2013-04-21

Probabilistic Explicit Topic Modeling

Joshua Aaron Hansen
Brigham Young University - Provo

Probabilistic Explicit Topic Modeling

Joshua A. Hansen

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Eric Ringger, Chair
Kevin Seppi
Charles Knutson

Department of Computer Science

Brigham Young University

April 2013

ABSTRACT

Probabilistic Explicit Topic Modeling

Joshua A. Hansen
Department of Computer Science, BYU
Master of Science

Latent Dirichlet Allocation (LDA) is widely used for automatic discovery of latent topics in document corpora. However, output from analysis using an LDA topic model suffers from a lack of identifiability between topics not only across corpora, but across runs of the algorithm. The output is also isolated from enriching information from knowledge sources such as Wikipedia and is difficult for humans to interpret due to a lack of meaningful topic labels.

This thesis introduces two methods for probabilistic explicit topic modeling that address these issues: Latent Dirichlet Allocation with Static Topic-Word Distributions (LDA-STWD), and Explicit Dirichlet Allocation (EDA). LDA-STWD directly substitutes precomputed counts for LDA topic-word counts, leveraging existing Gibbs sampler inference. EDA defines an entirely new explicit topic model and derives the inference method from first principles. Both of these methods approximate topic-word distributions *a priori* using word distributions from Wikipedia articles, with each article corresponding to one topic and the article title being used as a topic label. By this means, LDA-STWD and EDA overcome the nonidentifiability, isolation, and unintepretability of LDA output.

We assess the effectiveness of LDA-STWD and EDA by means of three tasks: document classification, topic label generation, and document label generation. Label quality is quantified by means of user studies. We show that a competing non-probabilistic explicit topic model handily beats both LDA-STWD and EDA as a dimensionality reduction technique in a document classification task. Surprisingly, we find that topic labels from another approach using LDA and *post hoc* topic labeling (called LDA+Lau) are on one corpus preferred over topic labels prespecified from Wikipedia. Finally, we show that LDA-STWD improves substantially upon the performance of the state of the art in document labeling.

Keywords: topic modeling, machine learning, Wikipedia

ACKNOWLEDGMENTS

Many an author has declared that those deserving thanks are too numerous to be listed. That's probably true. But that won't keep me from now trying to do exactly that:

To God,

- for a mind that works
- for truth on which to train it
- for a life full of blessings...

To my committee chair and advisor, Eric Ringger,

- for giving me a chance
- for humoring my crazy ideas
- for pulling my head down from the clouds sometimes...

To my family,

- for your love
- for moral, emotional, and spiritual support
- for being flexible about graduation dates!...

To my other committee members, Kevin Seppi and Charles Knutson,

- for reading this
- for your honest feedback
- for signing those papers!...

To Jen Bonnett,

- for welcoming me even when my paperwork quagmire must not have felt welcome

- for going to bat for me
- for support in stressful moments…

To my lab-mates Robbie Haertel, Dan Walker, Paul Felt, Kevin Cook, and Jeff Lund,

- for your friendship through the years
- for indispensible assistance
- for not erasing my volatile storage unless you had to…

…all my thanks, love, and gratitude.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The world is awash with data. One estimate puts the total number of books (not *copies* of books) ever published at some 129 million. The World Wide Web is estimated to contain at least 8.2 billion indexed pages. Twitter recently claimed 340 million tweets per day (124 billion annually), and Wikipedia boasts an aggregate 23 million articles across hundreds of languages. The data surge reaches beyond text, with 72 hours of video uploaded to YouTube each minute, and one report estimating that 1.8 *zettabytes* of enterprise server data were created in 2010 [13, 21, 25–27, 30].

Such a collection of data representing such a variety of activity did not exist prior to the emergence of the Internet as a mass technocultural phenomenon starting in the mid-1990s. Now these new data sources present ripe fruit for all manner of analysis, with insights in linguistics, anthropology, sociology, literature, organizational behavior, economics, and many other areas of human endeavor merely waiting to be discovered.

However, our own human limitations stand as the chief obstacle to advances at this scale. Even the most voracious human mind could never hope to take in more than the minutest fraction of this endless informatic ocean. Fortunately, the same digital systems that initially fostered the seemingly endless flood of data also prove useful in taming the deluge. One computational tool increasingly used to understand large datasets is probabilistic topic modeling, which "enables us to organize and summarize electronic archives at a scale that would be impossible" by human effort alone [1].

1.1 Modeling Topics with Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a probabilistic, generative model of document collections that models documents as mixtures over latent topics [2] where topics are defined as categorical distributions over some vocabulary. These topics can be discovered without supervision using any of a number of inference methods [2, 8, 19] and are useful for exploring document corpora in terms of the themes or topics that they contain [1, 7].

In order to give a better appreciation for the nature of topic model output, we now give an example taken from LDA inference on a corpus of excerpts from State of the Union messages. These messages have been given by United States presidents over the course of the country's history. Topic modeling allows for discovery of unifying themes across presidents and across time. Inference yielded these five topics (among others):

Topic 0	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**
defense	act	post	government	world
military	congress	service	great	peace
forces	law	mail	country	nations
strength	bill	department	means	nation
security	session	postal	experience	war

The following two documents from the corpus have been color-coded to represent the topic of some of the words:

Document 1 [Harry Truman]: We are working toward the time when the United Nations will control weapons of mass destruction and will have the forces to preserve international law and order. While the world remains unsettled, however, and as long as our own security and the security of the free world require, we will maintain a strong and well-balanced defense organization.... Under the principles of the United Nations Charter we must continue to share in the common defense of free nations against aggression. At

> the last session this Congress laid the basis for this joint effort. We now must put into effect the common defense plans that are being worked out.

> **Document 2 [Zachary Taylor]:** By an act of Congress . . . provision was made for extending post-office and mail accommodations to California and Oregon. Exertions have been made to execute that law, but the limited provisions of the act . . . render those exertions in a great degree ineffectual. More particular and efficient provision by law is required on this subject. The act of 1845 reducing postage has now. . . produced results fully showing that the income from such reduced postage is sufficient to sustain the whole expense of the service of the Post-Office . . . .

The topic model shows a strong connection between Truman's message and the national security and global peace movement topics (Topic 0 and Topic 4, respectively). Taylor's message is largely focused on the Postal Service, captured as Topic 2. This diversity corresponds well to the difference in concerns of mid-19th-century and mid-20th-century presidents. However, both had to deal with the Congress, which the topic model captures in the form of Topic 1, scattered through both documents. This example illustrates the ability of a topic model to discover thematic patterns in a corpus. In this case, those patterns largely correspond to what is known about the historical circumstances under which the documents in the corpus were created.

This example illustrates the ability of topic models to summarize document collections in terms of the topics discussed therein. This provides a level of semantic abstraction in many ways more instantly useable by the human mind than the documents themselves. Additionally, mapping documents into a reduced semantic space enables automated processing

1.2 A Critique of LDA

Topics discovered by latent topic models such as LDA usually did not originate in the corpus in question but were already in use in other contexts by the language community. Thus latent topic models often simply *re*discover topics already generally known, while being completely ignorant of the preexisting topics. When latent topic models encounter the same topic in different corpora, the implied linkage between the different instantiations of the topic can only be discovered by time-consuming *post hoc* comparisons. We term this the *inter-run identifiability problem*.

The inter-run identifiability problem can be thought of as an extension of the phenomenon of label switching in stochastic inference for mixture models, which demonstrates that the makeup of a topic is not guaranteed to persist across sampler iterations. This property also extends across iterations produce in samplers with different (as well as identical) initializations.

Topics generated by LDA and other latent topic models are also unsatisfactory in requiring post hoc labeling for easier human interpretation. Fast, automatic methods such as concatenating the top N words per topic produce labels that over-represent high-weighted terms and are often hard to interpret. Manual labeling may require substantial human effort to understand the contexts of usage and introduces obvious subjectivity. And better-performing labeling methods still require separate, often complicated processing in addition to the original topic modeling.

This thesis describes two approaches to overcoming LDA's deficiencies in inter-run identifiability and topic labeling. Both approaches involve adapting LDA for use with *explicit* rather than *latent* topics. The first approach, known as Latent Dirichlet Allocation with Static Topic-Word Distributions (LDA-STWD), simply substitutes precomputed topic-word counts into the standard LDA complete conditional, allowing inference by Gibbs sampling to be used essentially unmodified. The second approach, known as Explicit Dirichlet Allocation (EDA), is a new model similar to LDA but rederived from first principles after defining document-topic counts as explicit and not conditioned on a Dirichlet prior.

In both of these approaches, the explicit topics are estimated in advance from Wikipedia articles, one topic per article. Such a connection to Wikipedia provides a number of advantages in interpretation of topic model output:

- Each topic has a pre-existing, human-defined label (the article title).
- Display of topic model output can be enriched using Wikipedia article text, categories, links, etc.
- All topic model output—even across corpora—is instantly comparable because it exists in a single global space rather than many different spaces.

We exploit the first of these advantages to implement a document labeling algorithm that outperforms the current state of the art.

1.3 Thesis Statement

Probabilistic topic models that incorporate pre-specified topics with their accompanying labels can assign labels to topics and documents that reflect the meaning of the topics and the content of the individual documents in a collection more effectively than a combination of the LDA topic model and a *post hoc* labeling method.

1.4 Roadmap for Thesis

In the remainder of this thesis, we detail relevant related work in Chapter 2 and then define the two models in Sections 3.1 and 3.2. In Chapter 4 we describe the methods used to evaluate model output, and give results and analysis. Chapter 5 gives conclusions and future work.

Chapter 2

Related Work

In this chapter we situate the new LDA-STWD (Section 3.1) and EDA (Section 3.2) models within the general realm of topic modeling. As these models will be evaluated (Chapter 4) using a document labeling task in which document labels are generated from the topic label of the document's plurality topic, we also review the literature on automatic topic labeling.

2.1 Topical Representations of Documents

Numerous approaches to topical analysis of texts have been developed. These can be categorized according to whether the topic-document relationship is discovered using non-probabilistic techniques such as singular value decomposition (SVD) or through probabilistic inference methods, and according to whether topics are treated as latent and awaiting discovery or as pre-specified and explicit. A typology of these existing approaches—as well as of the approaches described in this thesis—is given in Table 2.1.

	Latent Topics	**Explicit Topics**
Non-Probabilistic	LSA†	ESA‡
Probabilistic	PLSA§, LDA$^\natural$	LDA-STWD$^\sharp$, EDA$^\flat$

Table 2.1: A typology of topical representations of documents. Note that PLSA is probabilistic but not fully generative, i.e. the model cannot be used to generate new documents. **Key to Abbreviations:** †Latent Semantic Analysis ‡Explicit Semantic Analysis §Probabilistic Latent Semantic Analysis $^\natural$Latent Dirichlet Allocation $^\sharp$Latent Dirichlet Allocation with Static Topic-Word Distributions (introduced in Section 3.1). $^\flat$Explicit Dirichlet Allocation (introduced in Section 3.2)

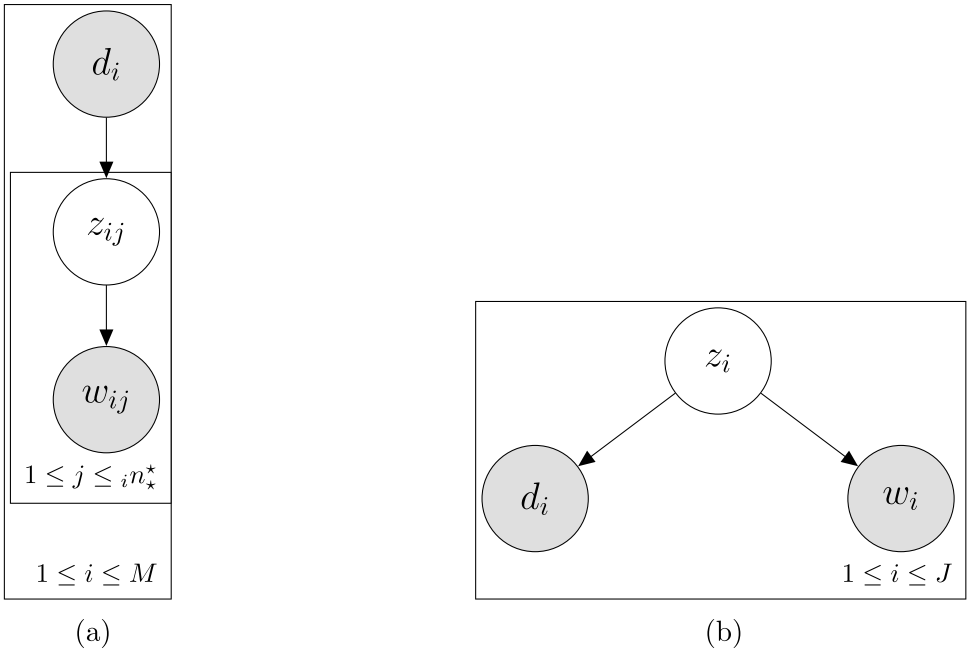
Latent Semantic Analysis [5] was the first in an important line of automatic topic discovery methods. It discovers topics or “concepts” by constructing a document-word matrix and reducing its dimensionality using singular value decomposition. This results in a model of the corpus consisting of k concepts corresponding to the k largest singular values in the decomposition.

