News Topic Modeling with Latent Dirichlet Allocation

Alex Jian

April 21, 2023

1 Introduction

In this report, I investigate a Bayesian hierarchical model called Latent Dirichlet Allocation (LDA) and use it to build a topic model for a collection of news articles. News articles are written independently, but they generally share common topics and themes - news can be grouped into categories such as politics, health, and finance, and two distinct articles may cover the the same news event from different perspectives or multiple events from the same developing story. Humans reading news articles can easily identify those that cover the same topic, and Latent Dirichlet Allocation allows programs to approximate the same classifications automatically by analyzing the statistical properties of the observed words. This report will both explain how LDA works and apply it to construct a topic model over a large set of news articles.

2 Latent Dirichlet Allocation

2.1 Overview

Latent Dirichlet Allocation is a generative statistical model which explains a large number of observations through a smaller set of unobserved groups. In the context of machine learning and natural language processing, it is used to discover common topics underlying a collection of documents and then classify each individual document based on which of these topics it contains. In particular, we define a topic as a set of words that collectively suggest a common theme and use LDA to discover these topics based on the repeated appearance of these words. For example, in a news dataset, the words *election, president* or *law* might suggest a politics-related topic, while the words *technology, breakthrough* or *experiment* may suggest a science-related theme. Note that LDA only finds the words that best represent each topic - assigning meaningful labels to these topics requires domain knowledge on the user's part.

2.2 Probabilistic Graph Model

As a generative model, Latent Dirichlet Allocation assumes a statistical process by which documents are generated, then conducts Bayesian inference to learn what the parameters of this statistical process are most likely to be. The parameters of the process define the per-document topic and word distributions, which allow for the construction of the topic model over a full document collection.

The probabilistic graph model below shows the generative, topic-based process by which LDA assumes documents are generated. Before showing the diagram, we define variables as follows:

M is the total number of documents in the collection
N_i is total number of words in document i
α is a vector specifying the Dirichlet prior for the per-document topic distribution
β is a vector specifying the Dirichlet prior for the word distribution per topic
θ_i is the topic distribution for a single document (document i)
ψ_k is the word distribution for a single topic (topic k)
z_{i_j} is the topic referred to by the jth word in document i
w_{i_j} is the jth observed word appearing in document i

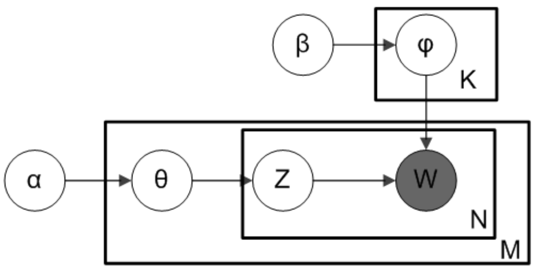

Figure 1:
Graph Model of LDA document generation process

Based on this diagram, we assume that the words for each document in a collection are generated as follows:

1) For each document i, choose $\theta_i \sim Dir(\alpha)$. Intuitively, we can think of this as the author deciding which topics the document will cover.

2) For each topic k, choose $\psi_k \sim Dir(\beta)$. Intuitively, once the topic is chosen,

certain words are much more likely to appear than others, a property specified by this per-topic word distribution.

3) We assume that each word in each document refers to one singular topic. For each word of each document, first choose its topic, $z_{ij} \sim Categorical(\theta_i)$. Then, choose the word from the topic's word distribution, $w_{i_j} \sim Categorical(\psi_{z_{i_j}})$.

Figure 1 illustrates this process at a high level. Note that W is grayed out, emphasizing that the words are the only observable variable, and the rest are latent variables which must be inferred through Bayesian Inference. Since each node in the graph is conditionally dependent only on its immediate predecessors, we can treat LDA as a Hierarchical Model and conduct inference using Bayesian methods. Note that this model assumes that all words within a document can be treated as exchangable, independent and identically distributed random variables. This assumption is largely untrue - words in sentences depend on each other and have context - but for the purposes of topic modeling, this assumption has still led to good results.

2.3 Prior Distributions

In LDA, it is important that the prior distributions accurately reflect our existing knowledge of text documents and their topic and word distributions. In this case, we observe that most documents (especially news articles) are rather focused, and contain only a few topics at most. We can account for this by a constructing sparse symmetric prior ($Dir(\alpha)$, $\alpha < 1$) for the per-document topic distribution, so that while there is no bias against any specific topic, there is bias against more uniform distributions that treat many topics as reasonably likely. The same holds true for the per-topic word distribution ($Dir(\beta)$, $\beta < 1$): for a given topic, certain words are more likely than others, a fact the prior should reflect.

