Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce


Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.
Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce

Ke Zhai
Computer Science
University of Maryland
College Park, MD, USA
zhaike@cs.umd.edu

Jordan Boyd-Graber
iSchool and UMIACS
University of Maryland
College Park, MD, USA
jbg@umiacs.umd.edu

Nima Asadi
Computer Science
University of Maryland
College Park, MD, USA
nima@cs.umd.edu

Mohamad Alkhouja
iSchool
University of Maryland
College Park, MD, USA
khouja@umd.edu

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Clustering—topic models, scalability, mapreduce

Abstract
Latent Dirichlet Allocation (LDA) is a popular topic modeling technique for exploring document collections. Because of the increasing prevalence of large datasets, there is a need to improve the scalability of inference for LDA. In this paper, we introduce a novel and flexible large scale topic modeling package in MapReduce (Mr. LDA). As opposed to other techniques which use Gibbs sampling, our proposed framework uses variational inference, which easily fits into a distributed environment. More importantly, this variational implementation, unlike highly tuned and specialized implementations based on Gibbs sampling, is easily extensible. We demonstrate two extensions of the models possible with this scalable framework: informed priors to guide topic discovery and extracting topics from a multilingual corpus. We compare the scalability of Mr. LDA against Mahout, an existing large scale topic modeling package. Mr. LDA out-performs Mahout both in execution speed and held-out likelihood.

In addition to being noisy, data from the web are big. The MapReduce framework for large-scale data processing [8] is simple to learn but flexible enough to be broadly applicable. Designed at Google and open-sourced by Yahoo, Hadoop MapReduce is one of the mainstays of industrial data processing and has also been gaining traction for problems of interest to the academic community such as machine translation [9], language modeling [10], and grammar induction [11].

In this paper, we propose a parallelized LDA algorithm in the MapReduce programming framework (Mr. LDA). Mr. LDA relies on variational inference, as opposed to the prevailing trend of using Gibbs sampling. We argue for using variational inference in Section 2. Section 3 describes how variational inference fits naturally into the MapReduce framework. In Section 4, we discuss two specific extensions of LDA to demonstrate the flexibility of the proposed framework. These are an informed prior to guide topic discovery and a new inference technique for discovering topics in multilingual corpora [12]. Next, we evaluate Mr. LDA’s ability to scale in Section 5 before concluding with Section 6.

1. INTRODUCTION

Because data from the web are big and noisy, algorithms that process large document collections cannot solely depend on human annotations. One popular technique for navigating large unannotated document collections is topic modeling, which discovers the themes that permeate a corpus. Topic modeling is exemplified by Latent Dirichlet Allocation (LDA), a generative model for document-centric corpora [1]. It is appealing for noisy data because it requires no annotation and discovers, without any supervision, the thematic trends in a corpus. In addition to discovering which topics exist in a corpus, LDA also associates documents with these topics, revealing trends in a corpus. In addition to discovering which topics exist in a corpus, LDA also associates documents with these topics, revealing trends in a corpus.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.

ACM 978-1-4503-1229-5/12/04.

In practice, probabilistic models work by maximizing the log-likelihood of observed data given the structure of an assumed probabilistic model. Less technically, generative models tell a story of how your data came to be with some pieces of the story missing; inference fills in the missing pieces with the best explanation of the missing variables. Because exact inference is often intractable (as it is for LDA), complex models require approximate inference.

2. SCALING OUT LDA

In practice, probabilistic models work by maximizing the log-likelihood of observed data given the structure of an assumed probabilistic model. Less technically, generative models tell a story of how your data came to be with some pieces of the story missing; inference fills in the missing pieces with the best explanation of the missing variables. Because exact inference is often intractable (as it is for LDA), complex models require approximate inference.

2.1 Why not Gibbs Sampling?

One of the most widely used approximate inference techniques for such models is Markov chain Monte Carlo (MCMC) sampling, where one samples from a Markov chain whose stationary distribution is the posterior of interest [20, 21]. Gibbs sampling, where the Markov chain is defined by the conditional distribution of each latent variable, has found widespread use in Bayesian models [20, 22, 23, 24]. MCMC is a powerful methodology, but it has drawbacks. Convergence of the sampler to its stationary distribution is difficult to diagnose, and sampling algorithms can be slow to converge in high dimensional models [21].

