Applications of Topic Models

Jordan Boyd-Graber
Department of Computer Science, UMIACS, Language Science
University of Maryland[1]
jbg@umiacs.umd.edu

Yuening Hu
Google, Inc.[2]
ynhu@google.com

David Mimno
Information Science
Cornell University
mimno@cornell.edu

now

Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

J. Boyd-Graber, Y. Hu and D. Mimno. *Applications of Topic Models*. Foundations and Trends® in Information Retrieval, vol. 11, no. 2-3, pp. 143–296, 2017.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-308-9

Foundations and Trends® in Information Retrieval

Volume 11, Issue 2-3, 2017

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2017, Volume 11, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in Information Retrieval
Vol. 11, No. 2-3 (2017) 143–296

DOI: 10.1561/1500000030

Applications of Topic Models

Jordan Boyd-Graber
Department of Computer Science, UMIACS, Language Science
University of Maryland[1]
jbg@umiacs.umd.edu

Yuening Hu
Google, Inc.[2]
ynhu@google.com

David Mimno
Information Science
Cornell University
mimno@cornell.edu

[1]Work completed while at University of Colorado
[2]Work completed while at Yahoo!

Contents

1 The What and Wherefore of Topic Models **2**

1.1 Tell Me about Your Haystack 2
1.2 What is a Topic Model 5
1.3 Foundations 6
1.4 Latent Dirichlet Allocation 11
1.5 Inference 13
1.6 The Rest of this Survey 19

2 Ad-hoc Information Retrieval **21**

2.1 Document Language Modeling 23
2.2 Topic-based Document Language Models 26
2.3 Query Expansion 27
2.4 Beyond Relevance—Search Personalization 34
2.5 Summary 36

3 Evaluation and Interpretation **38**

3.1 Displaying Topics 38
3.2 Labeling Topics 40
3.3 Displaying Models 43
3.4 Evaluation, Stability, and Repair 45
3.5 Summary 47

4 Historical Documents **49**
4.1 Newspapers 50
4.2 Historical Records 54
4.3 Scholarly Literature 56
4.4 Summary 58

5 Understanding Scientific Publications **60**
5.1 Understanding Fields of Study 62
5.2 How Fields Change 64
5.3 Innovation 66
5.4 Summary 68

6 Fiction and Literature **69**
6.1 Topic Models in the Humanities 69
6.2 What is a Document? 71
6.3 People and Places 72
6.4 Beyond the Literal 76
6.5 Comparison to Stylometric Analysis 78
6.6 Operationalizing "Theme" 78
6.7 Summary 79

7 Computational Social Science **81**
7.1 Topic Models for Qualitative Analysis 84
7.2 Sentiment Analysis 84
7.3 Upstream and Downstream Models 86
7.4 Understanding Stance and Polarization 87
7.5 Social Networks and Media 88
7.6 Summary 91

8 Multilingual Data and Machine Translation **92**
8.1 Document-level Alignment from Multilingual Corpora . . . 94
8.2 Word-level Alignment from Lexical Data 96
8.3 Alignment from Parallel Corpora and Lexical Information . 98
8.4 Topic Models and Machine Translation 99
8.5 The Components of Statistical Machine Translation 100
8.6 Topic Models for Phrase-level Translation 102

8.7 Topic Models for Sentence-level Language Modeling 106
8.8 Reordering with Topic Models 110
8.9 Beyond Domain Adaptation 111
8.10 Summary 112

9 Building a Topic Model **113**
9.1 Designing a Model 114
9.2 Implementing the Model 117
9.3 Debugging and Validation 123
9.4 Communicating Your Model 125
9.5 Summary 126

10 Conclusion **127**
10.1 Coping with Information Overload 127
10.2 Deeper Representations 128
10.3 Automatic Text Analysis for the People 129
10.4 Coda 131

References **132**

Abstract

How can a single person understand what's going on in a collection of millions of documents? This is an increasingly common problem: sifting through an organization's e-mails, understanding a decade worth of newspapers, or characterizing a scientific field's research. Topic models are a statistical framework that help users understand large document collections: not just to find individual documents but to understand the general themes present in the collection.

This survey describes the recent academic and industrial applications of topic models with the goal of launching a young researcher capable of building their own applications of topic models. In addition to topic models' effective application to traditional problems like information retrieval, visualization, statistical inference, multilingual modeling, and linguistic understanding, this survey also reviews topic models' ability to unlock large text collections for qualitative analysis. We review their successful use by researchers to help understand fiction, non-fiction, scientific publications, and political texts.

J. Boyd-Graber, Y. Hu and D. Mimno. *Applications of Topic Models*. Foundations and Trends® in Information Retrieval, vol. 11, no. 2-3, pp. 143–296, 2017.

DOI: 10.1561/1500000030.

1

The What and Wherefore of Topic Models

Imagine that you are an intrepid reporter with an amazing scoop: you have twenty-four hours of exclusive access three decades of e-mails sent within a corrupt corporation. You know there's dirt and scandal there, but it has been well-concealed by the corporation's political friends. How are you going to understand this haystack well enough to explain it to your devoted readers under such a tight deadline?

1.1 Tell Me about Your Haystack

Unlike the vignette above, interacting with large text data sets is often posed as a needle in a haystack problem. The poor user—faced with documents that would take a decade to read—is looking for a single needle: a document (or at most a handful of documents) that matches what the user is looking for: a "smoking gun" e-mail, the document that best represents a concept [Salton, 1968] or the answer to a question [Hirschman and Gaizauskas, 2001].

These questions are important. The discipline of information retrieval is built upon systematizing, solving, and evaluating this problem. Google's search service is built on the premise of users typing a few

keywords into a search engine box and seeing quick, consistent search results. However, this is not the only problem that confronts those interacting with large text datasets.

A different, but related problem is *understanding* large document collections, common in science policy [Talley et al., 2011], journalism, and the humanities [Moretti, 2013a]. The haystack has more than one precious needle. At the risk of abusing the metaphor, *sometimes you care about the straw.* Instead of looking for a smoking gun alerting to you some crime that was committed, perhaps you are looking for a sin of omission: did this company never talk about diversity in its workforce? Instead of a single answer to a question, perhaps you are looking for a diversity of responses: what are the different ways that people account for rising income inequality? Instead of looking for one document, perhaps you want to provide population level statistics: what proportion of Twitter users have ever talked about gun violence?

At first, it might seem that answering these questions would require building an extensive ontology or categorization scheme. For every new corpus, you would need to define the buckets that a document could fit into, politely ask some librarians and archivists to put each document into the correct buckets, perhaps automate the process with some supervised machine learning, and then collect summary statistics when you are done.

Obviously, such laborious processes are possible—they have been done for labeling congressional speeches[1] and understanding emotional state [Wilson and Wiebe, 2005]—and remain an important part of social science, information science, library science, and machine learning. But these processes are not always possible, fast, or even the optimal outcome if we had infinite resources. First, they require a significant investment of time and resources. Even creating the *list* of categories is a difficult task and requires careful deliberation and calibration. Even if it were possible, a particular question might not warrant the time or effort: the œuvre of a minor author (only of interest to a few), or the tweets of a day (not relevant tomorrow).

[1] `www.congressionalbills.org/`

Table 1.1: Five topics from a twenty-five topic model fit on Enron e-mails. Example topics concern financial transactions, natural gas, the California utilities, federal regulation, and planning meetings. We provide the five most probable words from each topic (each topic is a distribution over all words).

Topic	Terms
3	trading financial trade product price
6	gas capacity deal pipeline contract
9	state california davis power utilities
14	ferc issue order party case
22	group meeting team process plan

This survey explores the ways that humans and computers make sense of document collections through tools called topic models. Topic models allow us to answer big-picture questions quickly, cheaply, and without human intervention. Once trained, they provide a framework for humans to understand document collections both directly by "reading" models or indirectly by using topics as input variables for further analysis. For readers already comfortable with topic models, feel free to skip this chapter; we will mostly cover the definitions and implementations of topic models.

