Latent Dirichlet Allocation (LDA)

CPSC 503 - Pedagogical Project Report

Nazlı Özüm Kafaee
Sauder School of Business, University of British Columbia

April 23, 2019

1 Introduction

Looking at the thematic structure provides a way to explore and digest the content of a group of documents. Humans can accomplish such exploration smoothly. Indeed, researchers in the past mostly relied on manual qualitative analysis to extract information from unstructured textual data. However, the vast amount of digitized knowledge available online has encouraged the development of new computational tools that can extract the thematic structure of documents automatically, which enable researchers to produce the same knowledge while saving significantly in time and cost. Specifically for such purpose, machine learning scholars have developed what is called **probabilistic topic modelling**, a group of algorithms that do not require human-annotated data and simply discover the themes that emerge in a document by analyzing its words.

Topic modelling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives (Blei, 2009). In other words, it is used to discover the structure or topical patterns in a collection of documents, label new documents according to these topics, and use the labels to organize and search textual data. For example, a topic model can automatically divide a collection of customer reviews posted on a website into clusters according to the use of words within each review, which provides valuable insights more efficiently and at lower cost in comparison with analyses conducted by human labour. Such automation offers a useful method to explore and structure a large set of documents, offering great opportunities for research in various disciplines. These exploratory models are not limited to text corpora as they can be used for collaborative filtering, content-based image retrieval, and bioinformatics (Boyd-Graber, Hu, and Mimno 2017). Under the purpose of exploration of topic models within larges sets of data, automated topic modeling has been used in a variety of contexts including financial fraud detection (Dong, Liao, and Zhang 2018), social media analytics (Cummings and Dennis 2018), information security (Abbasi and Chen 2008; Samtani et al. 2017; Yue, Wang, and Hui 2019), e-commerce (Adamopoulos, Ghose, and Todri 2018), online communities behaviors, and even systematic literature reviews (Sidorova et al. 2008).

Latent Dirichlet Allocation model (LDA; Blei et al. 2003) is a hierarchical Bayesian approach to topic modelling that describes a generative process of document creation. The goal of LDA is to infer topics as latent variables from the observed distribution of words in each document.

2 Motivation to Learn LDA

David Blei explains in one of his lectures that, from a machine learning perspective, he thinks of topic modelling as a case study in applying hierarchical Bayesian models to grouped data. He mentions that topic modelling touches on and brings together multiple methods in statistical learning such as Hierarchical Bayesian methods, fast approximate posterior inference, modelling with graphs, conjugate priors and nonconjugate priors, etc. Therefore, focusing on a topic model can be an opportunity to understand the application of these various methods.

Within the list of topic modelling methods, LDA is the most widely used one, especially in research disciplines relying heavily on textual data to extract information. For example, applications of topic models in information systems literature include the analysis of blog content (Singh et al., 2014), stock recommendation messages (Aral et al., 2011), and firms' financial reports (Bao and Datta, 2014). The widespread adoption of LDA is not surprising considering that previous literature shows

humans tend to agree with the coherence of topics generated by LDA, providing strong support for the use of topic models for information extraction purposes (Chang et al., 2009). In the computer science literature, the research paper by Blei et al. (2003) that proposed LDA for the first time is one of the most cited papers in machine learning. Overall, given its popularity and widespread use, and importance in machine learning literature, I believe teaching LDA and discussing the various opportunities that LDA offers for data collection is a valuable contribution to NLP education, in particular for a group of students with differing backgrounds and conducting research in varying scientific disciplines.

3 Designing the Lecture

3.1 Learning Goals

The learning goals are provided both as a guidance to the lecturer when designing the lecture and as a tool for students' self-assessment following the lecture. The assignment described in the Learning Requirements section is designed to provide an objective assessment of whether the learning goals are achieved. Specifically, following the lecture and the requirements for the lecture, students are expected to be able to:

LG #1: Describe the LDA model to another person who is familiar with Bayesian learning but unknowledgeable about topic modelling

LG #2: Identify at least one example of how LDA model has been used in research within any scientific discipline or business setting

LG #3: List at least one advantage and disadvantage of LDA compared to other methods of topic modelling

LG #4: Apply an LDA model within the programming language of choice

3.2 Resources

Required

Primary resource: Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," Journal of Machine Learning Research (3), pp. 993-1022.