1Download the code at http://mrlda.cc.
Blei, Ng, and Jordan presented the first approximate inference technique for LDA based on variational methods [1], but the collapsed Gibbs sampler proposed by Griffiths and Steyvers [23] has been more popular in the community because it is easier to implement. However, such methods inevitably have intrinsic problems that lead to difficulties in moving to web-scale: shared state, randomness, too many short iterations, and lack of flexibility.

**Shared State.**

Unless the probabilistic model allows for discrete segments to be statistically independent of each other, it is difficult to conduct inference in parallel. However, we want models that allow specialization to be shared across many different corpora and documents when necessary, so we typically cannot assume this independence.

At the risk of oversimplifying, collapsed Gibbs sampling for LDA is essentially multiplying the number of occurrences of a topic in a document by the number of times a word type appears in a topic across all documents. The former is a document-specific count, but the latter is shared across the entire corpus. For techniques that scale out collapsed Gibbs sampling for LDA, the major challenge is keeping these second counts for collapsed Gibbs sampling consistent when there is not a shared memory environment.

Newman et al. [25] consider a variety of methods to achieve consistent counts: creating hierarchical models to view each slice as independent or simply syncing counts in a batch update. Yan et al. [14] first cleverly partition the data using integer programming (an NP-Hard problem). Wang et al. [17] use message passing to ensure that different slices maintain consistent counts. Smola and Narayanamurthy [18] use a distributed memory system to achieve consistent counts in LDA, and Ahmed et al. [26] extend the approach specifically for LDA, which restricts extensions and enhancements, unless the probabilistic model allows for discrete segments to be statistically independent of each other, it is difficult to conduct inference in parallel. However, we want models that allow specialization to be shared across many different corpora and documents when necessary, so we typically cannot assume this independence.

At the risk of oversimplifying, collapsed Gibbs sampling for LDA is essentially multiplying the number of occurrences of a topic in a document by the number of times a word type appears in a topic across all documents. The former is a document-specific count, but the latter is shared across the entire corpus. For techniques that scale out collapsed Gibbs sampling for LDA, the major challenge is keeping these second counts for collapsed Gibbs sampling consistent when there is not a shared memory environment.

Newman et al. [25] consider a variety of methods to achieve consistent counts: creating hierarchical models to view each slice as independent or simply syncing counts in a batch update. Yan et al. [14] first cleverly partition the data using integer programming (an NP-Hard problem). Wang et al. [17] use message passing to ensure that different slices maintain consistent counts. Smola and Narayanamurthy [18] use a distributed memory system to achieve consistent counts in LDA, and Ahmed et al. [26] extend the approach specifically for LDA, which restricts extensions and enhancements, unless the probabilistic model allows for discrete segments to be statistically independent of each other, it is difficult to conduct inference in parallel. However, we want models that allow specialization to be shared across many different corpora and documents when necessary, so we typically cannot assume this independence.

At the risk of oversimplifying, collapsed Gibbs sampling for LDA is essentially multiplying the number of occurrences of a topic in a document by the number of times a word type appears in a topic across all documents. The former is a document-specific count, but the latter is shared across the entire corpus. For techniques that scale out collapsed Gibbs sampling for LDA, the major challenge is keeping these second counts for collapsed Gibbs sampling consistent when there is not a shared memory environment.

Newman et al. [25] consider a variety of methods to achieve consistent counts: creating hierarchical models to view each slice as independent or simply syncing counts in a batch update. Yan et al. [14] first cleverly partition the data using integer programming (an NP-Hard problem). Wang et al. [17] use message passing to ensure that different slices maintain consistent counts. Smola and Narayanamurthy [18] use a distributed memory system to achieve consistent counts in LDA, and Ahmed et al. [26] extend the approach specifically for LDA, which restricts extensions and enhancements, unless the probabilistic model allows for discrete segments to be statistically independent of each other, it is difficult to conduct inference in parallel. However, we want models that allow specialization to be shared across many different corpora and documents when necessary, so we typically cannot assume this independence.
Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.