The intended audience of this book is a reader with some knowledge of document processing (e.g., knows what "tokens" and "documents" are), basic understanding of some probability (e.g., what a distribution is), and interested in many application domains. We discuss the information needs of each application area, and how those specific needs affect models, curation procedures, and interpretations.

By the end of the book (Chapter 9), we hope that readers will be excited enough to attempt to embark on building their own topic models. In this chapter, we go deeper into more of the implementation details. Readers who are already topic model experts will likely not learn much technically, but we hope our coverage of diverse applications will expose a topic modeling expert to models and approaches they had not seen before.

Yesterday, SDG&E filed a motion for adoption of an electric procurement cost recovery mechanism and for an order shortening time for parties to file comments on the mechanism. The attached email from SDG&E contains the motion, an executive summary, and a detailed summary of their proposals and recommendations governing procurement of the net short energy requirements for SDG&E's customers. The utility requests a 15-day comment period, which means comments would have to be filed by September 10 (September 8 is a Saturday). Reply comments would be filed 10 days later.

Topic	Probability
9	0.42
11	0.05
8	0.05

Figure 1.1: Example document from the Enron corpus and its association to topics. Although it does not contain the word "California", it discusses a single California utility's dissatisfaction with how much it is paying for electricity.

1.2 What is a Topic Model

Returning to our motivating example, consider the e-mails from Enron, the prototypical troubled corporation of the turn of the century. A source has provided you with a trove of emails, and your editor is demanding an article by yesterday. You know that wrongdoing happened, but you do not know who did it or how it was planned and carried out. You have suspicions (e.g., around the California energy spot market), but you are curious about other skeletons in the closet and you are highly motivated to find them.

So you run a topic model on the data. True to its name, a topic model gives you "topics", each of which is a ranking of all the distinct words in the e-mails by relevance to a topic. Taking the top five most relevant words in each topic results in collections of words that make sense together (Table 1.1). For example, one topic seems to have words relating to finance and trading. Another seems to involve to gas pipelines, their capacity, and deals or contracts relating to those pipelines. This all makes sense: Enron was an energy trading company. Others seem to involve language used in any business, such as meetings and plans.

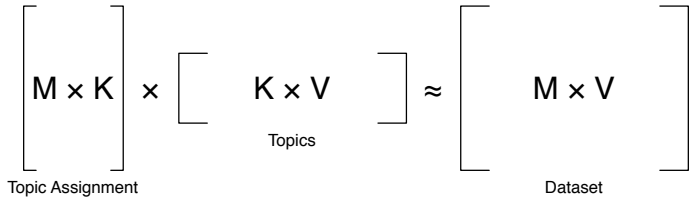

Figure 1.2: A matrix formulation of finding K topics for a dataset with M documents and V unique words. While this view of topic modeling includes approaches such as latent semantic analysis (LSA, where the approximation is based on SVD), we focus on probabilistic techniques in the rest of this survey.

The first half of a topic model connects topics to a jumbled "bag of words". When we say that a topic is about X, we are manually assigning a *post hoc* label (more on this in Chapter 3.1). It remains the responsibility of the human consumer of topic models to go further and make sense of these piles of straw (we discuss labeling the topics more in Chapter 3).

Making sense of one of these word piles by itself can be difficult. The second half of a topic model links topics to individual documents. For example, the document in Figure 1.1 is about a California utility's reaction to the short-term electricity market and exemplifies Topic 9 from Table 1.1. Considering examples of documents that are strongly connected to a topic, along with the words associated with the topic, can give us a more complete representation of the topic. If we get a sense that Topic 9 is of interest, we can explore deeper to find other documents.

1.3 Foundations

You might notice that we are using the general term "topic model". There are many mathematical formulations of topic models and many algorithms that learn the parameters of those models from data. Although we will focus on particular models and algorithms, we choose

our terminology to emphasize that the similarities between formulations, models, and algorithms are often greater than their differences.

Topic modeling began with a linear algebra approach [Deerwester et al., 1990] called latent semantic analysis (LSA): find the best low rank approximation of a document-term matrix (Figure 1.2). While these approaches have seen a resurgence in recent years [Anandkumar et al., 2012, Arora et al., 2013], we focus on probabilistic approaches [Hofmann, 1999a, Papadimitriou et al., 2000, Blei et al., 2003], which are intuitive, work well, and allow for easy extensions (as we see later in many of our later chapters).

The two foundational probabilistic topic models are latent Dirichlet allocation [Blei et al., 2003, LDA] and probabilistic latent semantic analysis [Hofmann, 1999a, pLSA]. We describe the former in significant detail in Chapter 1.4, but we want to take a moment to address some of the historical connection between these two models.

pLSA was historically first and laid the foundation for LDA. pLSA was used extensively in many applications such as information retrieval. However, this survey focuses on LDA because more researchers have not just *used* LDA—they have also *extended* it. LDA is not just widely used, but it is also widely modified. Because of these prolific modifcations, we focus on the mechanics of LDA, which many researchers have used as the foundations of new models. However, as we explain below (Chapter 1.5.4), the similarities between pLSA and LDA outweigh the differences.

In any technical field it is common for general terms to take on specific, concrete meanings, and this can be a source of confusion. In topic modeling the word "topic" takes on the specific meaning of a probability distribution over words, while still alluding the to more general meaning of a theme or subject of discourse. Because other areas of information retrieval have similarly developed specific meanings for the word "topic", we distinguish them here. The most common definition is a specific information need, as in the TREC evaluation corpora developed by NIST [Voorhees and Harman, 2005]. TREC topics are generally much more specific than topic model topics, and may relate to particular aspects or perspectives on a subject. An example from

the 2003 TREC Robust Track is "Identify positive accomplishments of the Hubble telescope since it was launched in 1991" [Voorhees, 2003]. Similarly to information retrieval, the related field of topic detection and tracking also has a specific technical definition of "topic" [Allan, 2002]. In TDT, a "topic" is usually closer to an event or an individual story. In contrast, topic models tend to identify more abstract latent factors. For example, a TDT topic might include an earthquake in Haiti, whereas a topic model might represent the same event as a combination of topics such as Haiti, natural disasters, and international aid.

There has been some work on using topic models to detect emerging events by searching for changes in topic probability [AlSumait et al., 2008]. But these methods tend to identify mainly the fact that an event has occurred, without necessarily identifying the specific features of that event. Other work has found that more lexically specific methods than topic models are best for identifying memes and viral phrases [Leskovec et al., 2009].

1.3.1 Probabilistic Building Blocks

In probabilistic models we want to find values for unobserved model variables that do a good job of explaining the observed data. The first step in inference is to turn this process around, and assert a way to generate data given model variables. Probabilistic models thus begin with a generative story: a recipe listing a sequence of random events that creates the dataset we are trying to explain. Figure 1.3 lists some of the key players in these stories, how they are parameterized and what samples drawn from these distributions look like. We will briefly discuss them, as we will use them to build a wide variety of topic models later.

Gaussian If you know any probability distribution already, it is (probably) the Gaussian. This distribution does not have a role in the most basic topic models that we will discuss here, but it will later (e.g., Chapter 7). We include it because it is a useful point of comparison against the other distributions we *are* using (since it is perhaps the easiest to understand and best known). A Gaussian is a distribution over all real numbers (e.g., $0.0, 0.5, -4.2, \pi, \ldots$). You can ask it to spit

Distribution	Density	Example Parameters	Example Draws
Gaussian	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mu = 2, \sigma^2 = 1.1$	$x = 2.21$
Discrete	$\prod_i \phi_i^{\mathbb{1}[w=i]}$	$\phi = \begin{bmatrix} 0.1 \\ 0.6 \\ 0.3 \end{bmatrix}$	$w = 2$
Dirichlet	$\frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^K \alpha_i\right)} \prod_{i=1}^K \theta_i^{\alpha_i - 1}$	$\alpha = \begin{bmatrix} 1.1 \\ 0.1 \\ 0.1 \end{bmatrix}$	$\theta = \begin{bmatrix} 0.8 \\ 0.15 \\ 0.05 \end{bmatrix}$

Figure 1.3: Examples of probability distributions used in the generative stories of topic models. In the case of the discrete draw, $w = 2$ denotes that the second element (the one with probability 0.6) was drawn.

out a number, and it will give you some real number between negative infinity and positive infinity. But not all numbers have equal probability. Gaussian distributions are parameterized by a mean μ and variance σ^2. Most samples from the distribution will be near the mean μ; how close is determined by the variance: higher variances will cause the samples to be more spread out.