In Probabilistic Latent Semantic Analysis [9] (Figure 2.1a), topics are still latent, but rather than use the tools of linear algebra to discover them, PLSA recasts LSA as a statistical inference problem. A partial generative process for documents is defined, allowing the latent document-topic and topic-word distributions to be estimated using an expectation maximization algorithm. The authors contend that this is more principled than using SVD to reduce document-word matrices. They also point out that by explicitly modeling “contexts” (topics), the model can represent polysemy (the occurrence of multiple senses for a given word), the inability to do so being a key failing of LSA.

The *aspect model* (Figure 2.1b) described by Hofmann and Puzicha is a probabilistic model in which both the document and the type of each word token is conditioned on a latent “aspect” variable. They show that the aspect model outperforms PLSA on an information retrieval task [10].

Latent Dirichlet Allocation [2] (Figure 2.2a) is likewise a probabilistic counterpart to LSA with latent topics, but unlike PLSA it is fully generative. This is achieved by specifying a Dirichlet prior distribution over the per-document mixing proportions—$P(\theta \mid \alpha)$—and another over the topic-word distributions—$P(\phi \mid \beta)$. In addition to the mathematical satisfaction of producing a fully generative model that can generate new documents and compute probabilities of previously unseen documents, LDA also consistently outperforms PLSA in empirical evaluations carried out by its creators. As a result, LDA has become the basis for a myriad of derivative models.

LDA is defined by the generative process for generating a corpus of documents given in Algorithm 1, where ${}_{d}n_{\star}^{\star}$ represents the number of tokens in document d. This process



(a)

(b)

Figure 2.1: Graphical models for (a) Probabilistic Latent Semantic Analysis and (b) the aspect model described by Hofmann and Puzicha. The d_i variables represent documents, z_{ij} and z_i represent latent token topic assignments, and w_{ij} and w_i represent observed word tokens.

implies the hierarchical Bayesian network in Figure 2.2.

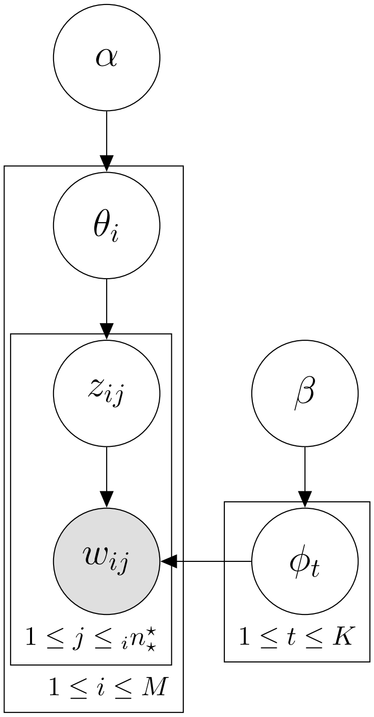
Gibbs sampling for LDA is principally defined by the complete conditional distribution for token-topic assignments [8]:

$$P\left(z_{ij} = k \;\middle|\; \overline{\overline{z}}_{\neg ij}, \overline{\overline{w}}\right) \propto \frac{{}_\star n^{ij}_k + \beta}{{}_\star n^\star_k + J\beta} \cdot \frac{{}_i n^\star_k + \alpha}{{}_i n^\star_\star + K\alpha} \tag{2.1}$$

where, excluding the current assignment of z_{ij} from the counts,

Algorithm 1 Generative process for LDA.

```
for each topic index $1 \leq k < K$ do
  Sample $\phi_k \sim Dirichlet(\beta_k)$
for each document index $1 \leq i < D$ do
  Sample length $K$ vector $\theta_d \sim Dirichlet(\alpha_i)$
  for each document token index $1 \leq j < {}_d n^\star_\star$ do
    Sample topic $z_{ij} \sim Categorical(\theta_i)$
    Sample word token $w_{ij} \sim Categorical(\phi_{z_{ij}})$
```



(a)

$$\begin{aligned}
w_{ij}|z_{ij}, \phi_{z_{ij}} &\sim Categorical(\phi_{z_{ij}}) \\
z_{ij}|\theta_i &\sim Categorical(\theta_i) \\
\theta_i|\alpha &\sim Dirichlet(\alpha) \\
\phi_i|\beta &\sim Dirichlet(\beta)
\end{aligned}$$

(b)

Figure 2.2: (a) Graphical model for Latent Dirichlet Allocation (b) Conditional probability distributions for Latent Dirichlet Allocation

- ${}_\star n_k^{ij}$ is the number of times words of the same type as token w_{ij} are assigned to topic k,
- ${}_\star n_t^\star$ is the number of times topic t is assigned to any word,
- J is the total number of tokens,
- ${}_i n_k^\star$ is the number of tokens in document i assigned to topic k,
- ${}_i n_\star^\star$ is the number of tokens in document i, and
- K is the number of topics

Explicit Semantic Analysis [6] represents documents in Wikipedia-derived semantic space. This is done by ranking each Wikipedia concept (corresponding to an article) by how well its terms are represented within a given document. The selected results are impressive, though evaluation is done only in terms of a word relatedness task. Though the algorithm is not formulated probabilistically—indeed, the authors make no mention of probabilistic approaches whatsoever—it may lend itself to a probabilistic reinterpretation.

Mr. LDA is a distributed variational inference method for LDA. In one application of their model, the authors provide "informed priors" for 100 softly prespecified topics, allowing active guidance of topic discovery while still keeping topics latent [31]. However, it is assumed that a small number of informed priors appropriate to the corpus are chosen in advance—an impractical requirement in many scenarios due to the lack of prior knowledge about a given corpus.

One analysis of LSA, LDA, and ESA finds significant benefit from explicit topics [4]. Explicit topics are intuitively appealing because they provide a semantic summary of documents in terms of human-defined concepts rather than machine-discovered topics. Yet ESA assumes a document consists of a single topic—or implies as much by seeking Wikipedia concepts that maximize the relatedness of the document as a whole—whereas documents are more effectively modeled as mixtures of topics—as in LDA—to be discovered jointly. This thesis addresses this issue by putting explicit topic corpus modeling in a probabilistic framework.

Labeled LDA is a variation on LDA in which document topics are assigned *a priori* [20]. The authors apply it to a corpus of web pages with labels taken from the *del.icio.us* social bookmarking site. The essential nature of the model is that topic-word distributions remain latent, while document-topic distributions are observed—a complementary approach to ours, which is to treat topic-word distributions as observed while document-topic distributions remain latent.

2.2 Topic Labeling

Topic labeling is the task of automatically generating a label given a topic, where *topic* is defined as a categorical distribution over the elements in some vocabulary. These labels are intended to aid human interpretation of topic model output. A number of prior approaches to topic labeling have been described. Mei, Shen, and Zhai first defined the topic labeling problem as the production of "a sequence of words which is semantically meaningful and covers the latent meaning of [the distribution over words in a topic]" [18]. The authors also describe a typical approach to the problem: generate candidate labels, rank the candidate labels according to some measure of relevance, and select a high-scoring candidate or candidates to serve as the label. Their candidates are generated from a corpus of relevant text, and are scored using various association measures computed on that corpus. Human annotators rated the quality of the selected high-scoring candidate labels, as well as the quality of baseline labels generated as the concatenation of the top k words. Humans preferred labels from the new method over those of the baseline.

Magatti, Calegari, Ciucci, *et al.* [17] vary this approach by taking candidate labels from the Open Directory Project hierarchy. Rather than choosing highly ranked candidates according to a single relevance score, they employ a complex scheme for combining the output of multiple relevance measures into a single answer. Unfortunately, they do not empirically validate the quality of the labels generated by their method.

Spitkovsky and Chang generate a mapping from phrases to Wikipedia URLs and back by observing the frequency of various anchor text phrases on hyperlinks pointing to Wikipedia pages [23]. The forward mapping could be applied to derive a distribution over Wikipedia article topics given the text of a document, but the authors do not develop such a method.

Lau, Newman, Karimi, *et al.* [15] follow much the same approach as Mei, Shen, and Zhai but define the task more narrowly as selecting a subset of the words contained in a topic, from which a label can be generated. They also introduce a supervised approach, with accompanying performance gains. Lau, Grieser, Newman, *et al.* [14] extend this work by generating label candidates from relevant Wikipedia article titles as well as from top words in a topic. Association measures computed on the entire English Wikipedia are then used to rank the candidates for final label selection. By outperforming the Mei, Shen, and Zhai approach (and presumably that of Lau, Newman, Karimi, *et al.* on account of generating candidate labels that are a superset of those in that paper) this approach distinguishes itself as the current state of the art in automatic topic labeling.

The Lau, Grieser, Newman, *et al.* topic labeling method is not without its difficulties. A flowchart representing the processing steps that make up the algorithm (Figure 2.3, page 13) reveals the complexity of the method. It was not tested on all of the topics output by LDA due to two significant filtering steps, designated in the flowchart as "Coherence >= 0.4" and "Half or More Of Top 10 Terms Default Nominal". The first is a requirement that the semantic relatedness of the top words in a topic must exceed a threshold of 0.4. The second is a requirement that half or more of the top 10 terms in a topic have a noun part of speech as determined by a POS tagger. However, the reliance on Google searches is the most problematic because of the opacity of that method. The details of Google's ever-changing algorithm are not generally known; thus any dependence on it renders part of the Lau, Grieser, Newman, *et al.* algorithm essentially undefined.

In all of the above approaches, topic labels are generated *post hoc.* By contrast, Latent Dirichlet Allocation with Static Topic-Word Distributions and Explicit Dirichlet Allocation both provide integrated approaches in which the topic label is included in the output of the topic model itself.