Variational methods enjoy clear convergence criterion, tend to be faster than MCMC in high-dimensional problems, and provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

With a variational method, we begin by positing a family of distributions $q \in Q$ over the same latent variables $Z$ with a simpler, fully factorized distribution with variational parameters $\gamma, \lambda,$ and $\phi$. The lack of inter-document dependencies in the variational distribution allows the parallelization of inference in MapReduce.
Algorithm 1 Mapper

Input:
KEY - document ID \(d \in [1, C]\), where \(C = |C|\).
VALUE - document content.

Configure
1: Load in \(\alpha\)'s, \(\lambda\)'s and \(\gamma\)'s from distributed cache.
2: Normalize \(\lambda\)'s for every topic.

Map
1: Initialize a zero \(V \times K\)-dimensional matrix \(\phi\).
2: Initialize a zero \(K\)-dimensional row vector \(\sigma\).
3: Read in document content \([w_1, w_2, \ldots, w_V]\)
4: repeat
5: for all \(v \in [1, V]\) do
6: for all \(k \in [1, K]\) do
7: Update \(\phi_{v,k} = \frac{\lambda_{v,k}}{\sum_k \lambda_{v,k}} \cdot \exp \left(\gamma_{d,k}\right)\).
8: end for
9: Normalize \(\phi\), set \(\sigma = \sum_v \phi_{v,*}\).
10: end for
11: Update row vector \(\gamma_{d,*} = \alpha + \sigma\).
12: until convergence
13: for all \(k \in [1, K]\) do
14: for all \(v \in [1, V]\) do
15: Emit \((k, v) : w_v \phi_{v,k}\).
16: end for
17: Emit \((\Delta, k) : \left(\psi_1(\gamma_{d,k}) - \psi_2\left(\sum_{i=1}^K \gamma_{d,i}\right)\right).\) \(\text{[Section 3.4]}\)
18: Emit \((k, d) : \gamma_{d,k}\) to file.
19: end for
20: Aggregate \(\mathcal{L}\) to global counter. \(\text{[ELBO, Section 3.5]}\)

algorithm: the mapper, which processes a single unit of data (in this case, a document); the reducer, which processes a single view of globally shared data (in this case, a topic parameter); the partitioner, which distributes the workload to reducers; and the driver, which controls the overall algorithm. The interconnections between the components of Mr. LDA are depicted in Figure 2.

3.1 Mapper: Update \(\phi\) and \(\gamma\)

Each document has associated variational parameters \(\gamma\) and \(\phi\). The mapper computes the updates for these variational parameters and uses them to create the sufficient statistics needed to update the global parameters. In this section, we describe the computation of these variational updates and how they are transmitted to the reducers.

Given a document, the updates for \(\phi\) and \(\gamma\) are

\[
\phi_{v,k} \propto \mathbb{E}_{q}[\beta_{v,k}] \cdot e^{q(\gamma_{d,k})}, \quad \gamma_{k} = \alpha_k + \sum_{v=1}^V \phi_{v,k},
\]

where \(v \in [1, V]\) is the term index and \(k \in [1, K]\) is the topic index. In this case, \(V\) is the size of the vocabulary \(V\) and \(K\) denotes the total number of topics. The expectation of \(\beta\) under \(q\) gives an estimate of how compatible a word is with a topic; words highly compatible with a topic will have a larger expected \(\beta\) and thus higher values of \(\phi\) for that topic.

Algorithm 1 illustrates the detailed procedure of the Map function. In the first iteration, mappers initialize variables, e.g. seed \(\lambda\) with the counts of a single document. For the sake of brevity, we omit that step here; in later iterations, global parameters are stored in distributed cache – a synchronized read-only memory that is shared among all mappers [35] – and retrieved prior to mapper execution in a configuration step.