Discrete While Gaussian distributions are over a continuous space, documents are combinations of discrete symbols, usually word tokens.[2] Thus, we need a distribution over discrete sets.

A useful metaphor for thinking about discrete distributions is a weighted die. The number of faces on the die is its dimension, and each face is associated with a distinct outcome. Each face has its own probability of how likely that outcome is; these probabilities are the parameters of a discrete distribution (Figure 1.3).

Topic models are described by discrete distributions (sometimes called multinomial distributions) that describe the connection between words and topics (the first half) and topics and documents (the second half). A distribution over words is called a topic distribution; each of

[2] An emerging trend in natural language processing research is to view words as embedded in a continuous space. We discuss these "representation learning" approaches and their connection to topic modeling in Chapter 10, but even then models are still defined over a discrete set of words.

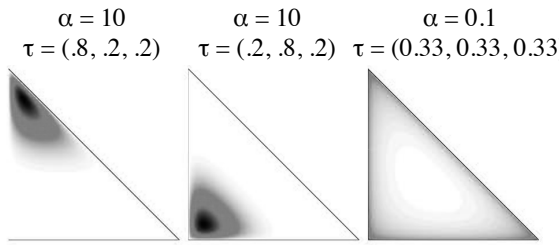

Figure 1.4: Given different Dirichlet parameters, the Dirichlet distribution can either be informative (left, middle) or sparse (right). Sparse distributions encourage distributions to favor a few elements but do not care which ones. This is consistent with our intuitions of how documents are written: they are only about a few things, and topics contain only a handful of words.

the topics gives higher weights to some words more than others (e.g., in Topic 9 from the Enron corpus, "state" and "california" have higher probability than other words). Each document also has an "allocation" for each topic: documents are about a small handful of topics, and most documents have very low weights for most of the possible topics.

Dirichlet Although discrete distributions are the star players in topic models, they are not the end of the story. We often begin with Dirichlet distributions. Just as Gaussians produce real numbers and discrete distributions produce symbols from a finite set, Dirichlet distributions produce probability vectors that can be used as the parameters of discrete distributions. Like the Gaussian distribution, they have parameters analogous to a mean and variance. The mean is called the "base measure" τ and is the expected value of the Dirichlet distribution: the values you would get if you averaged many draws from the Dirichlet. The concentration parameter α_0 controls how far away individual draws

are from the base measure. We often combine these parameters into a single value for each dimension: $\alpha_k = \alpha_0 \tau_k$.

If α_0 is very large, then the draws from a Dirichlet will be very close to τ (Figure 1.4, left). If α_0 is small, however, the discrete distributions become sparse (Figure 1.4, right). A sparse distribution is a distribution where only a few values have high probability and all other values are small.

Because topic models are meant to reflect the properties of real documents, modeling sparsity is important. When a person sits down to write a document, they only write about a handful of the topics that they could potentially use. They do not write about every possible topic, and the sparsity of Dirichlet distributions is the probabilistic tool that encodes this intuition.

There are several important special cases of the Dirichlet distribution. If the base measure τ is the same for every dimension, we call the resulting distribution *symmetric*. This case is appropriate when we do not expect any one element to be, on average, more likely than any other element across all samples from the distribution. In the symmetric case the distribution has only one parameter, the concentration α_0. If the base measure is uniform and the concentration parameter α_0 is equal to the number of dimensions K (or, equivalently, $\alpha_k = 1.0$ for all k), the distribution is uniform, placing equal probability on all K-dimensional probability distributions.

1.4 Latent Dirichlet Allocation

We now have all the tools we need to tell the complete story of the most popular topic model: latent Dirichlet allocation [Blei et al., 2003, LDA]. Latent Dirichlet allocation[3] posits a "generative process" about how the data came to be. We assemble the probabilistic pieces to tell this

[3]The name LDA is a play on LSA, its non-probabilistic forerunner (latent semantic analysis). Latent because we use probabilistic inference to infer missing probabilistic pieces of the generative story. Dirichlet because of the Dirichlet parameters encoding sparsity. Allocation because the Dirichlet distribution encodes the prior for each document's allocation over topics.

story about generating topics and how those topics are used to create diverse documents.

Generating Topics The first part of the story is to create the topics. The user specifies that there are K distinct topics. Each of the K topics is drawn from a Dirichlet distribution with a uniform base distribution and concentration parameter λ: $\phi_k \sim \text{Dir}(\lambda \boldsymbol{u})$. The discrete distribution ϕ_k has a weight for *every* word in the vocabulary.

However, when we summarize topics (as in Figure 1.1), we typically only use the top (most probable) words of a topic. The lower probability words are less relevant to the topic and thus are not shown.

Document Allocations Document allocations are distributions over topics for each document. This encodes what a document is about; the sparsity of the Dirichlet distribution's concentration parameter α_0 ensures that the document will only be about a few topics. Each document has a discrete distribution over topic: $\theta_d \sim \text{Dir}(\alpha \boldsymbol{u})$.

Words in Context Now that we know what each document is about, we create the words that appear in the document. We assume[4] that there are N_d words in document d. For each word n in the document d, we first choose a **topic assignment** $z_{d,n} \sim \text{Discrete}(\theta_d)$. This is one of the K topics that tells us which topic the word token is from, but not what the word is.

To select which word we will see in the document, we draw from a discrete distribution again. Given a word token's topic assignment $z_{d,n}$, we draw from that topic to select the word: $w_{d,n} \sim \phi_{z_{d,n}}$. The topic assignment tells you what the word is about, and then this selects which distribution over words we use to generate the word.

For example, consider the document in Figure 1.1. To generate it, we choose a distribution over all of the topics. This is θ. For this document, the distribution favors Topic 9 about California. The value for this topic

[4]We can model this in the generative story as well, e.g., with a Poisson distribution. However, we often do not care about document *lengths*—only what the document is about—so we can usually ignore this part of the story.

is higher than any other topic. For each word in the document, the generative process chooses a topic assignment z_n. For this document, any topic is theoretically possible, but we expect that most of those will be Topic 9.

Then, for each token in the document, we need to choose which word type will appear. This comes from Topic 9's distribution over words (multiple topics have word distributions shown in Figure 1.1). Each is a discrete draw from the topic's word distribution, which makes words like "California", "state", and "Sacramento" more likely.

It goes without saying that the generative story is a fiction [Box and Draper, 1987]. Nobody is sitting down with dice to decide what to type in on their keyboard. We use this story because it is *useful*. This fanciful story about randomly choosing a topic for each word can help us because if we assume this generative process, we can work backwards to find the topics that explain how a document collection was created: every word, every document, gets associated with these underlying topics.

This simple model helps us order our document collection: by assuming this story, we can discover *topics* (which certainly do not exist) so we can understand the common themes that people use to write documents. As we will see in later chapters, slight tweaks of this generative story allow us to uncover more complicated structures: how authors prefer specific topics, how topics change, or how topics can be used across languages.

1.5 Inference

Given a generative model and some data, the process of uncovering the hidden pieces of the probabilistic generative story is called *inference*. More concretely, it is a recipe for generating algorithms to go from data to *topics that explain a dataset*.

There are many flavors of algorithms for posterior inference: message passing [Zeng et al., 2013], variational inference [Blei et al., 2003], gradient descent [Hoffman et al., 2010], and Gibbs sampling [Griffiths and Steyvers, 2004]. All of these algorithms have their advocates and

reasons you should use them. In this survey, we focus on Gibbs sampling, which is simple, intuitive, and—with some clever tricks specific to topic models—fast [Yao et al., 2009]. (We discuss variational inference in Chapter 9.)