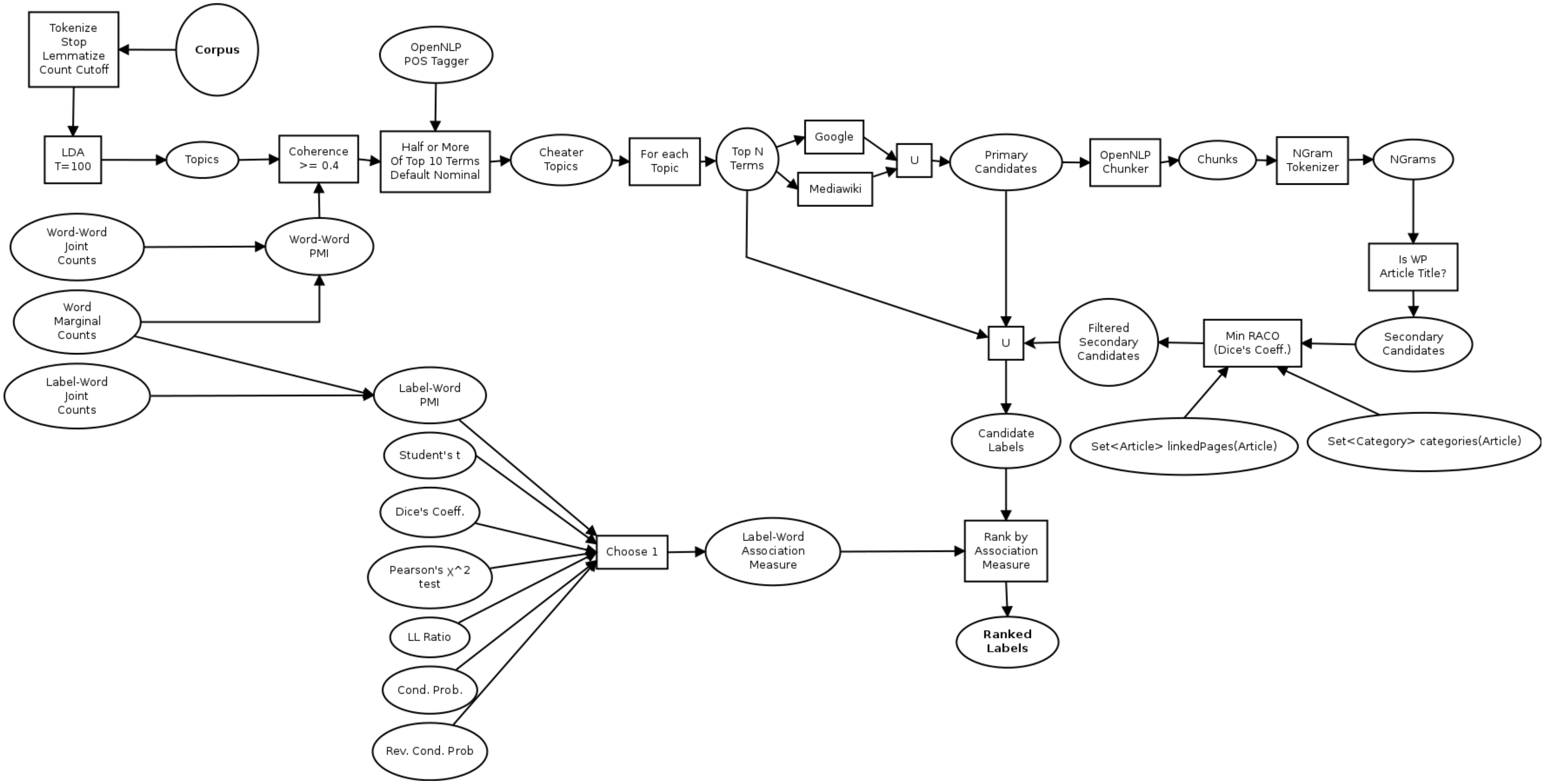


Figure 2.3: A flowchart representing the Lau, et al. algorithm. The complexity of the algorithm makes it difficult to implement and to apply.

Chapter 3

Probabilistic Explicit Topic Modeling

In this chapter we describe two probabilistic explicit topic models. The first is Latent Dirichlet Allocation with Static Topic-Word Distributions, or LDA-STWD, which is identical to LDA with the exception that topic-word distributions are prespecified. The second is Explicit Dirichlet Allocation, or EDA, which is a probabilistic explicit topic model derived from first principles.

3.1 Latent Dirichlet Allocation with Static Topic-Word Distributions (LDA-STWD)

As mentioned earlier, Ramage, Hall, Nallapati, *et al.* describe Labeled LDA, a variation on LDA in which document topics are assigned *a priori* [20]. They apply it to a corpus of web pages with labels taken from the *del.icio.us* social bookmarking site. The essential nature of the model is that topic-word distributions remain latent, while document-topic distributions are observed.

LDA can be adapted to use observed topic-word distributions but latent document-topic distributions. As a first, albeit *ad hoc* attempt, we adapt the Gibbs sampler for LDA developed by Griffiths and Steyvers [8] to model document corpora in terms of predefined topic-word distributions. This is done by reformulating the complete conditional in terms of the counts found in the *topic corpus* (such as Wikipedia) rather than in the *target corpus* (the corpus that the model is being applied to.) These counts do not vary during the course of sampling. Nevertheless, the d-separability implications of rendering ϕ observed are ignored—the topic corpus counts are directly substituted for the counts that are typically sampled.

We call this first approach Latent Dirichlet Allocation with Static Topic-Word Distributions. The LDA Bayesian network and conditional probability distributions are not changed, but the complete conditional distribution (eq. 5 of Griffiths and Steyvers) is reformulated in terms of precomputed topic-word counts:

$$P\left(z_{ij} = k \mid \overline{\overline{z}}_{\neg ij}, \overline{\overline{w}}\right) \propto \frac{{}_{\star}\lambda_{ij}^{k} + \beta}{{}_{\star}\lambda_{k}^{\star} + J\beta} \cdot \frac{{}_{i}n_{k}^{\star} + \alpha}{{}_{i}n_{\star}^{\star} + K\alpha} \tag{3.1}$$

where

- ${}_{\star}\lambda_{ij}^{k}$ is the number of times words of the same type as token w_{ij} are assigned to topic k in the topic corpus (excluding token w_{ij} itself),
- ${}_{\star}\lambda_{t}^{\star}$ is the number of times topic t is assigned to any word in the topic corpus (excluding token w_{ij} itself),
- J is the total number of tokens in the topic corpus,
- ${}_{i}n_{k}^{\star}$ is the number of tokens in document i assigned to topic k (excluding w_{ij} where applicable),
- ${}_{i}n_{\star}^{\star}$ is the number of tokens in document i (excluding w_{ij}), and
- K is the number of topics in the topic corpus

The resultant model insists that the documents in the target corpus were generated from a set of preexisting topics and topic word distributions. The sampler is thus expected to allocate the token-level topic assignments amongst those topics that best describe the corpus. In other words, the sampler simply chooses *which topics* to use, but not which words the topics will contain.

Examples of model output can be found in A.1 and A.2 in Appendix A.

3.2 Explicit Dirichlet Allocation

In Section 3.1 we described a variation on the standard LDA topic model in which topic-word distributions are specified *a priori* from distributions in a separate *topic corpus*. The resulting LDA-STWD model is somewhat *ad hoc*, being a repurposing of LDA rather than a purpose-built model *sui generis*. In this section we derive such a probabilistic explicit topic model—Explicit Dirichlet Allocation.[1]

Explicit Dirichlet Allocation is a probabilistic graphical model that adapts LDA for use with predefined, explicit topics. Comparison of the LDA graphical model (fig. 2.2a) with that of EDA (fig. 3.1) makes the differences between the models clear. In EDA, documents are still modeled as a probabilistic admixture of topics, where a topic is a categorical distribution over words in a vocabulary. However, the topics (ϕs) in EDA are treated as observed or explicit, whereas in LDA they are considered unobserved or latent. Additionally, EDA's ϕ is no longer conditioned on a parameter β.

The relevant conditional probability distributions can be specified as

$$\begin{aligned} w_{ij} | z_{ij}, \phi_{z_{ij}} &\sim Categorical(\phi_{z_{ij}}) \\ z_{ij} | \theta_i &\sim Categorical(\theta_i) \\ \theta_i | \alpha &\sim Dirichlet(\alpha) \end{aligned}$$

for all documents i and for all tokens j in that document.

Definitions:

- M is the number of documents in the corpus
- ${}_i n^{\star}_{\star}$ is the number of tokens in document i

[1] A note on nomenclature: as my esteemed colleague, Robbie Haertel, has observed, Explicit Dirichlet Allocation is something of a misnomer—*Latent* Dirichlet Allocation was so named on account of the latent nature of both the topic-word distributions and the document-topic distributions. *Explicit* Dirichlet Allocation, on the other hand, retains the latency of document-topic distributions, the topic-word distributions only being specified as explicit. Thus this model would be more properly termed Semi-latent Dirichlet Allocation. However, for purposes of analogy to Explicit Semantic Analysis, and to emphasize that the key difference in the new model is the rendering of a variable explicit rather than latent, the less accurate name of Explicit Dirichlet Allocation was retained.

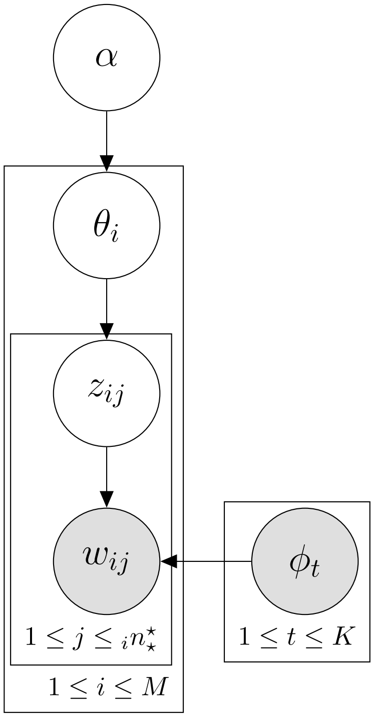


Figure 3.1: Graphical model for Explicit Dirichlet Allocation

- ${}_{i}n^{\star}_{k}$ is the number of tokens in document i assigned to topic k
- $\overline{\overline{z}}_{\neg mn}$ denotes the set of all token topic variables except z_{mn}.
- ${}_{i}n^{\star}_{k \neg mn}$ denotes the number of tokens in document i assigned to topic k, with the exception of token w_{mn} (if it occurs in the document)

The distribution of interest which we seek to derive is

$$\begin{aligned} P\left(\overline{\overline{w}}, \overline{\overline{z}} \mid \boldsymbol{\phi}, \boldsymbol{\alpha}\right) &= \prod_{i=1}^{M} P\left(\boldsymbol{w}_i, \boldsymbol{z}_i \mid \boldsymbol{\phi}, \boldsymbol{\alpha}\right) && (3.2) \\ &= \prod_{i=1}^{M} \int_{\theta_i} P\left(\boldsymbol{w}_i, \boldsymbol{z}_i, \theta_i \mid \boldsymbol{\phi}, \boldsymbol{\alpha}\right) d\theta_i && (3.3) \end{aligned}$$

Expanding according to the graphical model factorization:

$$P\left(\overline{\overline{w}}, \overline{\overline{z}} \mid \boldsymbol{\phi}, \boldsymbol{\alpha}\right) = \prod_{i=1}^{M} \int_{\theta_i} P\left(\theta_i \mid \alpha\right) \prod_{j=1}^{{}_{i}n^{\star}_{\star}} P\left(z_{ij} \mid \theta_i\right) P\left(w_{ij} \mid z_{ij}\right) d\theta_i \qquad (3.4)$$

As $P\left(w_{ij}\mid z_{ij}\right)$ is not bound by θ_i we move it outside of the integral:

$$P\left(\overline{\overline{w}},\overline{\overline{z}}\mid\boldsymbol{\phi},\boldsymbol{\alpha}\right)=\prod_{i=1}^{M}\left[\prod_{j=1}^{{}_in^\star_\star}P\left(w_{ij}\mid z_{ij}\right)\right]\int\limits_{\theta_i}P\left(\theta_i\mid\alpha\right)\prod_{j=1}^{{}_in^\star_\star}P\left(z_{ij}\mid\theta_i\right)d\theta_i \tag{3.5}$$

We now focus on the integral over θ_i, first expanding the probability distributions to their respective probability functions:

$$\int\limits_{\theta_i}P\left(\theta_i\mid\alpha\right)\prod_{j=1}^{{}_in^\star_\star}P\left(z_{ij}\mid\theta_i\right)d\theta_i=\int\limits_{\theta_i}\left[\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}\theta_{ik}^{\alpha_k-1}\right]\prod_{j=1}^{{}_in^\star_\star}\theta_{i,z_{ij}}d\theta_i \tag{3.6}$$

where $\Gamma\left(\cdot\right)$ is the gamma function. We reformulate the rightmost product topic-wise, revealing a function reminiscent of the Dirichlet distribution.