A document is represented as a term frequency sequence \(\vec{w} = [w_1, w_2, \ldots, w_V]\), where \(w_i\) is the corresponding term frequency in document \(d\). For ease of notation, we assume the input term frequency vector \(\vec{w}\) is associated with all the terms in the vocabulary, i.e., if term \(t_i\) does not appear at all in document \(d\), \(w_i = 0\).

Because the document variational parameter \(\gamma\) and the word variational parameter \(\phi\) are tightly coupled, we impose a local convergence requirement on \(\gamma\) in the Map function. This means that the mapper alternates between updating \(\gamma\) and \(\phi\) until \(\gamma\) stops changing.

3.2 Partitioner: Evenly Distribute Workloads

The Map function in Algorithm 1 emits sufficient statistics for updating the topic variational distribution \(\lambda\). These sufficient statistics are keyed by a composite key set \((\text{key}, \text{right})\). These keys can take two forms: tuple of topic and word identifier or, when the value represents the sufficient statistics for \(\alpha\) updating, a unique value \(\Delta\) and a topic identifier.

A partitioner is required to ensure that messages from the mappers are sent to the appropriate reducers. Each reducer is responsible for updating the per-topic variational parameter associated with a single topic indexed by \(k\). This is accomplished by ensuring the partitioner sorts on topic only. A consequence of this is that any reducers beyond the number of topics is superfluous. Given that the vast majority of the work is in the mappers, this is typically not an issue for LDA.

3.3 Reducer: Update \(\lambda\)

The Reduce function updates the variational parameter \(\lambda\) associated with each topic. It requires aggregation over all intermediate \(\phi\) vectors

\[
\lambda_{v,k} = \eta_{v,k} + \sum_{d=1}^C \left(w_{v}(d) \phi_{v,k}^{(d)}\right),
\]

where \(d \in [1, C]\) is the document index and \(w_{v}(d)\) denotes the number of appearances of term \(v\) in document \(d\). Similarly, \(C\) is the number of documents. Although the variational update for \(\lambda\) does not include a normalization, the expectation \(\mathbb{E}_{q}[\beta]\) requires the \(\lambda\) normalization. In Mr. LDA, the \(\lambda_{v,k}\) parameters are distributed to all mappers, and the normalization is taken care of by the mappers in a configuration step prior to every iteration.

To improve performance, we use combiners to facilitate the aggregation of sufficient statistics in mappers before they are transferred to reducers. This decreases bandwidth and saves the reducer computation.

3.4 Driver: Update \(\alpha\)

Effective inference of topic models depends on learning not just the latent variables \(\beta, \theta\), and \(z\) but also estimating the hyperparame-
ters, particularly \( \alpha \). The \( \alpha \) parameter controls the sparsity of topics in the document distribution and is the primary parameter to update in LDA, especially true if we launch a large number of reducers every iteration — this will result in a large number of small outputs, since each output is typically a small number of topics.

Updating hyperparameters is also important from the perspective of equalizing differences between inference techniques; as long as hyperparameters are optimized, there is little difference between the output of inference techniques [36].

The driver program marshals the entire inference process. On the first iteration, the driver is responsible for initializing all the model parameters \((K, V, C, \eta, \alpha)\); the number of topics \( K \) is user specified; \( C \) and \( V \), the number of documents and types, is determined by the data; the initial value of \( \alpha \) is specified by the user; and \( \lambda \) is randomly initialized or otherwise seeded.

The driver updates \( \alpha \) after each MapReduce iteration. We use a Newton-Raphson method which requires the Hessian matrix and the gradient,

\[
\alpha_{\text{new}} = \alpha_{\text{old}} - \mathcal{H}^{-1}(\alpha_{\text{old}}) \cdot g(\alpha_{\text{old}}),
\]

where the Hessian matrix \( \mathcal{H} \) and \( \alpha \) gradient are, respectively, as

\[
\mathcal{H}(k, l) = \delta(k, l) C \Psi'(\alpha_k) - C \Psi'(\sum_{i=1}^{K} \alpha_i),
\]

