Introduction to Probabilistic Topic Models

David M. Blei
Princeton University

Abstract

Probabilistic topic models are a suite of algorithms whose aim is to discover the hidden thematic structure in large archives of documents. In this article, we review the main ideas of this field, survey the current state-of-the-art, and describe some promising future directions. We first describe latent Dirichlet allocation (LDA) [8], which is the simplest kind of topic model. We discuss its connections to probabilistic modeling, and describe two kinds of algorithms for topic discovery. We then survey the growing body of research that extends and applies topic models in interesting ways. These extensions have been developed by relaxing some of the statistical assumptions of LDA, incorporating meta-data into the analysis of the documents, and using similar kinds of models on a diversity of data types such as social networks, images and genetics. Finally, we give our thoughts as to some of the important unexplored directions for topic modeling. These include rigorous methods for checking models built for data exploration, new approaches to visualizing text and other high dimensional data, and moving beyond traditional information engineering applications towards using topic models for more scientific ends.

1 Introduction

As our collective knowledge continues to be digitized and stored—in the form of news, blogs, web pages, scientific articles, books, images, sound, video, and social networks—it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search and understand these vast amounts of information.

Right now, we work with online information using two main tools—search and links. We type keywords into a search engine and find a set of documents related to them. We look at the documents in that set, possibly navigating to other linked documents. This is a powerful way of interacting with our online archive, but something is missing.

Imagine searching and exploring documents based on the themes that run through them. We might "zoom in" and "zoom out" to find specific or broader themes; we might look at how those themes changed through time or how they are connected to each other. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme.

For example, consider using themes to explore the complete history of the New York Times. At a broad level some of the themes might correspond to the sections of the newspaper—foreign policy, national affairs, sports. We could zoom in on a theme of interest, such as foreign policy, to reveal various aspects of it—Chinese foreign policy, the conflict in the Middle East, the United States's relationship with Russia. We could then navigate through time to reveal how these specific themes have changed, tracking, for example, the changes in the conflict in the Middle East over the last fifty years. And, in all of this exploration, we would be pointed to the original articles relevant to the themes. The thematic structure would be a new kind of window through which to explore and digest the collection.

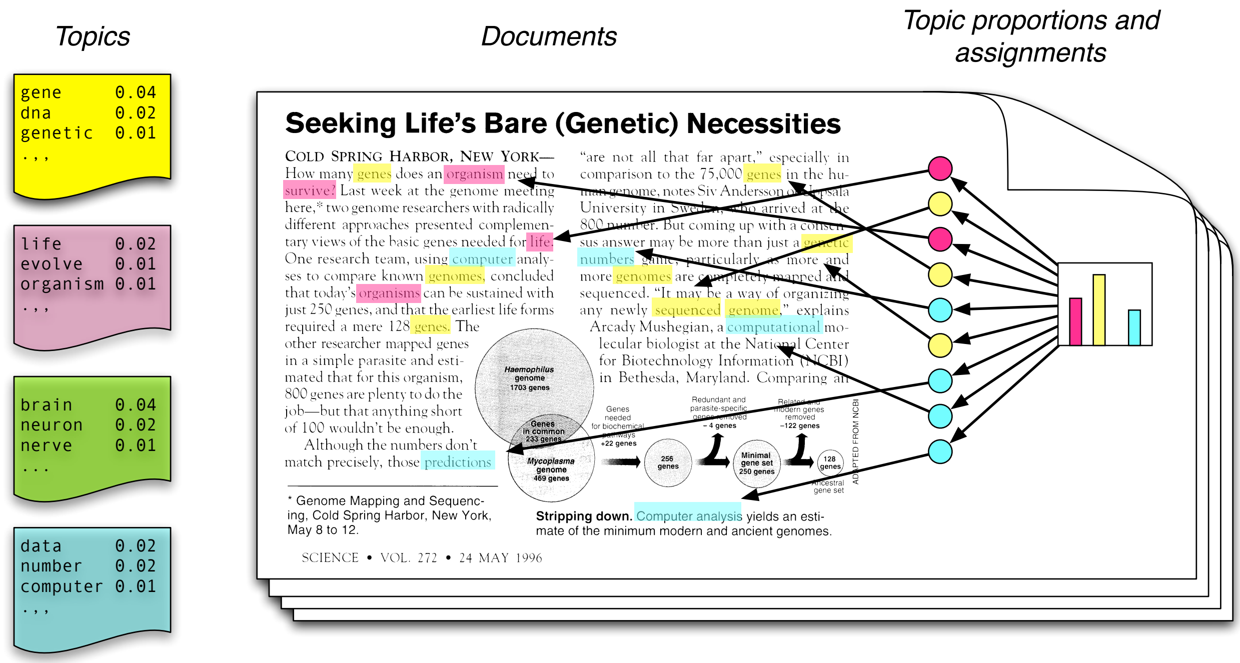
But we don't interact with electronic archives in this way. While more and more texts are available online, we simply do not have the human power to read and study them to provide the kind of browsing experience described above. To this end, machine learning researchers have developed *probabilistic topic modeling*, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time. (See, for example, Figure 3 for topics found by analyzing the *Yale Law Journal*.) Topic modeling algorithms do not require any prior annotations or labeling of the documents—the topics emerge from the analysis of the original texts. Topic modeling enables us to organize and summarize electronic archives at a scale that would be impossible by human annotation.