$$=\int\limits_{\theta_i}\left[\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}\theta_{ik}^{\alpha_k-1}\right]\prod_{k=1}^{K}\theta_{ik}^{{}_in_k^\star}d\theta_i \tag{3.7}$$

In other words, we have restated the product over tokens $\prod_{j=1}^{{}_in^\star_\star}\theta_{i,z_{ij}}$ as a product over topics $\prod_{k=1}^{K}\theta_{ik}^{{}_in_k^\star}$ by exponentiating the document-topic probability θ_{ik} by the number of tokens in document i assigned to topic k, ${}_in_k^\star$. This transformation enables the merger of the two products over K:

$$=\int\limits_{\theta_i}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}\theta_{ik}^{{}_in_k^\star+\alpha_k-1}d\theta_i \tag{3.8}$$

This integral can be transformed into the Dirichlet probability density function and thus eliminated because it integrates to 1, as follows:

$$\int\limits_{\theta_i}P\left(\theta_i\mid\alpha\right)\prod_{j=1}^{{}_in^\star_\star}P\left(z_{ij}\mid\theta_i\right)d\theta_i=\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\int\limits_{\theta_i}\prod_{k=1}^{K}\theta_{ik}^{{}_in_k^\star+\alpha_k-1}d\theta_i$$

$$
\begin{aligned}
&= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \cdot 1 \cdot 1 \cdot \int_{\theta_i} \prod_{k=1}^{K} \theta_{ik}^{{}_i n_k^\star + \alpha_k - 1} d\theta_i \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)} \frac{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)} \int_{\theta_i} \prod_{k=1}^{K} \theta_{ik}^{{}_i n_k^\star + \alpha_k - 1} d\theta_i \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)} \frac{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)} \int_{\theta_i} \prod_{k=1}^{K} \theta_{ik}^{{}_i n_k^\star + \alpha_k - 1} d\theta_i \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)} \int_{\theta_i} \frac{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)} \prod_{k=1}^{K} \theta_{ik}^{{}_i n_k^\star + \alpha_k - 1} d\theta_i
\end{aligned}
$$

$$
\int_{\theta_i} P\left(\theta_i \mid \alpha\right) \prod_{j=1}^{{}_i n_\star^\star} P\left(z_{ij} \mid \theta_i\right) d\theta_i = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)} \tag{3.9}
$$

Substituting the right side of eq. (3.9) for the integral in eq. (3.5) and then rearranging:

$$
\begin{aligned}
P\left(\overline{\overline{w}}, \overline{\overline{z}} \mid \boldsymbol{\phi}, \boldsymbol{\alpha}\right) &= \prod_{i=1}^{M} \left[\prod_{j=1}^{{}_i n_\star^\star} P\left(w_{ij} \mid z_{ij}\right) \right] \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)} && (3.10) \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \prod_{i=1}^{M} \frac{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)} \prod_{j=1}^{{}_i n_\star^\star} P\left(w_{ij} \mid z_{ij}\right) && (3.11)
\end{aligned}
$$

Replacing $P\left(w_{ij} \mid z_{ij}\right)$ with its probability function gives the full collapsed joint distribution:

$$
\begin{aligned}
P\left(\overline{\overline{w}}, \overline{\overline{z}} \mid \boldsymbol{\phi}, \boldsymbol{\alpha}\right) &= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \prod_{i=1}^{M} \frac{\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)} \prod_{j=1}^{{}_i n_\star^\star} \phi_{z_{ij}, w_{ij}} && (3.12) \\
&\propto \prod_{i=1}^{M} \left[\prod_{k=1}^{K} \Gamma\left({}_i n_k^\star + \alpha_k\right) \right] \prod_{j=1}^{{}_i n_\star^\star} \phi_{z_{ij}, w_{ij}} && (3.13)
\end{aligned}
$$

For Gibbs sampling, we derive the complete conditional distribution for any token topic assignment given all other model parameters:[2]

[2] Special thanks to Robbie Haertel for a correction in this section.

Algorithm 2 Gibbs Sampler for Explicit Dirichlet Allocation.

```
Input: Word vector $\overline{\overline{w}}$ where $w_{ij}$ is the index of the word type at position $j$ of document $i$.
  Randomly initialize each element of $\overline{\overline{z}}$ to values in $\{1, 2, \ldots, K\}$
  for $1 \leq i < M$ do
    for $1 \leq j < {}_i n_\star^\star$ do
      Sample $z_{ij} \sim UniformCategorical(K)$

  Sample topic assignments from the complete conditional (eq. (3.18)) for $N$ iterations
  for $N$ iterations do
    for $1 \leq i < M$ do
      for $1 \leq j < {}_i n_\star^\star$ do
        Sample $z_{ij} \sim P\left(z_{ij} = k \mid \overline{\overline{z}}_{\neg ij}, \overline{\overline{w}}\right)$
```

$$P\left(z_{mn} = t \mid \overline{\overline{z}}_{\neg mn}, \overline{\overline{w}}, \boldsymbol{\phi}, \boldsymbol{\alpha}\right) \propto P\left(z_{mn} = t, \overline{\overline{z}}_{\neg mn}, \overline{\overline{w}} \mid \boldsymbol{\phi}, \boldsymbol{\alpha}\right) \tag{3.14}$$

$$\begin{aligned} &\propto \cancel{\left[\prod_{i=1, i\neq m}^{M} \Gamma\left(\sum_{k=1}^{K} {}_i n_k^\star + \alpha_k\right)\right]} \prod_{j=1}^{{}_i n_\star^\star} \phi_{z_{ij}, w_{ij}} \\ &\times \Gamma\left({}_m n_t^\star + \alpha_t + 1\right) \prod_{k=1, k\neq t}^{K} \Gamma\left({}_m n_k^\star + \alpha_k\right) \\ &\times \cancel{\left[\prod_{j=1, j\neq n}^{{}_i n_\star^\star} \phi_{t, w_{mn}}\right]} \phi_{t, w_{mn}} \end{aligned} \tag{3.15}$$

$$\propto \left[\Gamma\left({}_m n_t^\star + \alpha_t + 1\right) \prod_{k=1, k\neq t}^{K} \Gamma\left({}_m n_k^\star + \alpha_k\right) \cdot \frac{\Gamma\left({}_m n_t^\star + \alpha_t\right)}{\Gamma\left({}_m n_t^\star + \alpha_t\right)}\right] \times \phi_{t, w_{mn}} \tag{3.16}$$

$$\propto \frac{\Gamma\left({}_m n_t^\star + \alpha_t + 1\right)}{\Gamma\left({}_m n_t^\star + \alpha_t\right)} \cancel{\prod_{k=1}^{K} \Gamma\left({}_m n_k^\star + \alpha_k\right)} \times \phi_{t, w_{mn}} \tag{3.17}$$

Finally, a well-known gamma function identity completes the reduction:

$$P\left(z_{mn} = t \mid \overline{\overline{z}}_{\neg mn}, \overline{\overline{w}}, \boldsymbol{\phi}, \boldsymbol{\alpha}\right) \propto \left({}_m n_t^\star + \alpha_t\right) \phi_{t, w_{mn}} \tag{3.18}$$

Because the topic-word counts in EDA are precomputed, the Gibbs sampler algorithm is somewhat simpler than LDA's. Pseudocode for the sampling algorithm is given in Algorithm 2.

Examples of model output can be found in A.3 and A.4 in Appendix A.

Chapter 4

Experiments

In this chapter we describe experiments, results, and analysis designed to assess the quality of the Latent Dirichlet Allocation with Static Topic-Word Distributions (LDA-STWD) and Explicit Dirichlet Allocation (EDA) topic models. We begin with a preliminary discussion of data (4.1), a description of our general approach to evaluating label quality using elicited human judgments (4.2), and our implementation of the topic labeling algorithm described by Lau, Grieser, Newman, *et al.* [14]. We then describe experiments used to set the topic count parameter for LDA+Lau, which is the combination of the Lau, Grieser, Newman, *et al.* labeling method with LDA topic modeling (4.4). We then describe an evaluation of topic label quality, comparing the $\langle \mathit{topic}, \mathit{label} \rangle$ pairs used by LDA-STWD and EDA to those generated by LDA+Lau by means of human annotation (4.5). We next describe a similar comparison regarding *document* label quality (4.6). Finally, we describe an assessment of the quality of the generated topics themselves, pitting LDA-STWD and EDA against the non-probabilistic, explicit ESA topic in a secondary text classification task.

4.1 Data

Two datasets are used: Reuters 21578 and SOTU Chunks.

Reuters 21578 is a widely-used newswire dataset consisting of 11 367 news reports in 82 business-centric categories [16]. These categories serve as natural labels for text classification. We tokenize the dataset's documents using Mallet's default stopwords list, with the addition of the words “blah”, “reuter”, and “reuters”. This yields a dataset of 827 841 tokens. Evaluations performed using this dataset also omit documents with fewer than 100 characters, fewer than 80 tokens, or more than 20% of characters being numeric.

The SOTU Chunks dataset was derived from the corpus of State of the Union messages[1] delivered once a year (with minor variations) by United States presidents beginning with George Washington’s first in 1790 and continuing to the present day. The messages are topically diverse due to the wide range of issues, times, and circumstances addressed. We split 223 publicly available State of the Union messages into 11 413 two-paragraph chunks to aid comprehension by human judges in the document label quality task (4.6).

4.2 Human Judgments of Label Quality

To assess LDA-STWD’s and EDA’s ability to find appropriate labeled topics given an input corpus, we compare document and topic labels from those algorithms to labels generated by the LDA+Lau combination of LDA and the topic labeling algorithm described by Lau, Grieser, Newman, *et al.* To do this, we run all three algorithms on the same target corpus. Their output was then used to create annotation tasks for Amazon Mechanical Turk to evaluate how well LDA-STWD or EDA labels fit their topics and documents, respectively. Amazon Mechanical Turk is a popular crowdsourcing platform that allows *requesters* to submit tasks for human *workers* to complete. Mechanical Turk has been shown to be an effective tool for a variety of data annotation tasks [22]. In particular it has been successfully used to evaluate the output of topic models [3][11].

4.2.1 Reimplementation of Lau, et al.

The topic labeling algorithm described in Lau, Grieser, Newman, *et al.* [14] is key to our evaluation strategy as it represents the state of the art in automatic generation of topic labels. As a working implementation of the algorithm was not readily available, it was necessary

[1] Originally downloaded from `http://en.wikisource.org/wiki/Portal:State_of_the_Union_Speeches_by_United_States_Presidents`.

to write our own implementation. Due to ambiguities in the description of the algorithm, it was not possible to guarantee that our implementation is functionally identical to that of the original authors. Our implementation is different (or potentially different) in the following ways:

- Lau, Grieser, Newman, *et al.* originally used Google searches to expand the set of candidate labels. Google's enforced use policy now prevents any more than a very small number of queries. To compensate, we use our own index of the documents, created using the Lucene search engine.
- Our association measure was calculated on the Wikipedia article page dump from 3 April 2012, whereas Lau, Grieser, Newman, *et al.* used an earlier version.
- We only use pointwise mutual information (PMI) for rating of label candidates, whereas Lau, Grieser, Newman, *et al.* use a large number of association measures. We justify this restriction on the basis of their results, which show PMI as the strongest performer.
- Up to five fallback candidates taken from the top 20 words in a topic are added to the label candidate set, but only when the word happens to also be the title of a Wikipedia article. Lau, Grieser, Newman, *et al.* use the top five words from the topic without any such restriction.
- Lau, Grieser, Newman, *et al.* resolved disambiguation pages to all of the pages list as potential disambiguators. However, we ignore disambiguation pages as they contain little unique semantic content.