\[
g(k) = C \left( \Psi\left( \sum_{i=1}^{K} \alpha_i \right) - \Psi\left( \sum_{l=1}^{K} \alpha_l \right) \right),
\]

computed in driver

\[
+ \sum_{d=1}^{C} \Psi\left( \gamma_{d,k} \right) - \Psi\left( \sum_{l=1}^{K} \gamma_{d,l} \right),
\]

computed in mapper

The Hessian matrix \( \mathcal{H} \) depends entirely on the vector \( \alpha \), which changes during updating \( \alpha \). The gradient \( g \), on the other hand, can be decomposed into two terms: the \( \alpha \)-tokens (i.e., \( \Psi\left( \sum_{i=1}^{K} \alpha_i \right) - \Psi\left( \sum_{i=1}^{K} \alpha_i \right) \)) and the \( \gamma \)-tokens (i.e., \( \sum_{d=1}^{C} \Psi\left( \gamma_{d,k} \right) - \Psi\left( \sum_{l=1}^{K} \gamma_{d,l} \right) \)). We can remove the dependence on the number of documents in the gradient computation by computing the \( \gamma \)-tokens in mappers. This observation allows us to optimize \( \alpha \) in the MapReduce environment.

Because LDA is a dimensionality reduction algorithm, there are typically a small number of topics \( K \) even for a large document collection. As a result, we can safely assume the dimensionality of \( \alpha, \mathcal{H}, \) and \( g \) are reasonably low, and additional gains come from the diagonal structure of the Hessian [37]. Hence, the updating of \( \alpha \) is efficient and will not create a bottleneck in the driver.

### 3.5 Likelihood Computation

The driver monitors the ELBO to determine whether inference has converged. If not, it restarts the process with another round of mappers and reducers. To compute the ELBO we expand Equation 1, which gives us

\[
\mathcal{L}(\gamma, \phi, \lambda; \eta) = \sum_{d=1}^{C} \Phi(\alpha) + \sum_{d=1}^{C} \left( \mathcal{L}_d(\gamma, \phi) + \mathcal{L}_d(\phi) - \Phi(\gamma) \right)
\]

computed in driver

\[
+ \sum_{k=1}^{K} \Phi(\eta, k) - \sum_{k=1}^{K} \Phi(\lambda, k)
\]

computed in reducer

where

\[
\Phi(\mu) = \log \Gamma\left( \sum_{i=1}^{K} \mu_i \right) - \sum_{i=1}^{K} \log \Gamma(\mu_i)
\]

\[
+ \sum_{d=1}^{C} \left( \sum_{i=1}^{K} \mu_i \right) \Psi(\mu_i) - \Psi\left( \sum_{i=1}^{K} \mu_i \right),
\]

\[
\mathcal{L}_d(\gamma, \phi) = \sum_{k=1}^{K} \sum_{v=1}^{V} \phi_{v,k} w_{v} \Psi(\gamma_{d,k}) - \Psi\left( \sum_{i=1}^{K} \gamma_{d,i} \right),
\]

\[
\mathcal{L}_d(\phi) = \sum_{v=1}^{V} \sum_{k=1}^{K} \phi_{v,k} \left( \sum_{i=1}^{V} \log \sum_{j,k} \lambda_{j,k} - \log \phi_{v,k} \right),
\]

Almost all of the terms that appear in the likelihood term can be computed in mappers; the only term that cannot are the terms that depend on \( \alpha \), which is updated in the driver, and the variational parameter \( \lambda \), which is shared among all documents. All terms that depend on \( \alpha \) can be easily computed in the driver, while the terms that depend on \( \lambda \) can be computed in each reducer.

Thus, computing the total likelihood proceeds as follows: each mapper computes its contribution to the likelihood bound \( \mathcal{L} \) and emits a special key that is unique to likelihood bound terms and then aggregated in the reducer; the reducers add topic-specific terms to the likelihood; these final values are then combined with the contribution from \( \alpha \) in the driver to compute a final likelihood bound.