We present the results of Gibbs sampling without derivation, which—along with the history of its origin in statistical physics—are well described elsewhere.[5] We use a variety of Gibbs sampling called *collapsed* Gibbs sampling, which allows inference to side-step some of the pieces of the generative story: instead of explicitly representing the parameters of a discrete distribution, distinct from any observations drawn from that distribution, we represent the distribution solely through those observations. We can then recreate the topic and document distributions through simple formulas.

1.5.1 Random Variables

Topic Assignments Since every individual token is assumed to be generated from a single topic, we can consider the *topic assignment* of a token as a variable. For example, an instance of the word "compilation" might be in a Computer topic in one document and in an Arts topic in another document. Because each token has its own topic assignment, the same word might be assigned to *different* topics in the *same* document. To estimate *global* properties of the topic model we use aggregate statistics derived from token-level topic assignments.

Document Allocation The document allocation is a distribution over the topics for each document; in other words, it says how popular each topic is in a document. If we count up how often a document uses a topic, this gives us its popularity. We define $N_{d,i}$ as the number of times document d uses topic i. This is larger for more popular topics; however, it is not a probability because it is larger than one. We make it a probability by dividing by the number of words in a document

$$\frac{N_{d,i}}{\sum_k N_{d,k}}, \tag{1.1}$$

[5] We recommend Resnik and Hardisty [2009] for additional information on derivation.

but this is problematic because it can sometimes give us zero and ignores the influence of the Dirichlet distribution; a better estimate is[6]

$$\theta_{d,i} \approx \frac{N_{d,i} + \alpha_i}{\sum_k N_{d,k} + \alpha_k}. \tag{1.2}$$

This must never become zero because we do not want it to rule out the possibility that a topic is used in a particular document (hence, each α must be non-zero). This helps the sampler explore more of the possible combinations.

Topics Each topic is a distribution over words. To understand what a topic is about, we look at the profile of all of the tokens that have been assigned to that topic. We estimate the probability of a word in a topic as

$$\phi_{i,v} \approx \frac{V_{i,v} + \beta_v}{\sum_w V_{i,w} + \beta_w}, \tag{1.3}$$

where β is the Dirichlet parameter for the topic distribution.

1.5.2 Algorithm

The collapsed Gibbs sampling algorithm for learning a topic model is only based on the topic assignments, but we will use our estimates for the topics ϕ_k and the documents θ_d discussed above. We begin by setting topic assignments randomly: if we have K topics, each word has equal chance to be associated with any of the topics. These topics will be quite bad, looking like noisy copies of the overall corpus distribution. But we will improve them one word at a time.

The algorithm proceeds by sweeping over all word tokens in turn over and over. At each iteration we change the topic assignments for each word in a way the reflects the underlying probabilistic model of the data. On average, each pass over the data makes the topics slightly better until the model reaches a steady state. There is no easy way to tell when such a steady state has been reached, but eventually the topics will "converge" to reasonable themes and you can consider yourself done.

[6] To be technical, Equation 1.1 is a maximum likelihood estimate and Equation 1.2 is the maximum *a posteriori*, which incorporates the influence of both the prior and the data.

The equation for the probability of assigning a word to a particular topic combines information about words and about documents[7]

$$p(z_{d,n} = i \mid \dots) = \theta_d \phi_j i = \left(\frac{N_{d,i} + \alpha_i}{\sum_k N_{d,k} + \alpha_k} \right) \left(\frac{V_{i,w_{d,n}} + \beta_v}{\sum_w V_{i,w} + \beta_w} \right). \quad (1.4)$$

Computing this value for each topic will result in a probability distribution over the topic assignment for this word token, given all the other topic assignments. The next step is to randomly choose one of those indices with probability proportional to the vector value. You now assign that word to the topic, update $N_{d,\cdot}$ and $V_{\cdot,w_{d,n}}$, and move on to the next word and repeat. The two terms provide two "pressures", for global and local coherence. Sparsity in the topic-word distributions encourages tokens of the same word type to be assigned to a small number of topics, regardless of where they occur. Sparsity in the document-topic distributions encourages tokens in the same document to be assigned to a small number of topics, regardless of what type they are. For example, knowing that a word is "compilation" narrows down the number of potential topics considerably, but leaves ambiguity: is it *program* compilation or a *music* compilation? Knowing that the word occurs in a document with many other words in the Arts topic resolves this ambiguity, leaving the Arts topic as the most probable assignment.

At the very end of the algorithm, we can use the estimates of each topic (Equation 1.3) to summarize the main themes of the corpus and the estimates of each document's topic distribution (Equation 1.2) to start exploring the collection automatically (Chapter 2) or with a human in the loop (Chapter 3).

The algorithm that we have sketched here is the foundation of many of the more advanced models that we will discuss later in the survey. While we will not describe the algorithms in detail, we will occasionally reference this sketch to highlight challenges or difficulties in implementing topic models.

[7] To be theoretically correct, it is important not to include the count associated with the token you are sampling in these counts, which becomes more clear if the probability is written as $p(z_{d,n} = j \mid z_{d,1} \dots z_{d,n-1}, z_{d,n+1} \dots z_{d,N_d}, w_{d,n})$ to show the dependence on the topic assignments of *all other* tokens but not this token.

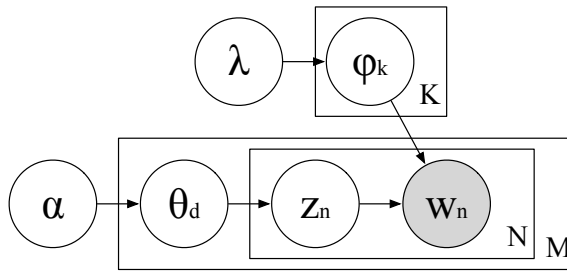

Figure 1.5: Plate diagram for LDA. Nodes show random variables, lines show (possible) probabilistic dependence, rectangles show repetition, and shading shows observation.

1.5.3 Plate Diagrams

Plate diagrams provide a shorthand for quickly explaining which random variables are associated with each other. If you look up many of the references used in this survey, you will likely see plate diagrams (we also use a plate diagram later in Figure 2.1b).

Let's begin with a plate diagram for LDA (Figure 1.5). You can compare these to the generative story in Chapter 1.4. All of the random variables are there, each in its own circle. The lines between random variables tell more of the story. You can see that if a random variable is conditioned on another, there is a line going from the variable that is *conditioned on* **to** the variable that is *conditionally dependent*. For example, a word depends on the token assignment $z_{d,n}$ and a topic ϕ_k, so we draw lines from both.

You can think about the rectangular boxes as repetition. The letter in the bottom right of the box shows how often what is inside the box is replicated. There is a box for each document (there are M in total) and each token (the box of words is inside the box for documents).

When a variable is shaded, this means that it is observed. These are the data we start with. The unshaded variables must either be inferred (e.g., topics ϕ) or are hyperparameters that must be set or inferred (e.g., Dirichlet parameter α).

Plate diagrams allow a reader to quickly see a "family resemblance" between related models, and once someone has become fully immersed in topic models, it is often possible to at a glance understand a model from its plate diagram. However, plate diagrams are imperfect; they lack some of the key information you need to understand the model. For instance, the exact probabilistic relationship between variables is underspecified.

1.5.4 What is so Great about Dirichlet?

Now that we have described what LDA is, we can return to its history. What is the innovation that separates LDA from pLSA, its predecessor? Naïvely, the difference is changing an "s" to a "d" (i.e., changing pL**S**A to L**D**A). The deeper story is about as consequential.

Instead of having a Dirichlet prior over θ, pLSA assumes that θ is a discrete parameter. In practice, this means that documents are not encouraged to focus on a limited number of topics and often "spread out" to have small weights for many different topics. In theory, this means that there is not as sound a generative story for how a document came to be: you cannot run the generative process forward from scratch if you must have θ as a parameter to start with.

These differences are relatively minor. LDA has slightly easier inference—particularly when it comes to tweaking the model—which has caused it to become the more popular of the two models. Thus, we will focus on comparing models to LDA. This is not to diminish from pLSA and its unquestionable place in the literature, but it helps us present a more unified narrative for our reader.