4.3 Convergence

We investigate the convergence properties of LDA-STWD by calculating log-likelihood of the data using the model at each iteration. Convergence plots of EDA were not generated and remain for future work. Figure 4.1 and Figure 4.2 illustrate that on SOTU Chunks and Reuters 21578, respectively, LDA-STWD sees rapid convergence, with the rate of change in log-likelihood dropping dramatically by the tenth iteration. This rapid convergence relative to LDA proper can be attributed to the lack of sampling of the topic-word distributions.

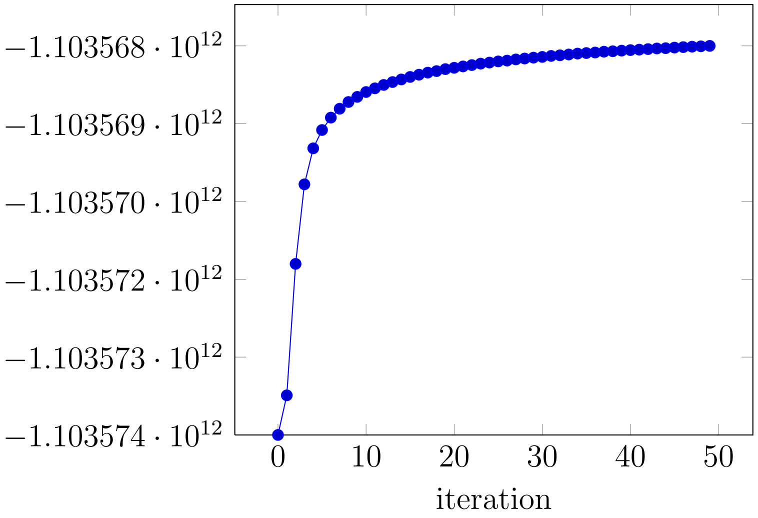


Figure 4.1: Log-likelihood convergence plot for LDA-STWD on SOTU Chunks across 50 iterations.

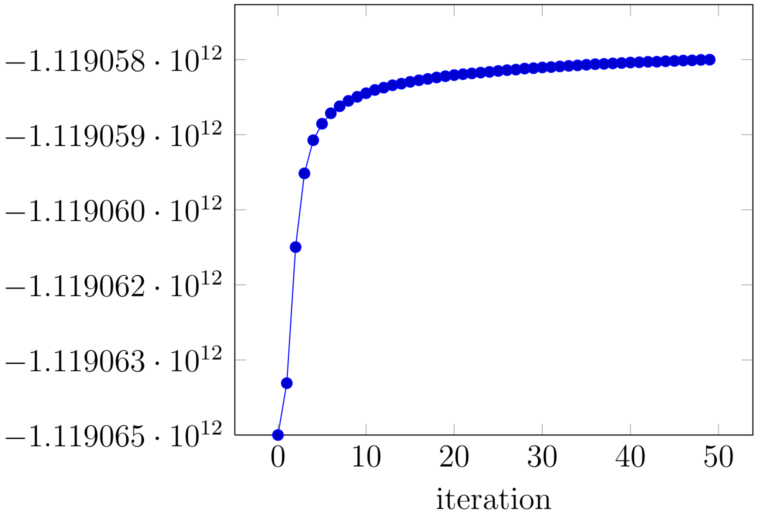


Figure 4.2: Log-likelihood convergence plot for LDA-STWD on Reuters 21578 across 50 iterations.

4.4 LDA+Lau Topic Count Calibration

In this section we describe experiments conducted to determine best-performing topic counts for use in the LDA+Lau algorithm.

4.4.1 Experiment

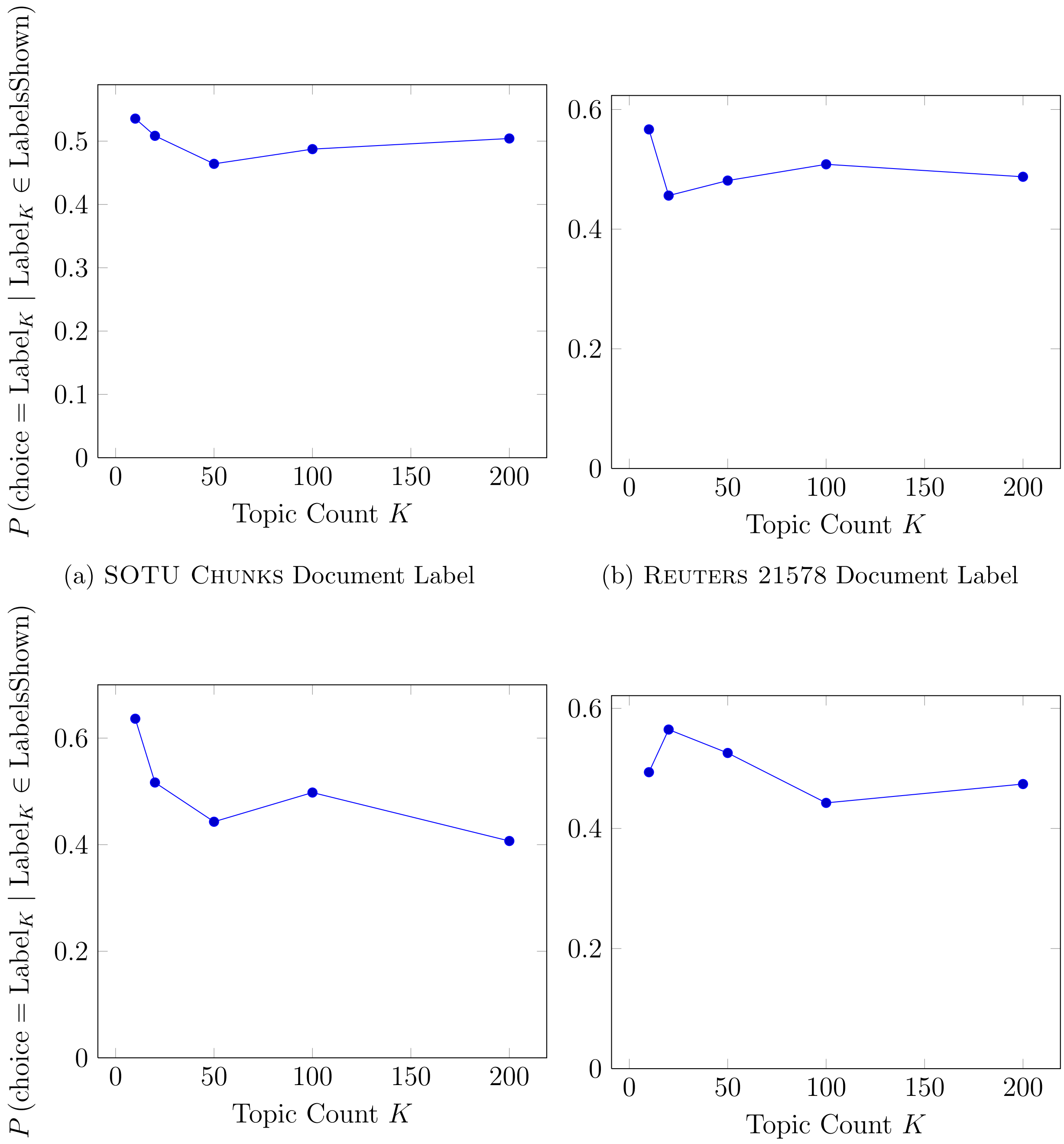
LDA is parameterized by a number of topics K. As LDA is a key component of LDA+Lau, it thus becomes necessary to set K in some manner. We do this by choosing the topic count from a predetermined set that is most often preferred over labels produced with other topic counts. Topic count calibration user studies were performed for the SOTU Chunks and Reuters 21578 datasets, on the same topic labeling and document labeling tasks used as the main form of validation for LDA-STWD and EDA (described in Section 4.5 and Section 4.6, respectively). In this case, however, LDA+Lau was compared to itself using different numbers of topics. Topic counts tried were those in $\{10, 20, 50, 100, 200\}$, with 5 runs each. 3 users annotated 25 comparisons per pair for each of the 10 unordered pairs of the possible topic counts.

4.4.2 Results

The results of the calibration experiments are illustrated in Figure 4.3, where $P\left(\text{choice} = \text{Label}_K \mid \text{Label}_K \in \text{LabelsShown}\right)$ indicates how often a human annotator selected the label generated using topic count K when shown side-by-side with a label generated using any other topic count.

In the SOTU Chunks document labeling calibration study (4.3a), the probability of participants choosing document labels produced by 10-topic LDA+Lau was $P\left(\text{choice} = \text{Label}_{10} \mid \text{Label}_{10} \in \text{LabelsShown}\right) \approx 0.54$—a higher rate of preference than for any other topic count.

The result for Reuters 21578 was nearly identical (4.3b), with $P\left(\text{choice} = \text{Label}_{10} \mid \text{Label}_{10} \in \text{LabelsShown}\right) \approx 0.57$, beating out all other topic counts



(a) SOTU Chunks Document Label

(b) Reuters 21578 Document Label

(c) SOTU Chunks Topic Label

(d) Reuters 21578 Topic Label

Figure 4.3: Topic count calibration results for the document label quality (a and b) and topic label quality (c and d) tasks.

evaluated.

On the topic labeling task, SOTU Chunks was most preferred at 10 topics, with $P\,(\text{choice} = \text{Label}_{10} \mid \text{Label}_{10} \in \text{LabelsShown}) \approx 0.64$ (4.3c). For Reuters 21578, preference was maximized at $K = 20$, with $P\,(\text{choice} = \text{Label}_{20} \mid \text{Label}_{20} \in \text{LabelsShown}) \approx 0.56$ (4.3d).

The small topic counts of $K = 10$ and $K = 20$ may appear suspect given that a more typical value for corpora of this size would be at least $K = 100$. The intention of the calibration was to allow LDA-STWD and EDA approaches to be compared to LDA+Lau at its best-performing setting, while not incurring the annotation cost required to perform the actual comparisons at all possible topic counts. The implication is that the validity of experiments carried out with the parameters chosen in the calibration experiments hinges on how well the performance of LDA+Lau-generated labels in comparison to other LDA+Lau-generated labels predicts performance of LDA+Lau-generated labels against other models.