### 3.6 Structural Optimization

In examining Mr. LDA’s performance, the two largest performance limitations were the large number of intermediate values being generated by the mappers and the time it takes for mappers to read in the current variational parameters during the mapper configuration phase.

**Reducer Caching.**

Recall that reducers sum over \( \phi \) contributions and emit the \( \lambda \) variational parameters, but mappers require a normalized form to compute the expectation with of the topic with respect to the variational distribution. To improve the normalization step, we compute the sum of the \( \lambda \) variational parameters in the reducer [38, 39], and then emit this sum before we emit the other \( \lambda \) terms.

Although this requires \( O(V) \) additional memory, it is strictly less than the memory required by mappers, so it in practice improves performance by allowing mappers to more quickly begin processing data.

**File Merge.**

Loading files in the distributed cache and configuring every mapper and reducer is another bottleneck for this framework. This is especially true if we launch a large number of reducers every iteration — this will result in a large number of small outputs, since
Mr. LDA is designed to distribute workload equally. These partial results would waste space if they are significantly smaller than HDFS block size. Moreover, they cause a overhead in file transfer through distributed cache. To alleviate this problem, we merge all relevant output before sending them to distributed cache for the next iteration.

4. FLEXIBILITY OF MR. LDA

In this section, we highlight the flexibility of Mr. LDA to accommodate extensions to LDA. These extensions are possible because of the modular nature of Mr. LDA’s design.

4.1 Informed Prior

The standard practice in topic modeling is to use a same symmetric prior (i.e., $\eta_{v,k}$ is the same for all topics $k$ and words $v$). However, the model and inference presented in Section 3 allows for topics to have different priors. Thus, users can incorporate prior information into the model.

For example, suppose we wanted to discover how different psychological states were expressed in blogs or newspapers. If this were our goal, we might create priors that captured psychological categories to discover how they were expressed in a corpus. The Linguistic Inquiry and Word Count (LIWC) dictionary [40] defines 68 categories encompassing psychological constructs and personal concerns. For example, the anger LIWC category includes the words “abuse,” “jerk,” and “jealous;” the anxiety category includes “afraid,” “alarm,” and “avoid;” and the negative emotions category includes “abandon,” “maddening,” and “sob.” Using this dictionary, we built a prior $\eta$ as follows:

$$
\eta_{v,k} = \begin{cases} 
10, & \text{if } v \in \text{LIWC category}_k \\
0.01, & \text{otherwise}
\end{cases}
$$

where $\eta_{v,k}$ is the informed prior for word $v$ of topic $k$. This is accomplished via a slight modification of the reducer (i.e. to make it aware of the values of $\eta$) and leaving the rest of the system unchanged.

4.2 Polylingual LDA

In this section, we demonstrate the flexibility of Mr. LDA by showing how its modular design allows for extending LDA beyond a single language. PolyLDA [12] assumes a document-aligned multilingual corpus. For example, articles in Wikipedia have links to the version of the article in other languages; while the linked documents are ostensibly on the same subject, they are usually not direct translations, and are often written with a culture-specific focus.

PolyLDA assumes that a single document has words in multiple languages, but each document has a common, language agnostic per-document distribution $\theta$ (Figure 3). Each topic also has different facets for language; these topics end up being consistent because of the links across language encoded in the consistent themes present in documents.

Because of the modular way in which we implemented inference, we can perform multilingual inference by embellishing each data unit with a language identifier $l$ and change inference as follows:

- Updating $\lambda$ happens $l$ times, once for each language. The updates for a particular language ignores expected counts of all other languages.
- Updating $\phi$ happens using only the relevant language for a word.
- Updating $\gamma$ happens as usual, combining the contributions of all languages relevant for a document.

From an implementation perspective, PolyLDA is a collection of monolingual Mr. LDA computations sequenced appropriately. Mr. LDA’s approach of taking relatively simple computation units, allowing them to scale, and preserving simple communication between computation units stands in contrast to the design choices made by approaches using Gibbs sampling.