2 Latent Dirichlet allocation

We first describe the basic ideas behind *latent Dirichlet allocation* (LDA), which is the simplest topic model [8]. The intuition behind LDA is that documents exhibit multiple topics. For example, consider the article in Figure 1. This article, entitled "Seeking Life's Bare (Genetic) Necessities," is about using data analysis to determine the number of genes that an organism needs to survive (in an evolutionary sense).

By hand, we have highlighted different words that are used in the article. Words about *data analysis*, such as "computer" and "prediction," are highlighted in blue; words about *evolutionary biology*, such as "life" and "organism", are highlighted in pink; words about *genetics,* such as "sequenced" and "genes," are highlighted in yellow. If we took the time to highlight every word in the article, you would see that this article blends genetics, data analysis, and evolutionary biology with different proportions. (We exclude words, such as "and" "but" or "if," which contain little topical content.) Furthermore, knowing that this article blends those topics would help you situate it in a collection of scientific articles.

LDA is a statistical model of document collections that tries to capture this intuition. It is most easily described by its generative process, the imaginary random process by which the



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
.,,
life 0.02
evolve 0.01
organism 0.01
.,,
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
.,,
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Figure 1: **The intuitions behind latent Dirichlet allocation.** We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

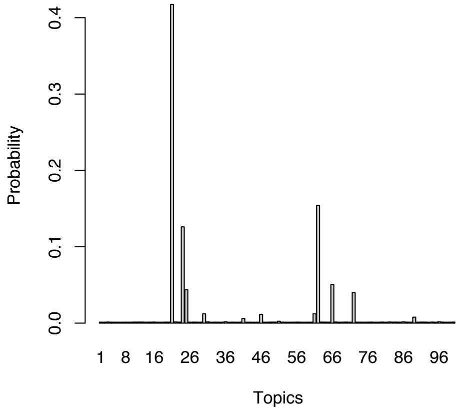
model assumes the documents arose. (The interpretation of LDA as a probabilistic model is fleshed out below in Section 2.1.)

We formally define a *topic* to be a distribution over a fixed vocabulary. For example the *genetics* topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability. We assume that these topics are specified before any data has been generated.[1] Now for each document in the collection, we generate the words in a two-stage process.

1. Randomly choose a distribution over topics.
2. For each word in the document
   (a) Randomly choose a topic from the distribution over topics in step #1.
   (b) Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits the topics with different proportion (step #1); each word in each document

[1] Technically, the model assumes that the topics are generated first, before the documents.



"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Figure 2: **Real inference with LDA.** We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left is the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.

is drawn from one of the topics (step #2b), where the selected topic is chosen from the per-document distribution over topics (step #2a).[2]

In the example article, the distribution over topics would place probability on *genetics*, *data analysis* and *evolutionary biology*, and each word is drawn from one of those three topics. Notice that the next article in the collection might be about *data analysis* and *neuroscience*; its distribution over topics would place probability on those two topics. This is the distinguishing characteristic of latent Dirichlet allocation—all the documents in the collection share the same set of topics, but each document exhibits those topics with different proportion.