Background reading (for students unfamiliar with Bayesian learning): TBD

Optional

Video lecture by David Blei on Topic Models

CPSC 540: Machine Learning slides by Mark Schmidt on Topic Models

Introduction to Latent Dirichlet Allocation on Edwin Chen's blog for a fun and intuitive explanation of LDA

3.3 Learning Requirements

Before Lecture

The students will be expected to come to class having done the background reading on Bayesian learning provided ahead of lecture. This reading is especially required for students who are unfamiliar with Bayesian learning or simply need a refresher on their knowledge.

After Lecture

The students will need to complete a self-assessed assignment to be submitted approximately a week after the lecture. This assignment will consist of qualitative and quantitative questions. Specifically, the structure of the assignment will be as follows:

- A one-paragraph summary of a use case of LDA in a research paper of the student's choice (most likely within their field of research)
- Various qualitative questions assessing the student's understanding of the LDA model
- Application of the LDA model in the programming language of the student's choice using data provided by the instructor

The assignment will be designed in a way that would enable students to complete it in at most five hours.

4 The Lecture Content

4.1 Topic Modelling

Topic modeling began with a linear algebra approach called Latent Semantic Analysis (LSA), which is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer and Dumais 1997). The goal is to find the best low rank approximation of a document-term matrix. In LSA, three are three major claims directing the analysis. Firstly, the semantic information in a document can be derived from a word-document co-occurrence matrix. Next, the dimensionality reduction is an essential part of this derivation. Finally, words and documents can be represented as points in Euclidean space. Topic models are consistent with the first two claims, but differ in the last in that semantic properties of words and documents are expressed in terms of probabilistic topics rather than points in space (Steyvers and Griffiths 2006).

Topic models are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. These models specify a simple probabilistic procedure by which documents can be generated, hence topic models are generative models. A generative model describes how words in documents might be generated on the basis of latent (random) variables. The goal is to find the best set of latent variables that can explain the observed data (i.e., observed words in documents), assuming that the model actually generated the data. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents. This generative process does not make any assumptions about the order of words as they appear in documents. The only information relevant to the model is

the frequency of words that are produced. Ignoring the word order is defined as the bag-of-words assumption, and is common to many statistical models of language including LSA. Topic models do not require any prior annotations or labelling of documents, i.e. the topics are uncovered via the analysis of solely the original texts. Discovered topics can be used to organize, summarize, and annotate documents at a scale that would require considerably more time to accomplish by human annotation.

There is a distinct advantage in the topic modelling approach in which the semantic properties of documents and words are expressed with probabilistic topics rather than spatial representation. That is, each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms (Steyvers and Griffiths 2006). A variety of probabilistic topic models have been used to analyze the content of documents and the meaning of words. These models all use the fundamental idea – that a document is a mixture of topics. In what follows, we explain this common base drawing heavily on the explanation provided by Steyvers and Griffiths (2006).

4.1.1 Core Process

Each word w_i in a document is generated by first sampling a topic from the topic distribution, then choosing a word from the topic-word distribution. So, the model specifies the following distribution over words within a document:

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j)\ P(z_i = j) \tag{1}$$

To simplify notation, let $\phi^{(j)} = P(w \mid z = j)$ refer to the multinomial distribution over words for topic j, and $\theta^{(d)} = P(z)$ refer to the multinomial distribution over topics for document d. Our notation is summarized in Table 1.