For example, Smola and Narayananmuthy [18] interleave the topic and document counts during the computation of the conditional distribution using Yao et al.’s “binning” approach [27]. While this improves performance, changing any of the modeling assumptions would potentially break this optimization.

In contrast, Mr. LDA’s philosophy allows for easier development of extensions of LDA. While we only discuss two extensions here,
other extensions are possible. For example, implementing supervised LDA [41] only requires changing the computation of $\phi$ and a regression; the rest of the model is unchanged. Implementing syntactic topic models [42] requires changing the mapper to incorporate syntactic dependencies.

5. EXPERIMENTS

We implemented Mr. LDA using Java with Hadoop 0.20.1 and ran all experiments on a cluster containing 16 physical nodes; each node has 16 2.4GHz cores, and has been configured to run a maximum of 6 map and 3 reduce tasks simultaneously. The cluster is usually under a heavy, heterogeneous load. In this section, we document the speed and likelihood comparison of Mr. LDA against Mahout LDA, another large scale topic modeling implementation based on variational inference. We report results on three datasets:

- TREC document collection (disks 4 and 5 [43]), newswire documents from the Financial Times and LA Times. It contains more than 300k distinct types over half a million documents. We remove types appearing fewer than 20 times, reducing the vocabulary size to approximately 60k.
- The BlogAuthorship corpus [44], which contains about 10 million blog posts from American users. In contrast to the newswire-heavy TREC corpus, the BlogAuthorship corpus is more personal and informal. Again, terms in fewer than 20 documents are excluded, resulting in 53k distinct types.
- Paired English and German Wikipedia articles (more than half a million in each language). As before, we ignore terms appearing in fewer than 20 documents, resulting in 170k English word types and 210k German word types. While each pair of linked documents shares a common subject (e.g. “George Washington”), they are usually not direct translations. The document pair mappings were established from Wikipedia’s interlingual links.

5.1 Informed Priors

In this experiment, we build the informed priors from LIWC [40] introduced in Section 4.1. We feed the same informed prior to both the TREC dataset and BlogAuthorship corpus. Throughout the experiments, we set the number of topics to 100, with a subset guided by the informed prior.

Table 2 shows topics for both TREC and BlogAuthorship. The prior acts as a seed, causing words used in similar contexts to become part of the topic. This is important for computational social scientists who want to discover how an abstract idea (represented by a set of words) is actually expressed in a corpus. For example, public news media (i.e. news articles like TREC) connect positive emotions to entertainment, such as music, film and TV, whereas social media (i.e. blog posts) connect it to religion. The Anxiety topic in news relates to middle east, but in blogs it focuses on illness, whereas social media (i.e. blog posts) connect it to religion. The Anxiety topic in news relates to middle east, but in blogs it focuses on illness.

Using informed priors can discover radically different words. While LIWC is designed for relatively formal writing, it can also discover Internet slang such as “lol” (“laugh out loud”) in Affective Process category. As a result, an informed prior might be helpful in aligning existing lexical resources with corpora with sparse and/or out-of-dictionary vocabularies, e.g., Twitter data.

On the other hand, some discovered topics do not have a clear relationship with the initial LIWC categories, such as the abbreviations and acronyms in Discrepancy category. In other cases, the LIWC categories were different enough from the dataset that model chose not to use topics with ill-fitting priors, e.g. the Cognitive Process category.

5.2 Polylingual LDA

As discussed in Section 4.2, Mr. LDA’s modular design allows us to consider models beyond vanilla LDA. To the best of our knowledge, we believe this is the first framework for variational inference for polylingual LDA [12], scalable or otherwise. In this experiment, we fit 50 topics to paired English and German Wikipedia articles. We let the program run for 35 iterations with 100 mappers and 50 reducers. Table 3 lists down some words from a set of randomly chosen topics.

The results listed indicates a general equivalent topic layout for both English and German corpus. For example, topic about Europe (“french”, “paris”, “russian” and “moscow”) in English is matched with the topic in German (“frankreich”, “paris”, “russischen” and “moskau”). Similar behavior was observed for other topics.