As we described in the introduction, the goal of topic modeling is to automatically discover the topics from a collection of documents. The documents themselves are observed, while the topic structure—the topics, per-document topic distributions, and the per-document per-word topic assignments—are *hidden structure*. The central computational problem for topic modeling is to use the observed documents to infer the hidden topic structure. This can be thought of as "reversing" the generative process—what is the hidden structure that likely generated the observed collection?

Figure 2 illustrates example inference using the same example document from Figure 1. Here, we took 17,000 articles from *Science* magazine and used a topic modeling algorithm to infer the hidden topic structure. (The algorithm assumed that there were 100 topics.) We

[2] We should explain the mysterious name, "latent Dirichlet allocation." The distribution that is used to draw the per-document topic distributions in step #1 (the cartoon histogram in Figure 1) is called a *Dirichlet distribution*. In the generative process for LDA, the result of the Dirichlet is used to *allocate* the words of the document to different topics. Why *latent*? Keep reading.

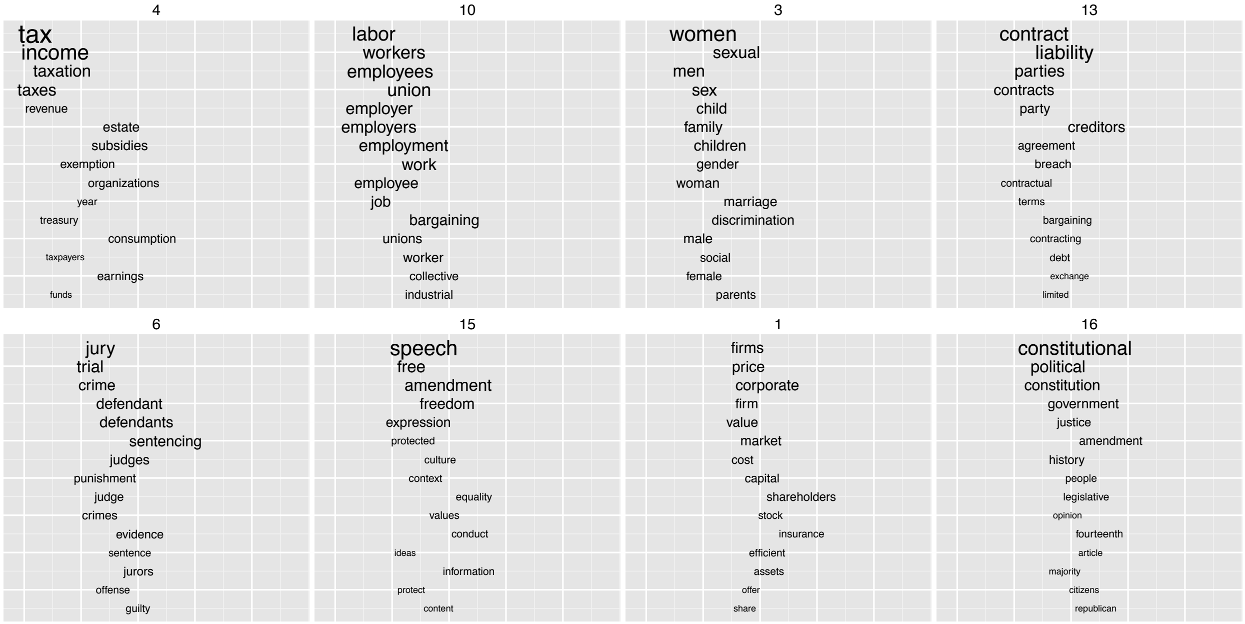


Figure 3: A topic model fit to the *Yale Law Journal*. Here there are twenty topics (the top eight are plotted). Each topic is illustrated with its top most frequent words. Each word's position along the x-axis denotes its specificity to the documents. For example "estate" in the first topic is more specific than "tax."

then computed the inferred topic distribution for the example article (Figure 2, left), the distribution over topics that best describes its particular collection of words. Notice that this topic distribution, though it can use any of the topics, has only "activated" a handful of them. Further, we can examine the most probable terms from each of the most probable topics (Figure 2, right). On examination, we see that these terms are recognizable as terms about genetics, survival, and data analysis, the topics that are combined in the example article.

We emphasize that the algorithms have no information about these subjects and the articles are not labeled with topics or keywords. The interpretable topic distributions arise by computing the hidden structure that likely generated the observed collection of documents.[3] For example, Figure 3 illustrates topics discovered from *Yale Law Journal*. (Here the number of topics was set to be twenty.) Topics about subjects like genetics and data analysis are replaced by topics about discrimination and contract law.