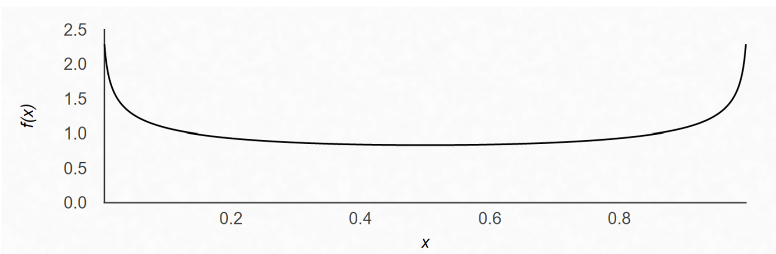

Figure 2:
Sparse Symmetric Prior for 1D Dirichlet Distribution

2.4 Inference

The mathematical problem that LDA tries to solve is computing the posterior distribution of the model's latent variables. We condition on the observed words ($\mathbf{w}$) as well as the α and β Dirichlet priors to compute the posterior per-document topic and topic-word distributions, θ and $\mathbf{z}$:

$$p(\theta, \mathbf{z}|\mathbf{w},\alpha,\beta) = p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta)/p(\mathbf{w}|\alpha,\beta)$$

As is often the case for Bayesian posteriors, it is impractical to analytically compute the denominator this distribution. To account for this, we can use several numerical/computational methods, including MCMC with Gibbs Sampling or Variational Bayes. Since the LDA model is typically trained on large datasets, Variational Bayes is the usual choice due to its advantage in efficiency.

3 Results

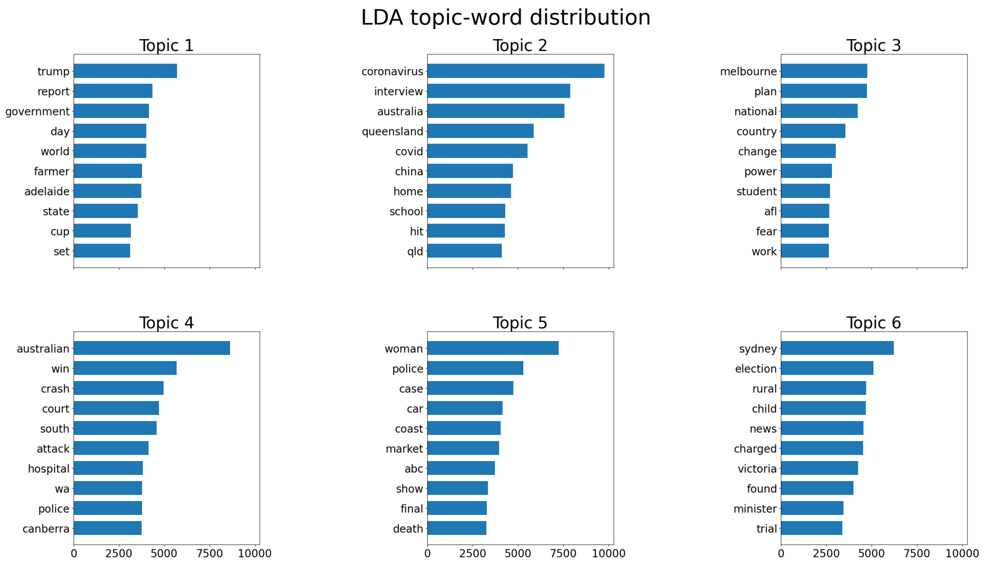

Figure 3:
Posterior Topic-Word Distribution

	publish_date	head_clean	topics
545241	20100620	call national standard driver licence	Topic 2: 0.594
618253	20110605	lake face afl mental hurdle eade	Topic 1: 0.417 Topic 2: 0.201 Topic 3: 0.213
843702	20131230	lebanon buries critic hezbollah	Topic 1: 0.291 Topic 2: 0.266 Topic 3: 0.259
839535	20131206	glory take point free flowing battle	Topic 2: 0.564 Topic 3: 0.213
181384	20050811	boy charged sexual assault	Topic 0: 0.227 Topic 4: 0.405
1225965	20201228	flood rain clermont queensland central quad	Topic 1: 0.490 Topic 5: 0.301
652951	20111116	dairy industry welcome public committment open	Topic 0: 0.340
814817	20130829	bundaberg flood money	Topic 3: 0.513 Topic 5: 0.241
241230	20060605	launceston lead ta population growth	Topic 0: 0.219 Topic 1: 0.233 Topic 3: 0.379
246294	20060629	union heart big turnout protest	Topic 0: 0.319 Topic 2: 0.261 Topic 3: 0.262
1203972	20200624	threatened native specie roam central australia	Topic 1: 0.278 Topic 2: 0.208 Topic 4: 0.357
939127	20150319	india win toss ; bat quarter final bangladesh	Topic 2: 0.214 Topic 3: 0.296 Topic 4: 0.332
1177263	20190925	ghost net northern gulf hotspot risk marine life	Topic 0: 0.243 Topic 2: 0.231 Topic 5: 0.299
862232	20140324	arnold palmer invitational live	Topic 1: 0.310 Topic 3: 0.235 Topic 4: 0.270

Figure 4:
Posterior per-document Topic Distribution, showing topics which contributed significantly to the observed words (>20%)