1.5.5 Implementations

Hopefully the previous algorithm sketch has convinced you that implementing topic models is not a Herculean task; most skilled programmers can complete a reasonable implementation of topic models in less than a day. However, we would suggest not trying to implement basic LDA if you just want the output of a topic model, many solid implementations can help users get to useful results more quickly, particularly as topic models often require extensive preprocessing.

Mallet is fast and is a widely used implementation in Java [McCallum, 2002]. This is where you should probably start, in our biased opinion. It runs in Java, uses highly-optimized Gibbs sampling implementations, and can work from a variety of text inputs. It is well documented, mature, and runs well on a multi-core machine, allowing it to process up to millions of documents. Variational inference is the other major option [Blei et al., 2003, Langford et al., 2007], but often requires a little more effort for new users to get a first result.

However, not all users are comfortable with Java; many implementations are available on other platforms and in many programming languages.[8] Many of these implementations are well-built, but check whether they have all of the features of mature implementations like Mallet so that you know what (if anything) you're missing.

However, if your corpus is truly large, consider techniques that can be parallelized over large computer clusters. These techniques can be based on variational inference [Narayanamurthy, 2011, Zhai et al., 2012] or on sampling [Newman et al., 2008].

While these implementations allow you to run *specific* topic models, other frameworks allow you to specify arbitrary generative models. This enables quick prototyping of topic models and integrating topic models with other probabilistic frameworks like regression or collaborative filtering. Examples of these general frameworks include Stan [Stan Development Team, 2014], Theano [Theano Development Team, 2016], and Infer.net [Minka et al., 2014].

If you cannot find the specific model that you want among these existing software packages, the flexibility and simplicity of topic models and inference makes it relatively simple to adapt topic models to model specific phenomena (as we describe in following chapters).

1.6 The Rest of this Survey

In each of the following chapters, we focus on an application of topic models, gradually increasing the complexity of the underlying models.

[8] So many that change so quickly; thus, we are reluctant endorse specific ones here.

The chapters do occasionally refer to each other, but a reader should be able to read each of the chapters independently.

The next chapter returns to the distinction between high level overviews and finding a needle in a haystack. We show how a high level overview can help users and algorithms find documents of interest. We show how a high level overview can help algorithms (Chapter 2) and users (Chapter 3) find documents of interest.

These tools help enable new applications of topic models: how understanding newspapers (Chapter 4) reveals the march of history, how the corpus of writers of fiction (Chapter 6) illuminates societal norms, how the writings of science reveal innovation (Chapter 5), or how politicians' speeches (Chapter 7) reveal schisms in political organizations.

Finally, the survey closes with thoughts about how interested researchers can start building their own topic models (Chapter 9) and how topic models may change in the future (Chapter 10).

References

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. Representing topics labels for exploring digital libraries. In *Proceedings of the IEEE/ACM Joint Conference on Digital Libraries*, pages 239–248, 2014.

Mark Algee-Hewitt, Ryan Heuser, and Franco Moretti. On paragraphs. scale, themes, and narrative form. *Stanford Literary Lab Pamphlets*, 1(10), October 2015.

James Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*. Springer, 2002.

Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *International Conference on Data Mining*, 2008.

Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi kai Liu. A spectral algorithm for latent Dirichlet allocation. In *Proceedings of Advances in Neural Information Processing Systems*, 2012.

David Andrzejewski and David Buttler. Latent topic feedback for information retrieval. In *Knowledge Discovery and Data Mining*, 2011.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the International Conference of Machine Learning*, 2009.

Sanjeev Arora, Rong Ge, Yoni Halpern, David M. Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the International Conference of Machine Learning*, 2013.

Anton Bakalov, Andrew Kachites McCallum, Hanna Wallach, and David Mimno. Topic models for taxonomies. In *Joint Conference on Digital Libraries*, 2012.

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Samuel Gershman. Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the Association for Computational Linguistics*, 2016.

Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 2017.

Jerome R. Bellegarda. A latent semantic analysis framework for large-span language modeling. In *European Conference on Speech Communication and Technology*, 1997.

Jerome R. Bellegarda. Statistical language model adaptation: review and perspectives. volume 42, 2004.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

Adam Berger and John Lafferty. Information retrieval as statistical translation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

Indrajit Bhattacharya. Collective entity resolution in relational data. *PhD Dissertation, University of Maryland, College Park*, 2006.

David M. Blei. Topic modeling and digital humanities. *Journal of Digital Humanities*, 2(1), 2012.

David M. Blei and John Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the International Conference of Machine Learning*, 2006.

David M. Blei and Jon D. McAuliffe. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*, 2007.

David M. Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):7:1–7:30, February 2010.

Shannon Bowen. Pseudo-events pay dividends from Cleopatra to Chipotle. *Public Relations Week*, 2016.

George E.P. Box and Norman R. Draper. *Empirical model-building and response surfaces*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1987.

Jordan Boyd-Graber and David M. Blei. Multilingual topic models for unaligned text. In *Proceedings of Uncertainty in Artificial Intelligence*, 2009.

Jordan Boyd-Graber and Philip Resnik. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of Empirical Methods in Natural Language Processing*, 2010.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. In *Proceedings of Empirical Methods in Natural Language Processing*, 2007.

Jordan Boyd-Graber, David Mimno, and David Newman. *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*. CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, Florida, 2014.

Percy Williams Bridgman. *The logic of modern physics*. Macmillan, New York, 1927.

Samuel Brody and Mirella Lapata. Bayesian word sense induction. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, 2009.

Andre David Broniatowski, Mark Dredze, J. Michael Paul, and Andrea Dugas. Using social media to perform local influenza surveillance in an inner-city hospital: A retrospective observational study. *JMIR Public Health and Surveillance*, 1(1):e5, May 2015.

John Burrows. Delta: a measure of stylistic difference and a guide to likely authorship. *Lit Linguist Computing*, 17(3):267–287, 2002.

Jaime G Carbonell, Yiming Yang, Robert E Frederking, Ralf D Brown, Yibing Geng, and Danny Lee. Translingual information retrieval: A comparative evaluation. In *International Joint Conference on Artificial Intelligence*, 1997.

Mark James Carman, Fabio Crestani, Morgan Harvey, and Mark Baillie. Towards query log based personalization using topic models. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2010.

Youngchul Cha and Junghoo Cho. Social-network analysis using topic models. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012.

Allison Chaney and David M. Blei. Visualizing topic models. In *International AAAI Conference on Weblogs and Social Media*, 2012.

Jonathan Chang and David M. Blei. Relational topic models for document networks. In *Proceedings of Artificial Intelligence and Statistics*, 2009.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems*, 2009.

Boxing Chen, George Foster, and Roland Kuhn. Adaptation of reordering models for statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, 2013.

Stanley Chen, Douglas Beeferman, and Ronald Rosenfeld. Evaluation metrics for language models. Technical report, Carnegie Mellon University School of Computer Science, 1998.

David Chiang, Steve DeNeefe, and Michael Pust. Two easy improvements to lexical weighting. In *Proceedings of the Human Language Technology Conference*, 2011.

Jaegul Choo, Changhyun Lee, Chandan K. Reddy, and Haesun Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001, 2013.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*, 2012.

Jason Chuang, Margaret E. Roberts, Brandon M. Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. TopicCheck: Interactive alignment for assessing topic model stability. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2015.

Philip R. Clarkson and Anthony J. Robinson. Language model adaptation using mixtures and an exponentially decaying cache. In *International Conference on Acoustics, Speech, and Signal Processing*, 1997.

Noah Coccaro and Daniel Jurafsky. Towards better integration of semantic predictors in statistical language modeling. In *International Conference on Acoustics, Speech, and Signal Processing*, 1998.

Ronan Collobert, Koray Kavukcuoglu, and Clement Farabet. Torch7: A Matlab-like environment for machine learning. In *NIPS Workshop on Big Learning (Biglearn)*, 2011.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

W. Bruce Croft and John Lafferty. Language modeling for information retrieval. In *Kluwer International Series on Information Retrieval*, 2003.