The utility of topic models stems from the property that the inferred hidden structure resembles the thematic structure of the collection. This interpretable hidden structure annotates each document in the collection—a task that is painstaking to perform by hand—and these annotations can be used to aid tasks like information retrieval, classification, and

[3] Indeed calling these models "topic models" is retrospective—the topics that emerge from the inference algorithm are interpretable for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA.

corpus exploration.[4] In this way, topic modeling provides an algorithmic solution to managing, organizing, and annotating large archives of texts.

2.1 LDA and probabilistic models

LDA and other topic models are part of the larger field of *probabilistic modeling*. In generative probabilistic modeling, we treat our data as arising from a generative process that includes *hidden variables*. This generative process defines a *joint probability distribution* over both the observed and hidden random variables. We perform data analysis by using that joint distribution to compute the *conditional distribution* of the hidden variables given the observed variables. This conditional distribution is also called the *posterior distribution*.

LDA falls precisely into this framework. The observed variables are the words of the documents; the hidden variables are the topic structure; and the generative process is as described above. The computational problem of inferring the hidden topic structure from the documents is the problem of computing the posterior distribution, the conditional distribution of the hidden variables given the documents.

We can describe LDA more formally with the following notation. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary (the distributions over words at left in Figure 1). The topic proportions for the dth document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d (the cartoon histogram in Figure 1). The topic assignments for the dth document are z_d, where $z_{d,n}$ is the topic assignment for the nth word in document d (the colored coin in Figure 1). Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary.

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables,

$$\begin{aligned} & p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) \\ & = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\textstyle\prod_{n=1}^{N} p(z_{d,n} \,|\, \theta_d) p(w_{d,n} \,|\, \beta_{1:K}, z_{d,n}) \right) . \end{aligned} \tag{1}$$

Notice that this distribution specifies a number of dependencies. For example, the topic assignment $z_{d,n}$ depends on the per-document topic proportions θ_d. As another example, the observed word $w_{d,n}$ depends on the topic assignment $z_{d,n}$ and *all* of the topics $\beta_{1:K}$. (Operationally, that term is defined by looking up which topic $z_{d,n}$ refers to and looking up the probability of the word $w_{d,n}$ within that topic.)

These dependencies define LDA. They are encoded in the statistical assumptions behind the generative process, in the particular mathematical form of the joint distribution, and—in a third way—in the *probabilistic graphical model* for LDA. Probabilistic graphical models provide

[4] See, for example, the browser of *Wikipedia* built with a topic model at `http://www.sccs.swarthmore.edu/users/08/ajb/tmve/wiki100k/browse/topic-list.html`.

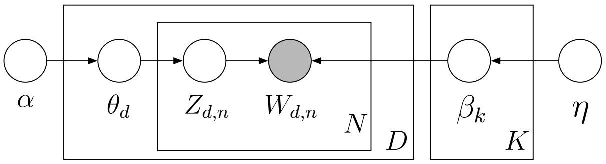


Figure 4: **The graphical model for latent Dirichlet allocation.** Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes–the topic proportions, assignments and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The N plate denotes the collection words within documents; the D plate denotes the collection of documents within the collection.

a graphical language for describing families of probability distributions.[5] The graphical model for LDA is in Figure 4. These three representations are equivalent ways of describing the probabilistic assumptions behind LDA.

In the next section, we describe the inference algorithms for LDA. However, we first pause to describe the short history of these ideas. LDA was developed to fix an issue with a previously developed probabilistic model *probabilistic latent semantic analysis* (pLSI) [21]. That model was itself a probabilistic version of the seminal work on *latent semantic analysis* [14], which revealed the utility of the singular value decomposition of the document-term matrix. From this matrix factorization perspective, LDA can also be seen as a type of principal component analysis for discrete data [11, 12].

2.2 Posterior computation for LDA

We now turn to the computational problem, computing the conditional distribution of the topic structure given the observed documents. (As we mentioned above, this is called the *posterior*.) Using our notation, the posterior is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \,|\, w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \tag{2}$$

The numerator is the joint distribution of all the random variables, which can be easily computed for any setting of the hidden variables. The denominator is the *marginal probability* of the observations, which is the probability of seeing the observed corpus under any topic model. In theory, it can be computed by summing the joint distribution over every possible instantiation of the hidden topic structure.