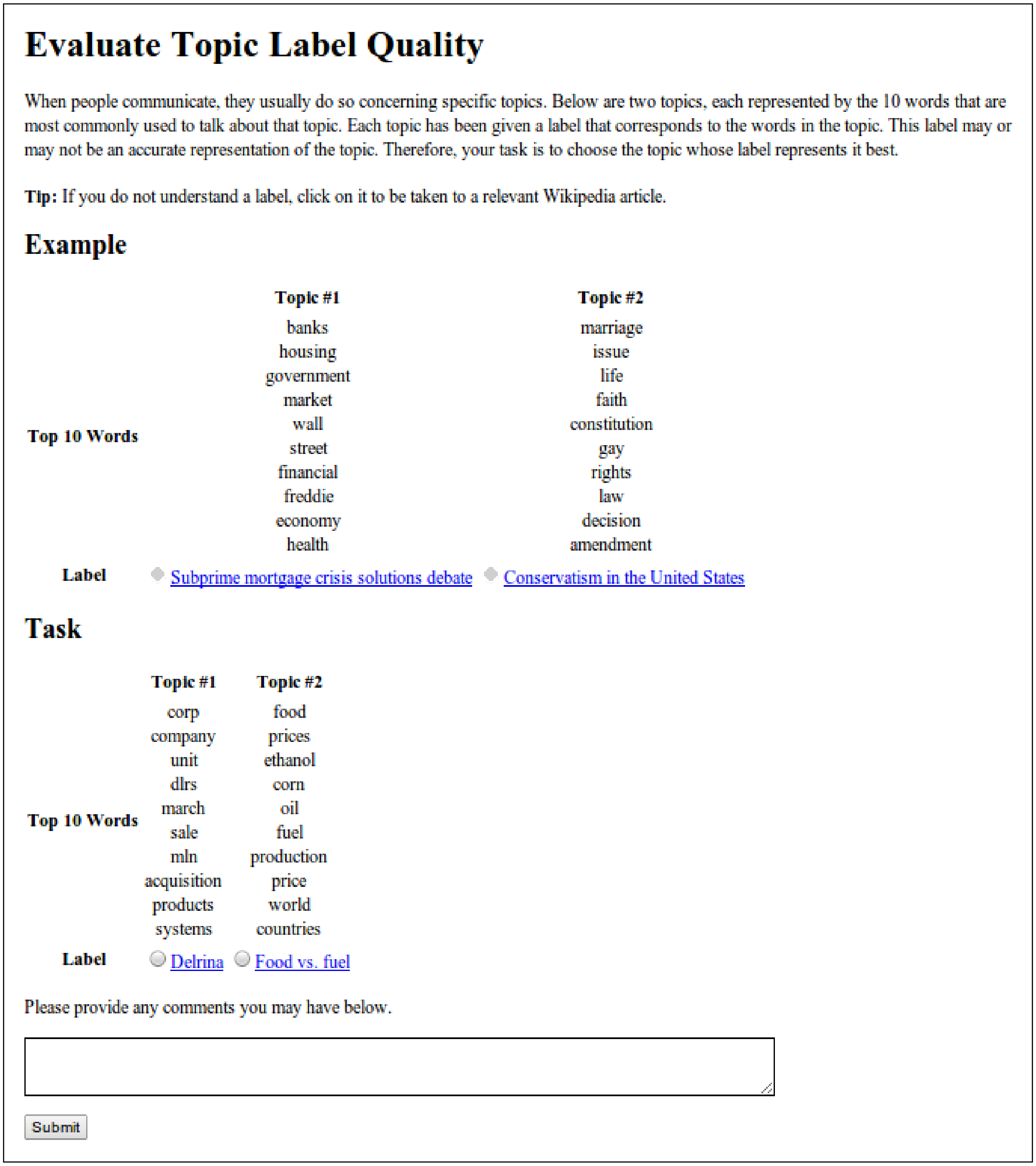
4.5 Topic Label Quality

To assess LDA-STWD's and EDA's ability to find appropriate labeled topics given an input corpus, we compare topic labels from those algorithms to labels generated by LDA+Lau.

4.5.1 Experiment

The $\langle topic, label \rangle$ pairs that LDA-STWD and EDA depend on are not products of the models, but rather are inherent in the Wikipedia topic corpus. Thus any evaluation of topic label quality does not evaluate the quality of these models *per se*, but assesses a critical property of the topic corpus, relative to the quality of $\langle topic, label \rangle$ pairs generated by LDA+Lau.

In the topic label quality task, the user is presented with two $\langle topic, label \rangle$ pairs (one from each model) side-by-side and asked to choose the label that they think best matches



Evaluate Topic Label Quality

When people communicate, they usually do so concerning specific topics. Below are two topics, each represented by the 10 words that are most commonly used to talk about that topic. Each topic has been given a label that corresponds to the words in the topic. This label may or may not be an accurate representation of the topic. Therefore, your task is to choose the topic whose label represents it best.

Tip: If you do not understand a label, click on it to be taken to a relevant Wikipedia article.

Example

	Topic #1	Topic #2
Top 10 Words	banks	marriage
	housing	issue
	government	life
	market	faith
	wall	constitution
	street	gay
	financial	rights
	freddie	law
	economy	decision
	health	amendment
Label	○ Subprime mortgage crisis solutions debate	○ Conservatism in the United States

Task

	Topic #1	Topic #2
Top 10 Words	corp	food
	company	prices
	unit	ethanol
	dlrs	corn
	march	oil
	sale	fuel
	mln	production
	acquisition	price
	products	world
	systems	countries
Label	○ Delrina	○ Food vs. fuel

Please provide any comments you may have below.

Figure 4.4: Topic label quality user study prompt. Participants in the study were asked to choose the topic for which correspondence between the topic's words and its label was greatest.

X	Y	X Preferred	Y Preferred	$\mathbf{BT}(\mathbf{X};\mathbf{Y},\mathbf{0.5})$
Wikipedia	LDA+Lau	201	207	0.805
Wikipedia	Random	64	31	0.000 924 6
LDA+Lau	Random	65	31	0.000 674 7

(a) On SOTU Chunks

X	Y	X Preferred	Y Preferred	$\mathbf{BT}(\mathbf{X};\mathbf{Y},\mathbf{0.5})$
Wikipedia	LDA+Lau	326	103	4.667×10^{-28}
Wikipedia	Random	77	34	5.495×10^{-5}
LDA+Lau	Random	30	30	1.0

(b) On Reuters 21578

Table 4.1: Outcome of topic label quality experiments

the topic with which it is paired. Topics are represented using the top 10 words in the topic by $P(w \mid z)$. Figure 4.4 shows a screenshot of the task as it might appear to a user.

To make $\langle \mathit{topic}, \mathit{label} \rangle$ pairs from the two models more comparable, the topics taken from the topic corpus were restricted to the 100 most frequent topics in an EDA analysis of the relevant target corpus.

4.5.2 Results

While acknowledging the well-known weaknesses of traditional null hypothesis significance tests, [12] we nevertheless employ the standard two-tailed binomial test in our analysis of results due to its presence in standard statistical analysis packages. Results of the topic label quality user study experiments are given in Table 4.1. This shows the number of times users preferred document labels from the Wikipedia topic corpus over those generated by LDA+Lau. For example, various users were simultaneously shown a $\langle \mathit{topic}, \mathit{label} \rangle$ pair from the Wikipedia topic corpus and a $\langle \mathit{topic}, \mathit{label} \rangle$ pair generated by LDA+Lau total of 408 times, preferring the Wikipedia label 201 times and preferring the label generated by LDA+Lau 207 times. The last column in the table shows the outcome of a two-tailed binomial test—the probability that these results would be produced if users were equally likely to choose one label or the other ($p = 0.5$, i.e. the null hypothesis). The second and third rows show how each option fared against randomly generated labels—a "sanity check" against a naive baseline.

In the case of SOTU Chunks, $BT(201; 207, 0.5) \approx 0.805$, leaving no room to conclude any statisticaly significant difference between the two options.

Reuters 21578, on the other hand, displays a clear distinction: $BT(326; 103, 0.5) \approx 4.667 \times 10^{-28}$, meaning that we reject the null hypothesis with confidence at the $9.334 \times 10^{-26}\,\%$ level, concluding that participants clearly preferred the Wikipedia labels. The integrity of this result is somewhat compromised by the presence of many confusing abbreviated terms in the Reuters 21578 corpus (e.g. "mln", "pct", and "bln") potentially prejudicing the annotators in favor of the Wikipedia labels. However, this can simply be understood as confirmation that the topics and labels used by LDA-STWD or EDA are more easily interpreted by humans.

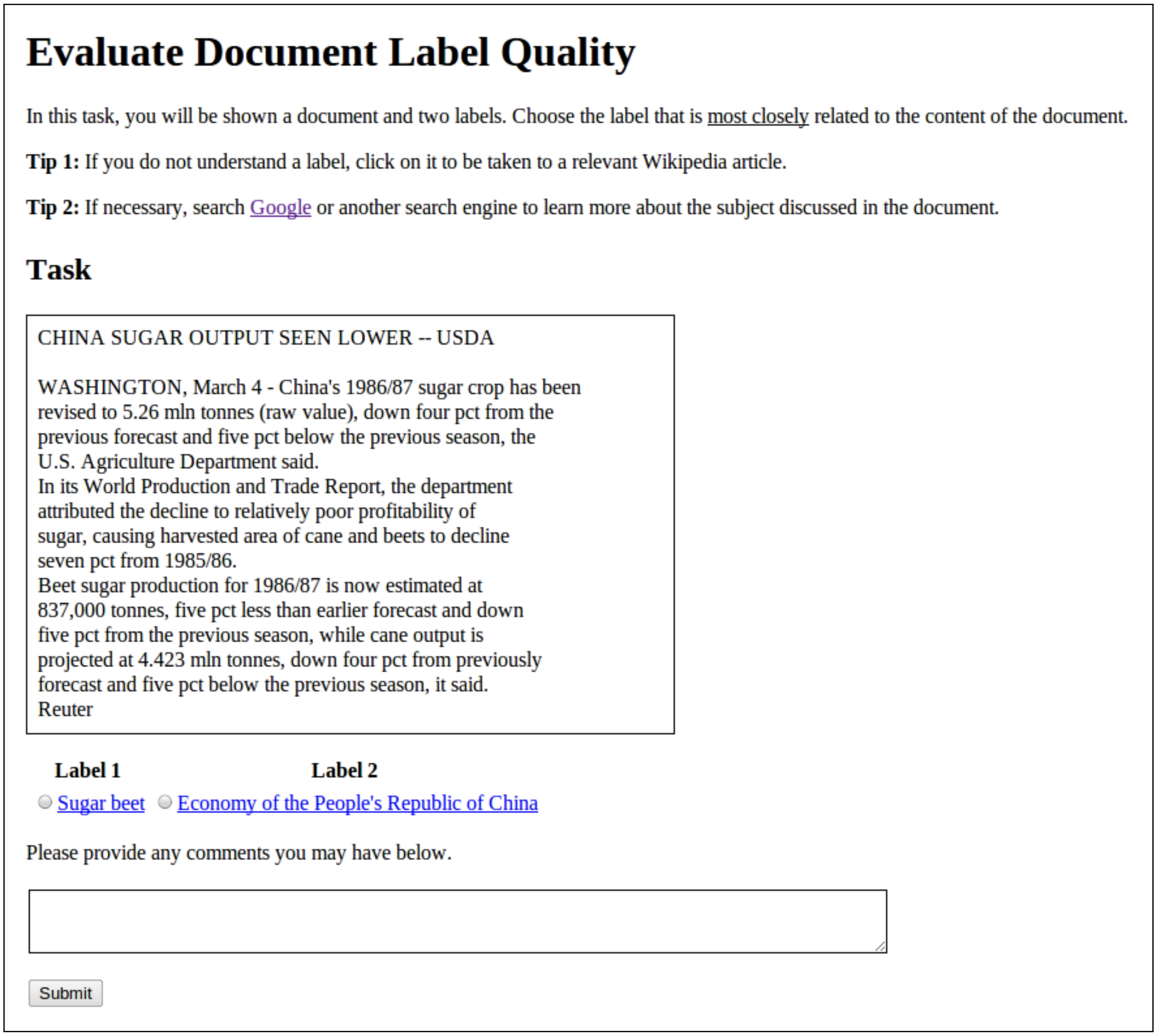
4.6 Document Label Quality

In this section we describe experiments conducted to assess LDA-STWD and EDA in terms of performance on a document labeling task.

4.6.1 Experiments

The user is presented with a randomly selected document from the target corpus. For each of the two topic models, the user is shown the labels of the top 10 topics in the document by $P\left(z \mid \theta\right)$ and asked to choose which of the two sets of labels best matches the content of the document.

In the document label quality task, participants were shown a short document and two possible labels for the document, one from LDA-STWD or EDA and one from LDA+Lau. They were then asked to choose which label best fit the document. The prompt seen by participants is shown in fig. 4.5. To account for positional bias, position (left or right) on screen was randomized. As an additional sanity check, occasionally one of the labels would



Evaluate Document Label Quality

In this task, you will be shown a document and two labels. Choose the label that is most closely related to the content of the document.

Tip 1: If you do not understand a label, click on it to be taken to a relevant Wikipedia article.

Tip 2: If necessary, search Google or another search engine to learn more about the subject discussed in the document.

Task

CHINA SUGAR OUTPUT SEEN LOWER -- USDA

WASHINGTON, March 4 - China's 1986/87 sugar crop has been revised to 5.26 mln tonnes (raw value), down four pct from the previous forecast and five pct below the previous season, the U.S. Agriculture Department said.
In its World Production and Trade Report, the department attributed the decline to relatively poor profitability of sugar, causing harvested area of cane and beets to decline seven pct from 1985/86.
Beet sugar production for 1986/87 is now estimated at 837,000 tonnes, five pct less than earlier forecast and down five pct from the previous season, while cane output is projected at 4.423 mln tonnes, down four pct from previously forecast and five pct below the previous season, it said.
Reuter

Label 1	Label 2
○ Sugar beet	○ Economy of the People's Republic of China

Please provide any comments you may have below.