Van Dang and W. Bruce Croft. Term level search result diversification. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013.

Hal Daumé III. Markov random topic fields. In *Proceedings of Artificial Intelligence and Statistics*, 2009.

Wim De Smet and Marie-Francine Moens. Cross-language linking of news stories on the web using interlingual topic modelling. In *Workshop on Social Web Search and Mining*, 2009.

Scott Deerwester, Susan Dumais, Thomas Landauer, George Furnas, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, 2014.

Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the International Conference of Machine Learning*, 2007.

Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of World Wide Web Conference*, 2007.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. Topic models for dynamic translation model adaptation. In *Proceedings of the Association for Computational Linguistics*, 2012.

Jacob Eisenstein. *Written dialect variation in online social media.* Wiley, 2017.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods in Natural Language Processing*, 2010.

Jacob Eisenstein, Duen Horng Chau, Aniket Kittur, and Eric Xing. TopicViz: interactive topic exploration in document collections. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*, 2012.

Jacob Eisenstein, Iris Sun, and Lauren F. Klein. Exploratory text analysis for large document archives. In *Digital Humanities*, 2014.

Matt Erlin. Topic modeling, epistemology, and the English and German novel. *Cultural Analytics*, May 2017.

George Foster and Roland Kuhn. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007.

Jianfeng Gao, Kristina Toutanova, and Wen tau Yih. Clickthrough-based latent semantic models for web search. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011.

Jianfeng Gao, Shasha Xie, Xiaodong He, and Alnur Ali. Learning lexicon models from search logs for query expansion. In *Proceedings of Empirical Methods in Natural Language Processing*, 2012.

Matthew Gardner, Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization*, 2010.

Sean Gerrish and David M. Blei. A language-based approach to measuring scholarly impact. In *Proceedings of the International Conference of Machine Learning*, 2010.

Daniel Gildea and Thomas Hofmann. Topic-based language models using EM. In *European Conference on Speech Communication and Technology*, 1999.

Barney G. Glaser and Anslem Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Aldine, 1967.

Andrew Goldstone and Ted Underwood. The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, 45(3), Summer 2014.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235, 2004.

Justin Grimmer. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*, 18(1):1–35, 2010.

Amit Gruber, Michael Rosen-Zvi, and Yair Weiss. Hidden topic Markov models. In *Artificial Intelligence and Statistics*, 2007.

Eric Hardisty, Jordan Boyd-Graber, and Philip Resnik. Modeling perspective using adaptor grammars. In *Proceedings of Empirical Methods in Natural Language Processing*, 2010.

Jacob Harris. Word clouds considered harmful. `http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/`, 2011.

Morgan Harvey, Fabio Crestani, and Mark James Carman. Building user profiles from topic models for personalised search. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2013.

Eva Hasler, Barry Haddow, and Philipp Koehn. Sparse lexicalised features and topic adaptation for SMT. In *Proceedings of International Workshop on Spoken Language Translation*, 2012.

Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: How can citations help? In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2009.

Lynette Hirschman and Rob Gaizauskas. Natural language question answering: The view from here. *Natural Language Engineering*, 7(4):275–300, December 2001.

Matthew Hoffman, David M. Blei, and Francis Bach. Online learning for latent Dirichlet allocation. In *Proceedings of Advances in Neural Information Processing Systems*, 2010.

Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence*, 1999a.

Thomas Hofmann. Probabilistic latent semantic indexing. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999b.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, 2010.

Diane Hu and Lawrence K. Saul. A probabilistic model of unsupervised learning for musical-key profiles. In *International Society for Music Information Retrieval Conference*, 2009.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine Learning Journal*, 95(3):423–469, June 2014a.

Yuening Hu, Ke Zhai, Vladimir Edelman, and Jordan Boyd-Graber. Polylingual tree-based topic models for translation domain adaptation. In *Association for Computational Linguistics*, 2014b.

Rukmini Iyer and Mari Ostendorf. Modeling long distance dependencies in language: topic mixtures versus dynamic cache models. *IEEE Transactions on Speech Audio Process*, 7:236–239, 1999.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Association for Computational Linguistics*, 2015.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *North American Association for Computational Linguistics*, 2016.

Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.

Fred Jelinek and Robert L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, 1980.

Shan Jiang, Yuening Hu, Changsung Kang, Tim Daly, Dawei Yin, and Yi Chang. Learning query and document relevance from a web-scale click graph. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016.

Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of ACM International Conference on Web Search and Data Mining*, 2011.

Matthew L. Jockers. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press, 2013.

Matthew L. Jockers and David Mimno. Significant themes in 19th century literature. *Poetics*, 41(6):750–769, December 2013.

Patrick Juola. Authorship attribution. *Foundations and Trends in information Retrieval*, 1(3):233–334, 2006.

Saurabh S. Kataria, Krishnan S. Kumar, Rajeev R. Rastogi, Prithviraj Sen, and Srinivasan H. Sengamedu. Entity disambiguation with hierarchical topic models. In *Knowledge Discovery and Data Mining*, 2011.

S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transaction on Acoustics, Speech and Signal Processing*, 1987.

Kirill Kireyev, Leysia Palen, and Kenneth Anderson. Applications of Topics Models to Analysis of Disaster-Related Twitter Data. December 2009.

Reinhard Kneser and Jochen Peters. Semantic clustering for adaptive language modeling. In *International Conference on Acoustics, Speech, and Signal Processing*, 1997.

Reinhard Kneser, Jochen Peters, and Dietrich Klakow. Language model adaptation using dynamic marginals. In *European Conference on Speech Communication and Technology*, 1997.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2003.

Thomas K Landauer and Michael L Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the UW Centre for the New Oxford English Dictionary*, 1990.

John Langford, Lihong Li, and Alex Strehl. Vowpal Wabbit, 2007.

Mark A. Largent and Julia I. Lane. STAR METRICS and the science of science policy. *Review of Policy Research*, 29(3):431–438, 2012.

Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. Best topic word selection for topic labelling. In *Proceedings of International Conference on Computational Linguistics*, 2010.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the Association for Computational Linguistics*, 2011.

Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, 2014.

Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Knowledge Discovery and Data Mining*, 2009.

Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. Reducing the sampling complexity of topic models. In *Knowledge Discovery and Data Mining*, 2014.

Percy Liang and Dan Klein. Structured Bayesian nonparametric models with variational inference (tutorial). In *Proceedings of the Association for Computational Linguistics*, 2007.

Shangsong Liang, Zhaochun Ren, and Maarten de Rijke. Personalized search result diversification via structured learning. In *Knowledge Discovery and Data Mining*, 2014.

Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2009.

Tianyi Lin, Wentao Tian, Qiaozhu Mei, and Hong Cheng. The dual-sparse topic model: Mining focused topics and focused terms in short text. In *Proceedings of World Wide Web Conference*, 2014.

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link LDA: Joint models of topic and author community. In *Proceedings of the International Conference of Machine Learning*, 2009.

Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of World Wide Web Conference*, 2008.

Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, 14(2):178–203, 2011.

Jeff Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Association for Computational Linguistics*, 2017.

David J. C. Mackay and Linda C. Bauman Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1:1–19, 1995.

Gideon Mann, David Mimno, and Andrew Kachites McCallum. Bibliometric impact measures leveraging topic analysis. In *Joint Conference on Digital Libraries*, 2006.

Xian-Ling Mao, Zhao-Yan Ming, Zheng-Jun Zha, Tat-Seng Chua, Hongfei Yan, and Xiaoming Li. Automatic labeling hierarchical topics. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2012.

Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Girish Maskeri, Santonu Sarkar, and Kenneth Heafield. Mining business topics in source code using latent dirichlet allocation. In *India Software Engineering Conference*, 2008.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. Discriminative corpus weight estimation for machine translation. In *Proceedings of Empirical Methods in Natural Language Processing*, 2009.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit, 2002. http://www.cs.umass.edu/ mccallum/mallet.

Andrew Kachites McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272, October 2007.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013.

Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Knowledge Discovery and Data Mining*, 2005.

Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of World Wide Web Conference*, 2006.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of World Wide Web Conference*, 2007a.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Knowledge Discovery and Data Mining*, 2007b.

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of World Wide Web Conference*, 2008.

Massimo Melucci. Contextual search: A computational framework. *Foundations and Trends in Information Retrieval*, 6:257–405, 2012.

Alessandro Micarelli, Fabio Gasparetti, Filippo Sciarrone, and Susan Gauch. Personalized search on the world wide web. In *The Adaptive Web*, volume 4321, 2007.

Ian Matthew Miller. Rebellion, crime and violence in qing china, 17221911: A topic modeling approach. *Poetics*, 41(6):626–649, December 2013.

David Mimno. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage*, 5(1):3:1–3:19, April 2012.

David Mimno and David M. Blei. Bayesian checking for topic models. In *Proceedings of Empirical Methods in Natural Language Processing*, 2011.

David Mimno and Andrew Kachites McCallum. Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. In *Proceedings of the 2008 Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew Kachites McCallum. Polylingual topic models. In *Proceedings of Empirical Methods in Natural Language Processing*, 2009.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew Kachites McCallum. Optimizing semantic coherence in topic models. In *Proceedings of Empirical Methods in Natural Language Processing*, 2011.

David Mimno, Matthew Hoffman, and David M. Blei. Sparse stochastic inference for latent Dirichlet allocation. In *Proceedings of the International Conference of Machine Learning*, 2012.

Tom Minka, John Winn, John Guiver, and David Knowles. Infer.NET 2.6, 2014. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

Franco Moretti. The slaughterhouse of literature. *Modern Language Quarterly*, 61(1):207–227, 2000.

Franco Moretti. *Distant Reading*. Verso, 2013a. URL `https://books.google.com/books?id=YKMCy9I3PG4C`.

Franco Moretti. Operationalizing, or the function of measurement in literary theory. *New Left Review*, 84, Nov/Dec 2013b.

Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, Mass., 1964.

Christof Müller and Iryna Gurevych. A study on the semantic relatedness of query and document terms in information retrieval. In *Proceedings of Empirical Methods in Natural Language Processing*, 2009.

Ramesh Nallapati and William Cohen. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *International Conference on Weblogs and Social Media*, 2008.

Shravan Narayanamurthy. Yahoo! LDA, 2011. URL `https://github.com/shravanmn/Yahoo_LDA/wiki`.

Radford M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto, 1993.

David Newman and Sharon Block. Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology*, 18(1):753–767, 2006.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed Inference for Latent Dirichlet Allocation. In *Proceedings of Advances in Neural Information Processing Systems*, 2008.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

Thang Nguyen, Jordan Boyd-Graber, Jeff Lund, Kevin Seppi, and Eric Ringger. Is your anchor going up or down? Fast and accurate supervised topic models. In *North American Association for Computational Linguistics*, 2015a.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. Lexical and hierarchical topic regression. In *Proceedings of Advances in Neural Information Processing Systems*, 2013.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Jonathan Chang. Learning a concept hierarchy from multi-labeled documents. In *Neural Information Processing Systems*, 2014.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. Tea party in the house: A hierarchical ideal point topic model and its application to Republican legislators in the 112th Congress. In *Association for Computational Linguistics*, 2015b.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. Mining multilingual topics from Wikipedia. In *Proceedings of World Wide Web Conference*, 2009.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Stanford Digital Library Working Paper SIDL-WP-1999-0120, Stanford University, 1999.

Bo Pang and Lillian Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc, 2008.

Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217 – 235, 2000.

Laurence A. Park and Kotagiri Ramamohanarao. The sensitivity of latent dirichlet allocation for information retrieval. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, 2009.

Michael Paul and Roxana Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2010.

James Pitkow, Hinrich Schütze, Todd Cass, Rob Cooley, Don Turnbull, Andy Edmonds, Eytan Adar, and Thomas Breuel. Personalized search. *Communications of the ACM*, 45(9):50–55, September 2002.

Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of Empirical Methods in Natural Language Processing*, 2009.

Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. Characterizing microblogs with topic models. In *International Conference on Weblogs and Social Media*, 2010a.

Daniel Ramage, Christopher D. Manning, and Daniel A. Mcfarland. Which universities lead and lag? toward university rankings based on scholarly output. In *NIPS Workshop on Computational Social Science and the Wisdom of the Crowds*, 2010b.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David M. Blei. Deep exponential families. In *Proceedings of Artificial Intelligence and Statistics*, 2015.

Philip Resnik and Eric Hardisty. Gibbs sampling for the uninitiated. Technical report, University of Maryland, 2009. URL `http://www.umiacs.umd.edu/~resnik/pubs/gibbs.pdf`.

Lia M. Rhody. Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1), 2012.

Allen Beye Riddell. *How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models*, pages 91–113. Camden House, 2012.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. STM: R package for structural topic models, 2014. URL `http://www.structuraltopicmodel.com`. R package version 1.0.8.

Joseph John Rocchio. *Relevance feedback in information retrieval*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall, 1971.

Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of Uncertainty in Artificial Intelligence*, 2004.

Nick Ruiz and Marcello Federico. Topic adaptation for lecture translation through bilingual latent semantic models. In *WMT Workshop on Statistical Machine Translation*, 2011.

Gerard. Salton. *Automatic Information Organization and Retrieval.* McGraw Hill Text, 1968.

Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Search result diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90, March 2015.

Tobias Schnabel, Igor Labutov, David M. Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of Empirical Methods in Natural Language Processing*, 2015.

Kristie Seymore and Ronald Rosenfeld. Using story topics for language model adaptation. In *European Conference on Speech Communication and Technology*, 1997.

Kristie Seymore, Stanley F. Chen, and Ronald Rosenfeld. Nonlinear interpolation of topic models for language model adaptation. In *International Conference on Spoken Language Processing*, 1998.

Alison Smith, Jason Chuang, Yuening Hu, Jordan Boyd-Graber, and Leah Findlater. Concurrent visualization of relationships between words and topics in topic models. In *ACL Workshop on Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014.

Alison Smith, Sana Malik, and Ben Shneiderman. *Visual analysis of topical evolution in unstructured text: Design and evaluation of topicflow*, pages 159–175. Springer, 2015.

Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Leah Findlater, Jordan Boyd-Graber, and Niklas Elmqvist. Evaluating visual representations for topic understanding and their effects on manually generated labels. *Transactions of the Association for Computational Linguistics*, 2016.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of Empirical Methods in Natural Language Processing*, 2012.

Fei Song and W. Bruce Croft. A general language model for information retrieval. In *International Conference on Information and Knowledge Management*, 1999.

Wei Song, Yu Zhang, Ting Liu, and Sheng Li. Bridging topic modeling and personalized search. In *Proceedings of International Conference on Computational Linguistics*, 2010.

Stan Development Team. Stan: A C++ library for probability and sampling, version 2.5.0, 2014. URL `http://mc-stan.org/`.

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Knowledge Discovery and Data Mining*, 2004.

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. Translation model adaptation for statistical machine translation with monolingual topic information. In *Proceedings of the Association for Computational Linguistics*, 2012.

Yizhou Sun, Jiawei Han, Jing Gao, and Yintao Yu. iTopicModel: Information network-integrated topic modeling. In *International Conference on Data Mining*, 2009.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Rick Szostak. Classifying science. *Classifying Science: Phenomena, Data, Theory, Method, Practice*, pages 1–22, 2004.

Edmund M. Talley, David Newman, David Mimno, Bruce W. Herr, Hanna M. Wallach, Gully A. P. C. Burns, A. G. Miriam Leenders, and Andrew Kachites McCallum. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, 8(6):443–444, May 2011.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. Bilingual-lsa based lm adaptation for spoken language translation. In *Proceedings of the Association for Computational Linguistics*, 2007.

Jian Tang, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the International Conference of Machine Learning*, 2014.

Timothy R. Tangherlini and Peter Leonard. Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics*, 41(6): 725–749, December 2013.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Association for Computational Linguistics*, 2008.