[5] The field of graphical models is actually more than a language for describing families of distributions. It is a field that illuminates the deep mathematical links between probabilistic independence, graph theory, and algorithms for computing with probability distributions [35].

That number of possible topic structures, however, is exponentially large; this sum is intractable to compute.[6] As for many modern probabilistic models of interest—and for much of modern Bayesian statistics—we cannot compute the posterior because of the denominator, which is known as the *evidence*. A central research goal of modern probabilistic modeling is to develop efficient methods for approximating it. Topic modeling algorithms—like the algorithms used to create Figure 1 and Figure 3—are often adaptations of general-purpose methods for approximating the posterior distribution.

Topic modeling algorithms form an approximation of Equation 2 by forming an alternative distribution over the latent topic structure that is adapted to be close to the true posterior. Topic modeling algorithms generally fall into two categories—sampling-based algorithms and variational algorithms.

Sampling based algorithms attempt to collect samples from the posterior to approximate it with an empirical distribution. The most commonly used sampling algorithm for topic modeling is *Gibbs sampling*, where we construct a *Markov chain*—a sequence of random variables, each dependent on the previous—whose limiting distribution is the posterior. The Markov chain is defined on the hidden topic variables for a particular corpus, and the algorithm is to run the chain for a long time, collect samples from the limiting distribution, and then approximate the distribution with the collected samples. (Often, just one sample is collected as an approximation of the topic structure with maximal probability.) See [33] for a good description of Gibbs sampling for LDA, and see `http://CRAN.R-project.org/package=lda` for a fast open-source implementation.

Variational methods are a deterministic alternative to sampling-based algorithms [22, 35]. Rather than approximating the posterior with samples, variational methods posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior.[7] Thus, the inference problem is transformed to an optimization problem. Variational methods open the door for innovations in optimization to have practical impact in probabilistic modeling. See [8] for a coordinate ascent variational inference algorithm for LDA; see [20] for a much faster online algorithm (and open-source software) that easily handles millions of documents and can accommodate streaming collections of text.

Loosely speaking, both types of algorithms perform a search over the topic structure. The collection of documents (the observed random variables in the model) are held fixed and serve as a guide towards where to search. Which approach is better depends on the particular topic model being used—we have so far focused on LDA, but see below for other topic models—and is a source of academic debate. For a good discussion of the merits and drawbacks of both, see [1].

[6] More technically, the sum is over all possible ways of assigning each observed word of the collection to one of the topics. Document collections usually contain observed words at least on the order of millions.

[7] Closeness is measured with *Kullback-Leibler divergence*, an information theoretic measurement of the distance between two probability distributions.

3 Research in topic modeling

The simple LDA model provides a powerful tool for discovering and exploiting the hidden thematic structure in large archives of text. However, one of the main advantages of formulating LDA as a probabilistic model is that it can easily be used as a module in more complicated models for more complicated goals. Since its introduction, LDA has been extended and adapted in many ways.

3.1 Relaxing the assumptions of LDA

LDA is defined by the statistical assumptions it makes about the corpus. One active area of topic modeling research is how to relax and extend these assumptions to uncover more sophisticated structure in the texts.

One assumption that LDA makes is the "bag of words" assumption, that the order of the words in the document does not matter. (To see this note that the joint distribution of Equation 1 remains invariant to permutation of the words of the documents.) While this assumption is unrealistic, it is reasonable if our only goal is to uncover the course semantic structure of the texts.[8] For more sophisticated goals—such as language generation—it is patently not appropriate. There have been a number of extensions to LDA that model words nonexchangeably. For example, [36] developed a topic model that relaxes the bag of words assumption by assuming that the topics generate words conditional on the previous word; [18] developed a topic model that switches between LDA and a standard HMM. These models expand the parameter space significantly, but show improved language modeling performance.

Another assumption is that the order of documents does not matter. Again, this can be seen by noticing that Equation 1 remains invariant to permutations of the ordering of documents in the collection. This assumption may be unrealistic when analyzing long-running collections that span years or centuries. In such collections we may want to assume that the *topics* change over time. One approach to this problem is the dynamic topic model [5]—a model that respects the ordering of the documents and gives a richer posterior topical structure than LDA. Figure 5 shows a topic that results from analyzing all of *Science* magazine under the dynamic topic model. Rather than a single distribution over words, a topic is now a sequence of distributions over words. We can find an underlying theme of the collection and track how it has changed over time.