The topics discovered by polylingual LDA are not exact matches, however. For example, the second to last column in Table 3 is about North America, but the English words focus on Canada, while the corresponding German topic focuses on the United States. Similarly, the forth last column in English contains keywords like “hong”, “kong” and “korean”, which did not appear in the top 10 words in German. Since this corpus is not a direct translation, these discrepancies might due to a different perspectives, different editorial styles, or different cultural norms.

5.3 Scalability

To measure the scalability and accuracy of Mr. LDA, we compare Mr. LDA with Mahout [19], another large scale topic modeling package based on variational inference. We use Mahout-0.4 as our baseline measure, with a comparable settings in Mr. LDA — we set the memory limit for every mapper and reducer to 2-GB, and start the hyper-parameter $\alpha$ from 1. Mr. LDA continuously updates vector $\alpha$ in the driver program, whereas Mahout does not. All experiments are carried out with 100 mapper instances and 20 reducer instances.

In these set of experiments, we use 90% of the entire TREC corpus as training data, and the rest as our testing data. We then plot the held-out log-likelihood of the test dataset against the training time. Our empirical results show that, with identical data and hardware, Mr. LDA out-performs Mahout LDA.

Both models had identical input and both models ran for 40 iterations. The held-out likelihood was computed using the variational distribution obtained after every iteration. Figure 4 shows the result for 50 topics. Mr. LDA runs faster than Mahout. In addition, Mr. LDA yields a better held-out likelihood than Mahout, probably as a consequence of hyper-parameter updating.

When we double the total number of topics to 100, the difference in processing time is magnified. Mr. LDA converges faster than Mahout, again due to the hyper-parameter updating.

6. CONCLUSION AND FUTURE WORK

Understanding large text collections such as those generated by social media requires algorithms that are unsupervised and scalable. In this paper, we present Mr. LDA, which fulfills both of these requirements. Beyond text, LDA is continually being applied to new fields such as music [45] and source code [46]. All of these domains struggle with the scale of data, and Mr. LDA could help them better cope with large data.

Mr. LDA represents a viable alternative to the existing scalable mechanisms for inference of topic models. Its design easily accommodates other extensions, as we have demonstrated with the
addition of informed priors and multilingual topic modeling, and the ability of variational inference to support non-conjugate distributions allows for the development of a broader class of models than could be built with Gibbs samplers alone. Mr. LDA, however, would benefit from many of the efficient, scalable data structures (“wiener” and “wien”), while the corresponding concept in English does not appear until the

### Table 2: Twelve Topics Discovered from TREC (top) and BlogAuthorship (bottom) collection with LIWC-derived informed prior.

The model associates TREC documents containing words like “arab”, “israel”, “palestinian” and “peace” with Anxiety. In the blog corpus, however, the model associates words like “iraq”, “america”, “militari”, “unit”, and “force” with the Anger category.

### Table 3: Extracted Polylingual Topics from the Wikipedia Corpus. While topics are generally equivalent (e.g. on “computer games”), some regional differences are expressed. For example, the “music” topic in German has two words referring to “Vienna”.

#### 7. ACKNOWLEDGMENTS

The authors would also like to thank Dr. Jimmy Lin for valuable comments throughout this project and making data available for the experiments. This research was supported by NSF grant #1018625. Jordan Boyd-Graber is also supported by the Army Research Laboratory through ARL Cooperative Agreement W911NF-09-2-0072. Comments throughout this project and making data available for the experiments.

#### 8. REFERENCES


Figure 4: Training Time vs. Held-out Likelihood on 50 topics. This figure shows the accumulated training time of the model against the held-out likelihood over Mr. LDA and Mahout. The time and held-out likelihood are measured over 40 iterations with 50 topics. Markers indicate the finishing point of an iteration. Mr. LDA outperforms Mahout both in speed and performance.

Figure 5: Training Time vs. Held-out Likelihood on 100 topics. Similar to Figure 4, this figure shows the accumulated training time of the model against the held-out likelihood for Mr. LDA and Mahout over 40 iterations, but for 100 topics. Markers indicate the finishing point of an iteration.