Kristina Toutanova and Mark Johnson. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Proceedings of Advances in Neural Information Processing Systems*, 2008.

David Vallet and Pablo Castells. Personalized diversification of search results. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012.

Fernanda B. Viégas and Martin Wattenberg. Tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, 2008.

Maximilian Viermetz, Michal Skubacz, Cai-Nicolas Ziegler, and Dietmar Seipel. Tracking topic evolution in news environments. In *IEEE International Conference on E-Commerce Technology*, 2008.

Ellen M. Voorhees. Overview of TREC 2003. In *Proceedings of the Text REtrieval Conference*, pages 1–13, 2003.

Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng. Collaborative personalized Twitter search with topic-language models. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014.

Ivan Vulić and Marie-Francine Moens. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, 2012.

Ivan Vulić and Marie-Francine Moens. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *Proceedings of Empirical Methods in Natural Language Processing*, 2014.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Cross-language information retrieval with latent topic models trained on a comparable corpus. In *Asia Information Retrieval Societies*, 2011a.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the Association for Computational Linguistics*, 2011b.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, 16(3):331–368, 2013.

Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. Probabilistic topic modeling in multilingal settings: An overview of its methodology and applications. *Information Processing and Management*, 51(1), 2015.

Hanna Wallach, David Mimno, and Andrew Kachites McCallum. Rethinking LDA: Why priors matter. In *Proceedings of Advances in Neural Information Processing Systems*, 2009a.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the International Conference of Machine Learning*, 2009b.

Chong Wang, David M. Blei, and David Heckerman. Continuous time dynamic topic models. In *Proceedings of Uncertainty in Artificial Intelligence*, 2008.

Quan Wang, Jun Xu, Hang Li, and Nick Craswell. Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Transactions on Information Systems*, 31(1):5:1–5:44, January 2013.

Shiliang Wang, J. Michael Paul, and Mark Dredze. Social media as a sensor of air quality and public response in China. *Journal of Medical Internet Research*, 17(3):e22, Mar 2015.

Xing Wang, Deyi Xiong, Min Zhang, Yu Hong, and Jianmin Yao. A topic-based reordering model for statistical machine translation. *Natural Language Processing and Chinese Computing*, 496:414–421, 2014.

Xing Wei. *Topic Models in Information Retrieval*. Ph.D. dissertation, University of Massachusetts Amherst, 2007.

Xing Wei and W. Bruce Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. TwitterRank: Finding topic-sensitive influential Twitterers. In *Proceedings of ACM International Conference on Web Search and Data Mining*, 2010.

Theresa Wilson and Janyce Wiebe. Annotating attributions and private states. In *CorpusAnno '05: Proceedings of the Workshop on Frontiers in Corpus Annotations II*, 2005.

Frank Wood and Yee Whye Teh. A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 12, 2009.

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, and Shouxun Lin. A topic similarity model for hierarchical phrase-based translation. In *Proceedings of the Association for Computational Linguistics*, 2012.

Deyi Xiong and Min Zhang. A topic-based coherence model for statistical machine translation. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2013.

Deyi Xiong, Qun Liu, and Shouxun Lin. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, 2006.

Tze-I Yang, Andrew J. Torget, and Rada Mihalcea. Topic modeling on historical newspapers. In *ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 2011.

Limin Yao, David Mimno, and Andrew Kachites McCallum. Efficient methods for topic model inference on streaming document collections. In *Knowledge Discovery and Data Mining*, 2009.

Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In *Proceedings of the European Conference on Information Retrieval*, volume 5478, 2009.

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. Geographical topic discovery and comparison. In *Proceedings of World Wide Web Conference*, 2011.

Heng Yu, Jinsong Su, Yajuan Lv, and Qun Liu. A topic-triggered language model for statistical machine translation. In *International Joint Conference on Natural Language Processing*, 2013.

Jia Zeng, W. K. Cheung, and Jiming Liu. Learning topic models by belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1121–1134, 2013.

Qing T. Zeng, Doug Redd, Thomas C. Rindflesch, and Jonathan R. Nebeker. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. In *American Medical Informatics Association Annual Symposium*, 2012.

ChengXiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001a.

ChengXiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2001b.

ChengXiang Zhai, Atulya Velivelli, and Bei Yu. A cross-collection mixture model for comparative text mining. In *Knowledge Discovery and Data Mining*, 2004.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of World Wide Web Conference*, 2012.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. Cross-lingual latent topic extraction. In *Proceedings of the Association for Computational Linguistics*, 2010.

Bing Zhao and Eric P. Xing. BiTAM: Bilingual topic admixture models for word alignment. In *Proceedings of the Association for Computational Linguistics*, 2006.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of Empirical Methods in Natural Language Processing*, 2010.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing Twitter and traditional media using topic models. In *Proceedings of the European Conference on Information Retrieval*, 2011.

Ding Zhou, Xiang Ji, Hongyuan Zha, and C. Lee Giles. Topic evolution and social interactions: How authors effect research. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2006.

Jun Zhu, Amr Ahmed, and Eric P. Xing. MedLDA: maximum margin supervised topic models for regression and classification. In *Proceedings of the International Conference of Machine Learning*, 2009.

Index

k-means clustering, 51
JSTOR, 56

American Revolution, 51
approval rating, 82
aspect model, 87

Bilingual topical admixture, 111

Chinese language, 54, 96
close reading, 69
comparable corpora, 95
continuous time dynamic topic model, 65
copycat model, 66

decoding (machine translation), 100
deep learning, 117, 128
diary, 56
Dickens, 73
dictionary, 96, 98
Dirichlet distribution, 10
  alternatives, 65
  conjugacy, 116
  parameter, 11
  role in LDA, 18
  smoothing in language models, 25
  tree, 97
disambiguation, 105
discrete distribution, 9, 65, 116
distant reading, 70
distributed representation, 128
document language modeling, 23
dynamic influence model, 67
dynamic topic model, 64

expectation maximization, 118

Facebook, 84
fully-factorized distribution, 119

Gaussian distribution, 8, 64, 86, 116
German language, 54, 56, 95
Gibbs sampling, 14, 122

for LDA, 15
grounded theory, 84

hidden topic Markov model, 109

Infer.Net, 118
influenza, 82
interactive TOpic Model and MEtadata (TOME), 44
interactivity
information retrieval, 33
topic models, 46
interpretability, 45, 83, 117, 125, 128

Japanese language, 54

labeled LDA, 42
language model
machine translation, 106
query expansion, 28
latent Dirichlet allocation, 7, 11
document language model, 26
generative process, 13
implementations, 18
latent semantic analysis, 7, 51, 108
likelihood evaluation, 45, 125
link latent Dirichlet allocation, 90

Mallet, 18
Martha Ballard, 55
mixed-membership block model, 89
Modern Language Association, 58

names of fictional characters, 72
National Institutes of Health, 62
nested Dirichlet process, 88
newspaper, 50
novels, 70

online latent Dirichlet allocation, 127

PageRank, 41
parallel corpus, 98
phrase (machine translation), 102
plate diagram, 17
personalized retrieval, 35
poetry, 77
pollution in China, 82
polylingual latent Dirichlet allocation, 94
polylingual tree-based Latent Dirichlet allocation, 98
posterior predictive checks, 113
prediction vs. interpretation, 82
probabilistic latent semantic analysis, 7, 26, 51

query expansion, 27, 95

relevance model, 29
reordering (machine translation), 110

search personalization, 34
sentiment analysis, 84
smoothing, 24
spectral learning, 128
Stan, 118
statistical machine translation, 92
domain adaptation, 99
stochastic block model, 89
stylometry, 78

supervised latent Dirichlet allocation, 86, 124
survey, 69, 81
synthetic data, 123

Termite, 44
Theano, 118
topic coherence, 46
topic detection and tracking, 7
topic labeling, 40
Topic Model Visualization Engine, 43
topical guide, 43
Torch, 118
tree-based latent Dirichlet allocation, 96
Twitter, 91

upstream vs. downstream models, 86

variational inference, 118

Wikipedia, 41, 93
WordNet, 96