A third assumption about LDA is that the number of topics is assumed known and fixed. The Bayesian nonparametric topic model [34] provides an elegant solution: The number of topics is determined by the collection during posterior inference, and furthermore new documents can exhibit previously unseen topics. Bayesian nonparametric topic models have been extended to hierarchies of topics, which find a tree of topics, moving from more general to more concrete, whose particular structure is inferred from the data [3].

[8] As a thought experiment, imagine shuffling the words of the article in Figure 1. Even when shuffled, you would be able to glean that the article has something to do with genetics.

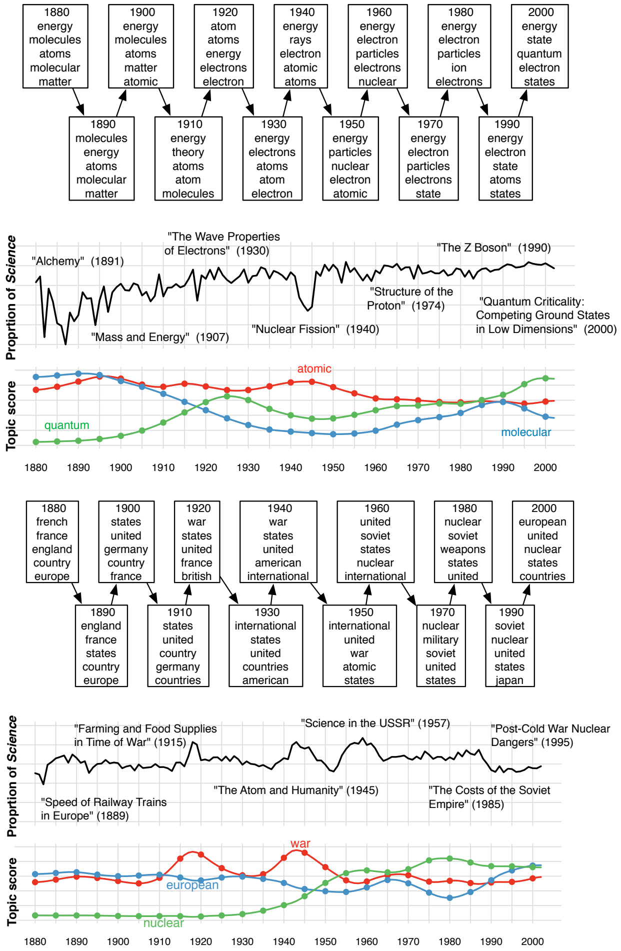


Figure 5: **Two topics from a dynamic topic model.** This model was fit to *Science* from (1880–2002). We have illustrated the top words at each decade.

There are still other extensions of LDA that relax various assumptions made by the model. The correlated topic model [6] and pachinko allocation machine [24] allow the occurrence of topics to exhibit correlation (for example a document about *geology* is more likely to also be about *chemistry* then it is to be about *sports*); the spherical topic model [28] allows words to be *unlikely* in a topic (for example, "wrench" will be particularly unlikely in a topic about *cats*); sparse topic models enforce further structure in the topic distributions [37]; and "bursty" topic models provide a more realistic model of word counts [15].

3.2 Incorporating meta-data

In many text analysis settings, the documents contain additional information—such as author, title, geographic location, links, and others—that we might want to account for when fitting a topic model. There has been a flurry of research on adapting topic models to include meta-data.

The author-topic model [29] is an early success story for this kind of research. The topic proportions are attached to authors; papers with multiple authors are assumed to attach each word to an author, drawn from a topic drawn from his or her topic proportions. The author-topic model allows for inferences about authors as well as documents. Rosen-Zvi et al. show examples of author similarity based on their topic proportions—such computations are not possible with LDA.

Many document collections are linked—for example scientific papers are linked by citation or web pages are linked by hyperlink—and several topic models have been developed to account for those links when estimating the topics. The *relational topic model* of [13] assumes that each document is modeled as in LDA and that the links between documents depend on the distance between their topic proportions. This is both a new topic model and a new network model. Unlike traditional statistical models of networks, the relational topic model takes into account node attributes (here, the words of the documents) in modeling the links.