Figure 4.5: Document label quality user study prompt. Participants in the study were asked to choose the label that best corresponds to the content of the document.

be replaced with a randomly chosen label. Clearly, if the random labels were to outperform either of the models, then there would likely be something wrong with the experiment.

4.7 LDA-STWD Results

Results of the document label quality user study experiments for Latent Dirichlet Allocation with Static Topic-Word Distributions are given in Table 4.2.

In the case of SOTU Chunks, $BT(256; 160, 0.5) \approx 2.905 \times 10^{-6}$. We can thus firmly

X	**Y**	**X Preferred**	**Y Preferred**	$\mathbf{BT}(\mathbf{X};\mathbf{Y},\mathbf{0.5})$
LDA-STWD	LDA+Lau	256	160	2.905×10^{-6}
LDA-STWD	Random	54	27	0.003 596
LDA+Lau	Random	91	11	7.967×10^{-17}

(a) On SOTU Chunks

X	**Y**	**X Preferred**	**Y Preferred**	$\mathbf{BT}(\mathbf{X};\mathbf{Y},\mathbf{0.5})$
LDA-STWD	LDA+Lau	233	205	0.1970
LDA-STWD	Random	76	20	7.319×10^{-9}
LDA+Lau	Random	52	14	2.822×10^{-6}

(b) On Reuters 21578

Table 4.2: Outcome of document label quality experiments with LDA-STWD

X	**Y**	**X Preferred**	**Y Preferred**	$\mathbf{BT}(\mathbf{X};\mathbf{Y},\mathbf{0.5})$
EDA	LDA+Lau	123	266	3.286×10^{-13}
EDA	Random	82	26	6.141×10^{-8}
LDA+Lau	Random	91	11	7.967×10^{-17}

(a) On SOTU Chunks

X	**Y**	**X Preferred**	**Y Preferred**	$\mathbf{BT}(\mathbf{X};\mathbf{Y},\mathbf{0.5})$
EDA	LDA+Lau	182	247	0.001 97
EDA	Random	60	21	1.694×10^{-5}
LDA+Lau	Random	79	11	7.774×10^{-14}

(b) On Reuters 21578

Table 4.3: Outcome of document label quality experiments with EDA

reject the null hypothesis, leaving us to conclude at the $5.810 \times 10^{-4}\,\%$ level that study participants preferred document labels generated by LDA-STWD over those generated by LDA+Lau on this dataset.

The results for Reuters 21578 are inconclusive: $BT(233; 205, 0.5) \approx 0.1970$, meaning the data are sufficiently well explained by the null model that there is not sufficient justification to conclude that either algorithm outperforms the other on this dataset.

4.8 EDA Results

Results of the document label quality user study experiments for Explicit Dirichlet Allocation are given in Table 4.3, revealing that EDA fared much worse than LDA-STWD. In the case of SOTU Chunks, $BT(123; 266, 0.5) \approx 3.286 \times 10^{-13}$, indicating that study participants preferred document labels generated by LDA+Lau over those generated by EDA on this dataset, with significance at the $6.571 \times 10^{-11}\,\%$ level.

Likewise on Reuters 21578, $BT(182; 247, 0.5) \approx 0.001\,97$, meaning that labels generated by LDA+Lau were preferred to those generated by EDA on this dataset, with significance at the $0.393\,\%$ level.

4.9 Topic Quality

We evaluate the quality of our algorithms' topics relative to ESA's by treating each algorithm as a dimensionality reduction method on the Reuters 21578 dataset, and evaluating the performance of an SVM classifier on the reduced data. Presumably the topic model whose output produces the best performance on the text classification task will be the same one that best reflects the essential meaning of the documents [28, 29].

Version 3.12 of the popular LIBSVM library was used for the evaluation.

4.9.1 Comparison to ESA

It is important to compare the quality of our models' output to the baseline of ESA, a non-probabilistic, explicit topic model. Though a number of implementations of ESA were already available, we chose to reimplement the algorithm to allow sharing of data and code with our algorithms' implementations. We had already expended substantial effort generating topic count and type-topic count indexes on Wikipedia article text for use in LDA-STWD and EDA. Implementing ESA ourselves allowed us to make direct use of those existing indexes, freeing us from substantial preprocessing effort.

The comparison of our models to ESA was done by means of a classification task serving as proxy for model quality. Because ESA can produce a score for all prespecified topics given a document, it was necessary to select features to bring ESA output down to a reasonable size. The feature selection algorithm was defined by:

$$Features(N) = \bigcup_{d \in Documents} TopNTopics(d, N) \tag{4.1}$$

$Features(1)$ was used in practice because the number of features selected becomes prohibitively large for higher values of N.

4.9.2 Sample Summarization

Each iteration of the LDA-STWD and EDA Gibbs samplers assigns topics to all tokens in the target corpus. These samples can then be "summarized" in various ways. We summarize by summing document-topic counts across post-burn iterations:

$$\forall_{d\in\{1,\ldots,M\}} \quad \forall_{z\in\{1,\ldots,K\}} \quad \hat{\theta}_{d,z} = \frac{\sum_i C_i\left(z \mid d\right)}{\sum_i \sum_{z'} C_i\left(z' \mid d\right)} \tag{4.2}$$

where $C_i\left(z \mid d\right)$ is the number of tokens in document d assigned to topic z in the ith iteration. This summarization method gives representation to any document-topic pair that occurs in at least one post-burn iteration, but gives greater weight to document-topic pairs that occur in many iterations or with high frequency within iterations. In practice, we only sum across 10 iterations to prevent the counts matrix from becoming impractically dense.

4.9.3 Results

Whereas the document label quality experiments sought to validate LDA-STWD and EDA as document labeling methods compared to an LDA+Lau baseline, we now compare the same algorithms' performance as dimensionality reduction methods on a classification task using the method described in Section 4.9, which shows classification accuracies for the var-

	ESA	LDA-STWD	EDA
Accuracy	80.6	47.3	51.9
		50.2	52.5
		51.3	52.2
		50.5	52.4
		50.2	52.7
Mean		49.9	52.3
Stdev		1.36	0.27

Table 4.4: Results of the topic label quality evaluation, comparing the utility of ESA, LDA-STWD, and EDA respectively as dimensionality reduction methods prior to classifying documents in the Reuters 21578 dataset.

ious algorithms. For LDA-STWD and EDA, . The results are given in Table 4.4. Because ESA is deterministic, classification on that model's output always gave the same result. Clearly ESA has an advantage over its probabilistic counterparts on this task: a one-sample t test comparing ESA's 80.6 to LDA-STWD gives $p = 1.44e-06$, while the comparison to EDA gives $p = 3.25e-09$. ESA's advantage is easily attributable to ESA's fully populated topic relevance output matrix, in contrast to the sparse output of the probabilistic models. More concretely, ESA can be computed and output on a per-topic basis, whereas the probabilistic methods' output must undergo memory-intensive sample summarization, limiting output density on a single core. A better comparison could be achieved using a relatively straightforward map/reduce algorithm on a small cluster, but we consider such work outside of the scope of this thesis.

The winner on this task between the two probabilistic algorithms is EDA, with a standard t test giving $p = 0.00789$. No obvious reason for the distinction presents itself, and it stands in interesting contrast to the superiority of LDA-STWD in the document labeling task.

Chapter 5

Conclusions and Future Work

This thesis introduces two methods for probabilistic explicit topic modeling that address specific weaknesses of LDA, namely the lack of useful topic labels and the inter-run identifiability problem. LDA-STWD does this by directly substituting precomputed counts for LDA topic-word counts, leveraging existing Gibbs sampler inference. EDA defines an entirely new explicit topic model and derives the inference method from first principles. Both of these methods approximate topic-word distributions *a priori* using word distributions from Wikipedia articles, with each article corresponding to one topic and the article title being used as a topic label.

LDA-STWD significantly outperforms LDA+Lau in labeling the documents in the SOTU Chunks corpus, and the topic labels derived from Wikipedia are vastly preferred by human annotators over those generated by LDA+Lau for the Reuters 21578 corpus.

A number of non-rejections of the null hypothesis also speak in favor of LDA-STWD and EDA as more-principled peers to LDA+Lau. The lack of significant difference between Wikipedia and LDA+Lau on SOTU Chunks shows the Wikipedia-derived topics performing no worse than the state of the art. Likewise, LDA-STWD is not found to be worse than incumbent LDA+Lau on Reuters 21578, even though it cannot be said to be significantly better. And finally, EDA is not found to perform significantly worse than LDA+Lau on Reuters 21578. Thus the principled, straightforward LDA-STWD and EDA algorithms can be seen performing at the same level of quality as the less-intuitive *post hoc* Lau, et al. method for this task.

The superiority of LDA-STWD to EDA demands explanation. Mathematically, the two models differ solely in the presence or absence of smoothing in the topic-word distributions ($\boldsymbol{\phi}$). This suggests that EDA's lack of smoothing could be harming performance.

LDA-STWD and EDA suffer from the need to maintain a topic count many orders of magnitude higher than usual in order to achieve generality. Typical applications of LDA set K in the hundreds at most, but LDA-STWD and EDA have K in the millions, reflecting the number of articles in Wikipedia. This clearly increases the amount of computation for each sampling iteration proportionally. Multithreaded performance on a single computer was slow but acceptable. But the prespecification of topic-word counts opens up opportunities for further parallelism because all token topic assignments for a given document are independent of all token topic assignments in other documents given the owning document's document-topic vector ($\boldsymbol{\theta}$).

The research in this thesis could be extended in a number of ways. One would be to implement distributed, parallelized versions of the algorithms introduced in this thesis. Because LDA-STWD's and EDA's topic-word distributions are fixed in advance, the models have independence properties that should allow easier distribution of the sampling workload amongst multiple compute nodes. This should lead to far shorter runtimes, allow more samples to be generated, and thus lead to richer output.

One evaluation of EDA still lacking is proper convergence plots. This is simply a matter of implementing a proper log-likelihood calculation and running it.

Another variation on EDA worth investigating is a model in which topic-word distributions are latent, but explicit topic information is encoded in the priors on those distributions. Such an approach was taken in Mr. LDA, but was not applied to a large number of topics.

Another avenue for future work would be to allow some topics in LDA-STWD or EDA to have latent topic-word distributions. Relative to standard LDA, the probabilistic, explicit topic models introduced in this thesis bias the resultant topics toward subjects of general interest,[1] at the expense of topics of only local interest. For example, a collection of corporate emails may deal with the topic of supply chain management (general interest) as well as the particulars of the corporation's internal politics (local interest). Only the former is likely to have a corresponding Wikipedia-derived topic, meaning the latter will be improperly accounted for by LDA-STWD and EDA. Such confusion would have at least two negative consequences: first, the blocking of the true local interest topic from model output; second, potential degradation of the quality of the topics chosen as they are forced to account for the unrecognized local interest topic.

Additional allowance must be made for local interest topics lest they be conflated with the *a priori* topics. A simple solution may be to treat a certain set of topics $\mathcal{L} \subseteq \mathcal{T}$ as latent. This can be achieved by using the original complete conditional (Equation (2.1)) when calculating the probability of topics in $\mathcal{L}$.

[1] http://en.wikipedia.org/wiki/Wikipedia:Notability

Appendix A

Sample Output

We now give listings of the 40 most frequent topics for the various dataset-algorithm combinations evaluated in this thesis. In each table, `rank` gives the topic's position when sorted by token count, `label` is the human-readable label associated with the topic, and `count` is the total number of tokens assigned the corresponding topic across 40 iterations.

