J., Sag, I.A., 1994. Head-Driven Phrase Structure Grammar. University of Chicago Press, Chicago.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ.

Radev, D.R., Muthukrishnan, P., Qazvinian, V., 2009. The ACL anthology network corpus. In: Kan, M.-Y., Teufel, S. (General Chairs), Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, Association for Computational Linguistics, Stroudsburg, PA, pp. 54–61.

Ramage, D., 2011. Studying People, Place, and the Web with Statistical Text Models. Computer Science, Stanford University (Doctoral Thesis).

Ramage, D., Hall, D., Nallapati, R., Manning, C.D., 2009a. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Koehn, P., Mihalcea, R. (Program Chairs), Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, pp. 248–256.

Ramage, D., Rosen, E., Chuang, J., Manning, C.D., McFarland, D.A., 2009b. Topic Modeling for the Social Sciences. In: Neural Information Processing Systems (NIPS) 2009 Workshop on Applications for Topic Models. Available at:In: http://nlp.stanford.edu/∼dramage//papers/tmt-nips09.pdf.

Ramage, D., Dumais, S., Liebling, D., 2010. Characterizing Microblogs with Topic Models. In: International AAAI Conference on Weblogs and Social Media (ICWSM). Available at:In: http://nlp.stanford.edu/∼dramage//papers/twitter-icwsm10.pdf.

Ramage, D., Manning, C.D., Dumais, S., 2011. Partially labeled topic models for interpretable text mining. In: Conference on Knowledge Discovery and Data Mining (KDD 2011). Available at:In: http://nlp.stanford.edu/∼dramage//papers/pldp-kdd11.pdf.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smith, P., 2004. The author-topic model for authors and documents. In: Meek, C., Halpern, J. (Eds.), Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, Arlington, VA, pp. 487–494.

Salton, G., McGill, M., 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York.

Snow, D.A., Rochford, E.B., Worden, S.K., Benford, R.D., 1986. Frame alignment processes, micro-mobilization and movement participation. American Sociological Review 51, 464–481.

Strehl, A., Ghosh, J., Mooney, R., 2000. Impact of similarity measures on webpage clustering. In: Bollaker, K. (Program Chair), AAAI Workshop on Artificial Intelligence for Web Search, Association for Advancement of Artificial Intelligence, Palo Alto, CA, pp. 58–64.

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet processes. Journal of the American Statistical Association 101, 1566–1581.

Vicari, S., 2010. Measuring collective action frames: a linguistic approach to frame analysis. Poetics 38, 504–525.

Vogel, A., Jurafsky, D., 2012. He Said, She Said: Gender in the ACL Anthology. In: ACL Workshop on Rediscovering 50 Years of Discoveries. Available at:In: http://www.stanford.edu/∼jurafsky/vogeljurafsky12.pdf.

Wallach, H.M., Mirray, I., Salakhutdinov, R., Mimno, D., 2009. Evaluation methods for topic models. In: Bottou, L., Littman, M. (Eds.), Proceedings of the 26th International Conference on Machine Learning. pp. 1105–1112.

Wang, C., Paisley, J., Blei, D., 2011. Online variational inference for the hierarchical Dirichlet process. In: Artificial Intelligence and Statistics 2011. Available at:In: http://www.cs.cmu.edu/∼chongw/papers/WangPaisleyBlei2011.pdf.

Weber, R.P., 1985. Basic Content Analysis. Sage, Beverly Hills, CA.

White, H., 1992. Identity and Control: A Structural Theory of Social Action. Princeton University Press, Princeton, NJ.

Daniel A. McFarland is an Associate Professor of Education, and Associate Professor by courtesy of Sociology and Organizational Behavior, at Stanford University. His research on social dynamics focuses on the coevolution of social networks and cultural systems.

Daniel Ramage is a research scientist at Google. His work focuses on statistical techniques for modeling language and user behavior.

Jason Chuang is a post-doctoral research in Computer Science at the University of Washington. His work focuses on the process of visual data analysis: combining information visualization, human-centered design, and machine learning to create effective workflows for making sense of large and complex data.

Jeffrey Heer is an Associate Professor of Computer Science at the University of Washington, where he works on human-computer interaction, visualization and social computing. His research investigates the perceptual, cognitive and social factors involved in making sense of large data collections, resulting in new interactive systems for visual analysis and communication.

Christopher Manning is a Professor of Computer Science and Linguistics at Stanford University. He is an AAAI Fellow and an ACL Fellow, and he has coauthored leading textbooks on statistical approaches to natural language processing and information retrieval. His recent research concentrates on machine learning approaches to various computational linguistic problems – including parsing, semantic similarity, and textual inference.

Dan Jurafsky is Professor of Linguistics, and Professor by courtesy of Computer Science, at Stanford University. His research in computational linguistics focuses on statistical models of human and machine language processing, particularly the automatic extraction of meaning and the application of natural language processing to the social sciences.