Other work that incorporates meta-data into topic models includes models of linguistic structure [10], models that account for distances between corpora [38], and models of named entities [26]. General purpose methods for incorporating meta-data into topic models include Dirichlet-multinomial regression models [25] and supervised topic models [7].

3.3 Other kinds of data

In LDA, the topics are distributions over words and this discrete distribution generates observations (words in documents). One advantage of LDA is that these choices for the topic parameter and data-generating distribution can be adapted to other kinds of observations with only small changes to the corresponding inference algorithms. As a class of models, LDA can be thought of as a *mixed-membership model* of grouped data—rather than associate each group of observations (document) with one component (topic), each group exhibits multiple components with different proportions. LDA-like models have been adapted to many kinds of data, including survey data, user preferences, audio and music, computer code, network logs, and social networks. We describe two areas where mixed-membership models have been particularly successful.

In population genetics, the same probabilistic model was independently invented to find ancestral populations (e.g., originating from Africa, Europe, the Middle East, etc.) in the genetic ancestry of a sample of individuals [27]. The idea is that each individual's genotype descends from one or more of the ancestral populations. Using a model much like LDA, biologists can both characterize the genetic patterns in those populations (the "topics") and identify how each individual expresses them (the "topic proportions"). This model is powerful because the genetic patterns in ancestral populations can be hypothesized, even when "pure" samples from them are not available.

LDA has been widely used and adapted in computer vision, where the inference algorithms are applied to natural images in the service of image retrieval, classification, and organization. Computer vision researchers have made a direct analogy from images to documents. In document analysis we assume that documents exhibit multiple topics and a collection of documents exhibits the same set of topics. In image analysis we assume that each image exhibits a combination of visual patterns and that the same visual patterns recur throughout a collection of images. (In a preprocessing step, the images are analyzed to form collections of "visual words.") Topic modeling for computer vision has been used to classify images [16], connect images and captions [4], build image hierarchies [2, 23, 31] and other applications.

4 Future directions

Topic modeling is an emerging field in machine learning, and there are many exciting new directions for research.

Evaluation and model checking. There is a disconnect between how topic models are evaluated and why we expect topic models are useful. Typically, topic models are evaluated in the following way. First, hold out a subset of your corpus as the test set. Then, fit a variety of topic models to the rest of the corpus and approximate a measure of model fit (e.g., probability) for each trained model on the test set. Finally, choose the the model that achieves the best held out performance.

But topic models are often used to organize, summarize and help users explore large corpora, and there is no technical reason to suppose that held-out accuracy corresponds to better organization or easier interpretation. One open direction for topic modeling is to develop evaluation methods that match how the algorithms are used. How can we compare topic models based on how interpretable they are?

This is the *model checking* problem. When confronted with a new corpus and a new task, which topic model should I use? How can I decide which of the many modeling assumptions are important for my goals? How should I move between the many kinds of topic models that have been developed? These questions have been given some attention by statisticians [9, 30], but they have been scrutinized less for the scale of problems that machine learning tackles. New computational answers to these questions would be a significant contribution to topic modeling.

Visualization and user interfaces. Another promising future direction for topic modeling is to develop new methods of interacting with and visualizing topics and corpora. Topic models provide new exploratory structure in large collections—how can we best exploit that structure to aid in discovery and exploration?

One problem is how to display the topics. Typically, we display topics by listing the most frequent words of each (see Figure 2), but new ways of labeling the topics—either by choosing different words or displaying the chosen words differently—may be more effective. A further problem is how to best display a document with a topic model. At the document-level, topic models provide potentially useful information about the structure of the document. Combined with effective topic labels, this structure could help readers identify the most interesting parts of the document. Moreover, the hidden topic proportions implicitly connect each document to the other documents (by considering a distance measure between topic proportions). How can we best display these connections? What is an effective interface to the whole corpus and its inferred topic structure?

These are user interface questions, and they are essential to topic modeling. Topic modeling algorithms show much promise for uncovering meaningful thematic structure in large collections of documents. But making this structure useful requires careful attention to information visualization and the corresponding user interfaces.