A.1 LDA-STWD on SOTU Chunks

rank	label	count
1	Judicial system of the People's Republic of China	601421
2	UK Immigration Service	346536
3	United States Constitution	268730
4	William Fitzwilliam, 4th Earl Fitzwilliam	190242
5	Origins of the American Civil War	142781
6	History of the English fiscal system	141606
7	First Stadtholderless Period	141174
8	History of rent control in England and Wales	139002
9	Human rights in the Democratic Republic of the Congo	133997
10	Historical powers	118912
11	A Program for Monetary Reform	113266
12	Mediation	110380
13	Government procurement in the United States	92626

rank	label	count
14	Capital, Volume I	90736
15	Foreign relations of India	84905
16	South African property law	83392
17	Reconstruction Era of the United States	77556
18	International child abduction in Mexico	71777
19	United States federal budget	68996
20	NSA warrantless surveillance controversy	67340
21	Criticism of the Israeli government	59453
22	2003 in Afghanistan	59239
23	Income tax in the United States	54231
24	International reactions to the 2006 Lebanon War	53998
25	Value-form	53568
26	History of Basilan	52469
27	Health care in the United States	50811
28	Opportunism	49623
29	Industrial and organizational psychology	49604
30	Maritime history of California	48129
31	Presidency of Ulysses S. Grant	48038
32	South African contract law	47433
33	Spanish-Moro Conflict	44803
34	History of United States diplomatic relations by country	43407
35	Critique of Pure Reason	41785
36	History of Eglin Air Force Base	41454
37	Trent Affair	40972
38	Lame duck session	40910
39	Palestine 194	40788

rank	label	count
40	Economic democracy	39463

A.2 LDA-STWD on Reuters 21578

rank	label	count
1	Russian Venture Company	389101
2	Economy of the People's Republic of China	370646
3	Nuclear program of Iran	287149
4	UK Immigration Service	273428
5	2003 in Afghanistan	245129
6	Fortis (finance)	235536
7	Chronology of world oil market events (19702005)	201052
8	Economy of Pakistan	150795
9	History of agriculture in the People's Republic of China	142420
10	A Program for Monetary Reform	133515
11	History of private equity and venture capital	108893
12	Blockade of Germany (19391945)	91701
13	Taxation in the People's Republic of China	87636
14	Pharmaceutical industry in the People's Republic of China	83760
15	Economic history of Portugal	79189
16	Saddam Hussein and al-Qaeda link allegations timeline	76763
17	Social Security debate in the United States	75065
18	Political debates about the United States federal budget	74175
19	Economy of Egypt	73797
20	Common Agricultural Policy	72504
21	Foreign trade of the Soviet Union	69154
22	Economic history of Brazil	68915

rank	label	count
23	National broadband plans from around the world	68441
24	International public opinion on the war in Afghanistan	66066
25	Subprime mortgage crisis	64409
26	Ethanol fuel in Brazil	62700
27	South African contract law	62623
28	DoddFrank Wall Street Reform and Consumer Protection Act	61619
29	MLN-29	61018
30	Emergency Economic Stabilization Act of 2008	59947
31	Rate-capping rebellion	59214
32	Automotive industry in India	58932
33	Deficit reduction in the United States	56065
34	Subprime crisis background information	55132
35	Enron scandal	54510
36	History of Germany	54401
37	Monetary policy	54216
38	Convertible bond	52986
39	Value-form	51883
40	Inflation	50900

A.3 EDA on SOTU Chunks

rank	label	count
1	UK Immigration Service	56475
2	Judicial system of the People's Republic of China	43070
3	Ages of consent in North America	37256
4	South African contract law	35168
5	Government procurement in the United States	31506

rank	label	count
6	Spanish-Moro Conflict	29330
7	William Fitzwilliam, 4th Earl Fitzwilliam	28084
8	The Idler (17581760)	23048
9	Nuclear program of Iran	22222
10	Mediation	21334
11	United States federal budget	19612
12	English contract law	18485
13	History of Basilan	17441
14	First Stadtholderless Period	16884
15	Taxation in the People's Republic of China	16793
16	Trent Affair	16791
17	International public opinion on the war in Afghanistan	16066
18	Blockade of Germany (19391945)	15703
19	History of United States diplomatic relations by country	15601
20	Social Security (United States)	15446
21	A Program for Monetary Reform	15312
22	History of the English fiscal system	14816
23	Humanitarianism	14361
24	Economic history of the United States	13854
25	History of rent control in England and Wales	13798
26	Gold standard	13292
27	International child abduction in Mexico	12843
28	South African property law	12517
29	Timeline of events leading to the American Civil War	12359
30	Subprime mortgage crisis solutions debate	12244
31	Tariffs in United States history	11851

rank	label	count
32	Capital, Volume I	11805
33	Presidency of Ulysses S. Grant	11380
34	Nullification (U.S. Constitution)	11352
35	Foreign relations of Romania	11326
36	2003 in Afghanistan	11208
37	Nicomachean Ethics	11190
38	Federal Reserve System	11106
39	Harold Wilson	10751
40	Article One of the United States Constitution	10270

A.4 EDA on Reuters 21578

rank	label	count
1	Kaluga Oblast	314706
2	Russian Venture Company	223512
3	Coal Company Zarechnaya	176941
4	Chronology of world oil market events (19702005)	143460
5	Deutsche Bundesbank	88719
6	Federal Reserve System	68958
7	Subprime mortgage crisis	61054
8	1990-1999 world oil market chronology	59478
9	Balance of payments	53290
10	Economy of the People's Republic of China	42438
11	2010 New England Revolution season	38956
12	Fortis (finance)	38400
13	University of Pittsburgh School of Health and Rehabilitation Sciences	38285
14	Soybean	36739

rank	label	count
15	Economic history of Portugal	34965
16	Economy of Honduras	33613
17	Carpal tunnel syndrome	33418
18	Audie Pitre	33409
19	Blockade of Germany (19391945)	31466
20	Clayton Keith Yeutter	29971
21	Economy of Pakistan	29091
22	Nuclear program of Iran	28008
23	MLN-29	27537
24	Conflict tactics scale	26502
25	Foreign trade of the Soviet Union	26356
26	Bretton Woods system	26241
27	Tobin tax	25826
28	Net capital rule	25585
29	Credit default swap	25486
30	History of private equity and venture capital	25111
31	Shearson	25046
32	Social Security (United States)	24957
33	2010 NCAA Division I baseball season	23773
34	US Airways	23676
35	Cadillac CTS	23454
36	A Program for Monetary Reform	22877
37	Occupational therapy in carpal tunnel syndrome	22867
38	DoddFrank Wall Street Reform and Consumer Protection Act	22382
39	History of agriculture in the People's Republic of China	22317
40	Subprime mortgage crisis solutions debate	22077

References

[1] D. M. Blei, “Introduction to probabilistic topic models,” *Communications of the ACM*, 2011. [Online]. Available: `http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf`.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, “Reading tea leaves: How humans interpret topic models,” in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, Neural Information Processing Systems Foundation, 2009.

[4] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab, “Explicit versus latent concept models for cross-language information retrieval,” in *Proceedings of the 21st International Joint Conference on Artifical intelligence*, International Joint Conferences on Artificial Intelligence, 2009, pp. 1513–1518.

[5] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[6] E. Gabrilovich and S. Markovitch, “Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis,” in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence, vol. 6, 2007, p. 12.

[7] M. J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi, “The Topic Browser: An interactive tool for browsing topic models,” in *NIPS Workshop on Challenges of Data Visualization*, Neural Information Processing Systems Foundation, 2010.

[8] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National Academy of Sciences*, vol. 101, suppl. 1, pp. 5228–5235, Jan. 2004. DOI: `10.1073/pnas.0307752101`. [Online]. Available: `http://www.pnas.org/content/101/suppl.1/5228.abstract`.

[9] T. Hofmann, “Unsupervised learning by Probabilistic Latent Semantic Analysis,” *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.

[10] T. Hofmann and J. Puzicha, “Unsupervised learning from dyadic data,” International Computer Science Institute, Tech. Rep., Dec. 1998. [Online]. Available: `ftp://ftp.icsi.berkeley.edu/pub/techreports/1998/tr-98-042.pdf`.

[11] Y. Hu, J. Boyd-Graber, and B. Satinoff, “Interactive topic modeling,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, vol. 1, 2011, pp. 248–257.

[12] J. Kruschke, “Bayesian estimation supersedes the t test,” *Journal of Experimental Psychology: General*, 2012. DOI: `10.1037/a0029146`.

[13] M. de Kunder, *The size of the world wide web (the internet)*, Website, 2012. [Online]. Available: `http://www.worldwidewebsize.com/` (visited on 08/28/2012).

[14] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, “Automatic labelling of topic models,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, vol. 1, Los Angeles, California, 2011, pp. 1536–1545.

[15] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin, "Best topic word selection for topic labelling," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10, Beijing, China: Association for Computational Linguistics, 2010, pp. 605–613. [Online]. Available: `http://dl.acm.org/citation.cfm?id=1944566.1944635`.

[16] D. D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," in *Proceedings of the 15th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, ser. SIGIR '92, Copenhagen, Denmark: Association for Computing Machinery, 1992, pp. 37–50, ISBN: 0-89791-523-2. DOI: `10.1145/133160.133172`. [Online]. Available: `http://doi.acm.org/10.1145/133160.133172`.

[17] D. Magatti, S. Calegari, D. Ciucci, and F. Stella, "Automatic labeling of topics," in *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, IEEE, 2009, pp. 1227–1232.

[18] Q. Mei, X. Shen, and C. X. Zhai, "Automatic labeling of multinomial topic models," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2007, pp. 490–499.

[19] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, Association for Uncertainty in Artificial Intelligence, 2002, pp. 352–359.

[20] D. Ramage, D. Hall, R. Nallapati, and C. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, vol. 1, 2009, pp. 248–256.

[21] J. E. Short, R. E. Bohn, and C. Baru, “How much information?,” Global Information Industry Center, Tech. Rep., Dec. 2010. [Online]. Available: http://hmi.ucsd.edu/pdf/HMI_2010_EnterpriseReport_Jan_2011.pdf.

[22] R. Snow, B. O’Connor, D. Jurafsky, and A. Ng, “Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2008, pp. 254–263.

[23] V. I. Spitkovsky and A. X. Chang, “A cross-lingual dictionary for english Wikipedia concepts,” in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey: European Language Resources Association (ELRA), May 23–25, 2012, ISBN: 978-2-9517408-7-7.

[24] V. Stodden, R. LeVeque, and I. Mitchell, “Reproducible research for scientific computing: tools and strategies for changing the culture,” *Computing in Science Engineering*, vol. 14, no. 4, pp. 13–17, Jul. 2012, ISSN: 1521-9615. DOI: 10.1109/MCSE.2012.38.

[25] L. Taycher, *Books of the world, stand up and be counted! all 129,864,880 of you.* Website. [Online]. Available: http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html (visited on 08/05/2010).

[26] @twitter, *Twitter turns six*, Blog post. [Online]. Available: http://blog.twitter.com/2012/03/twitter-turns-six.html (visited on 03/21/2012).

[27] Various, *List of Wikipedias*, Wiki page. [Online]. Available: http://meta.wikimedia.org/w/index.php?title=List_of_Wikipedias&oldid=4072100 (visited on 08/28/2012).

[28] D. Walker, W. Lund, and E. Ringger, “Evaluating models of latent document semantics in the presence of OCR errors,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2010, pp. 240–250.

[29] D. Walker, E. Ringger, and K. Seppi, “Evaluating supervised topic models in the presence of OCR errors,” in *IS&T/SPIE Electronic Imaging*, International Society for Optics and Photonics, 2013, pp. 865 812–865 812.

[30] YouTube, *Press statistics*, Website. [Online]. Available: `http://www.youtube.com/t/press_statistics` (visited on 08/28/2012).

[31] K. Zhai, J. Boyd-Graber, N. Asadi, and M. Alkhouja, “Mr. LDA: A flexible large scale topic modeling package using variational inference in map/reduce,” in *Proceedings of the 21st International Conference on World Wide Web*, Association for Computational Linguistics, Lyon, France, 2012.