Topic models for data discovery. Topic models have been developed with information engineering applications in mind. As a statistical model, however, topic models should be able to tell us something, or help us form a hypothesis, about the data. What can we *learn* about the language (and other data) based on the topic model posterior? Some work in this area has appeared in political science [19], bibliometrics [17] and psychology [32]. This kind of research adapts topic models to measure an external variable of interest, a difficult task for unsupervised learning which must be carefully validated.

In general, this problem is best addressed by teaming computer scientists with other scholars to use topic models to help explore, visualize and draw hypotheses from their data. In addition to scientific applications, such as genetics or neuroscience, one can imagine topic models coming to the service of history, sociology, linguistics, political science, legal studies, comparative literature, and other fields where texts are a primary object of study. By working with scholars in diverse fields, we can begin to develop a new interdisciplinary computational methodology for working with and drawing conclusions from archives of texts.

5 Summary

We have surveyed *probabilistic topic models*, a suite of algorithms that provide a statistical solution to the problem of managing large archives of documents. With recent scientific advances in support of unsupervised machine learning—flexible components for modeling, scalable algorithms for posterior inference, and increased access to massive data sets—topic models promise to be an important component for summarizing and understanding our growing digitized archive of information.

References

[1] A. Asuncion, M. Welling, P. Smyth, and Y. Teh. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence*, 2009.

[2] E. Bart, M. Welling, and P. Perona. Unsupervised organization of image collections: Unsupervised organization of image collections: Taxonomies and beyond. *Transactions on Pattern Recognition and Machine Intelligence*, 2010.

[3] D. Blei, T. Griffiths, and M. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010.

[4] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134. ACM Press, 2003.

[5] D. Blei and J. Lafferty. Dynamic topic models. In *International Conference on Machine Learning*, pages 113–120, New York, NY, USA, 2006. ACM.

[6] D. Blei and J. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

[7] D. Blei and J. McAuliffe. Supervised topic models. In *Neural Information Processing Systems*, 2007.

[8] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[9] G. Box. Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A*, 143(4):383–430, 1980.

[10] J. Boyd-Graber and D. Blei. Syntactic topic models. In *Neural Information Processing Systems*, 2009.

[11] W. Buntine. Variational extentions to EM and multinomial PCA. In *European Conference on Machine Learning*, 2002.

[12] W. Buntine and A. Jakulin. Discrete component analysis. In *Subspace, Latent Structure and Feature Selection*. Springer, 2006.

[13] J. Chang and D. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4(1), 2010.

[14] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[15] G. Doyle and C. Elkan. Accounting for burstiness in topic models. In *International Conference on Machine Learning*, pages 281–288. ACM, 2009.

[16] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. *IEEE Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[17] S. Gerrish and D. Blei. A language-based approach to measuring scholarly impact. In *International Conference on Machine Learning*, 2010.

[18] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 537–544, Cambridge, MA, 2005. MIT Press.

[19] J. Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1, 2010.

[20] M. Hoffman, D. Blei, and F. Bach. On-line learning for latent Dirichlet allocation. In *Neural Information Processing Systems*, 2010.

[21] T. Hofmann. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence (UAI)*, 1999.

[22] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[23] J. Li, C. Wang, Y. Lim, D. Blei, and L. Fei-Fei. Building and using a semantivisual image hierarchy. In *Computer Vision and Pattern Recognition*, 2010.

[24] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *International Conference on Machine Learning*, pages 577–584, 2006.

[25] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, 2008.

[26] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *Knowledge Discovery and Data Mining*, 2006.

[27] J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, June 2000.

[28] J. Reisinger, A. Waters, B. Silverthorn, and R. Mooney. Spherical topic models. In *International Conference on Machine Learning*, 2010.

[29] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smith. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494. AUAI Press, 2004.

[30] D. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

[31] J. Sivic, B. Russell, A. Zisserman, W. Freeman, and A. Efros. Unsupervised discovery of visual object class hierarchies. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[32] R. Socher, S. Gershman, A. Perotte, P. Sederberg, D. Blei, and K. Norman. A Bayesian analysis of dynamics in free recall. In *Neural Information Processing Systems*, 2009.

[33] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2006.

[34] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[35] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[36] H. Wallach. Topic modeling: Beyond bag of words. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[37] C. Wang and D. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1982–1989. 2009.

[38] C. Wang, B. Thiesson, C. Meek, and D. Blei. Markov topic models. In *Artificial Intelligence and Statistics*, 2009.