Table 1: Notation for topic modelling

$P(z)$	distribution over topics z in a particular document
$P(w \mid z)$	probability distribution over words w given topic z
w_i	ith word in a document
$P(z_i = j)$	probability that jth topic was sampled for the ith word token
$P(w_i \mid z_i = j)$	probability of word w_i under topic j
T	number of topics
D	number of documents in the text collection
N_d	number of word tokens in document d
$N = \sum N_d$	total number of word tokens

As mentioned earlier, various different topic models exist and although they follow the same fundamental idea, they make slightly different statistical assumptions. Here, we explore a few of these topic models that have shown to be used frequently in information systems research.

4.1.2 Algorithms of Topic Modelling

4.1.2.1 Probabilistic Latent Semantic Indexing (pLSI)

Hofmann (2001) introduced the Probabilistic Latent Semantic Indexing (pLSI) model, thus added probabilistic topic modelling to the existing document modelling techniques. The pLSI model posits that a document label d and a word w_n are conditionally independent given an unobserved topic z:

$$P(d, w_n) = P(d) \sum_z P(w_n \mid z)\ P(z \mid d) \tag{2}$$

As $p(z \mid d)$ specifies the mixture weights of the topics for a particular document d, pLSI is capable of capturing that a document may contain multiple topics. The d is a dummy index into the list of documents in the training set. pLSI laid the foundation for LDA (Boyd-Graber, Hu, and Mimno 2017), which we explain next.

4.1.2.2 Latent Dirichlet Allocation (LDA)

It is difficult to test the generalizability of the model to new documents using pLSI, which does not make any assumptions about how the mixture weights θ are generated. Since d is a multinomial random variable with as many possible values as there are training documents, the pLSI model learns the topic mixtures $p(z \mid d)$ only for those documents on which it is trained which makes the pLSI incapable of assigning probabilities to previously unseen documents. Blei, Ng, and Jordan (2003) address such limitation by extending pLSI with symmetric Dirichlet priors α and β on θ and ϕ, respectively. This new model, which is called Latent Dirichlet Allocation (LDA), is coined as the simplest topic model (Blei 2012).

A major assumption behind LDA is that each document consists of multiple topics and each topic is a distribution over a fixed vocabulary. The model assumes that these topics - distributions over the vocabulary - are generated prior to the documents. All the documents in the collection of documents share the same set of topics, but each document consists of these topics in different proportion. Blei (2012) mentions a two-stage process to explain how words are generated in this model:

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

Each word in each document is drawn from one of the topics (step #2b), which is chosen from the per-document distribution over topics (step #2a). Step #1 outlines that each document exhibits the topics in different proportion and this determines the topic selected for a given word in step #2a. The last part, step #2b, indicates that each word in each document is drawn from one of the topics. This model can be represented graphically as in Figure 1.

Due to its simplicity, LDA has been adopted and used widely in academic research within various disciplines. It has also been extended and improved in various ways. Nowadays, many of the probabilistic topic models can be seen as a variation of the LDA model.

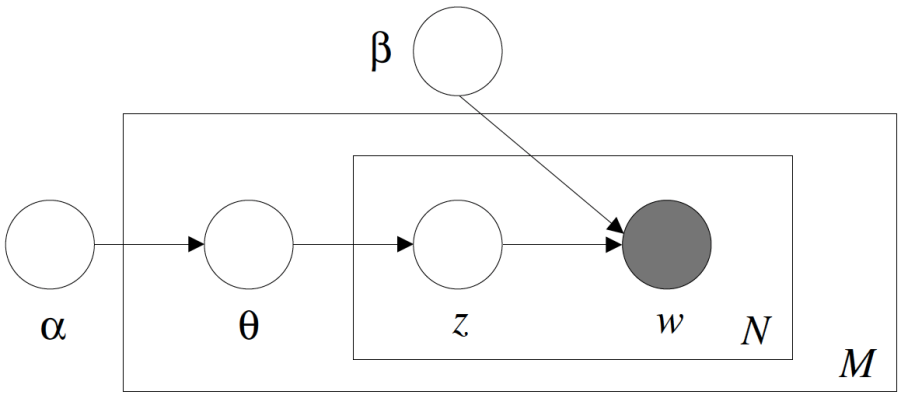

Figure 1: Graphical representation of LDA (Blei et al. 2003)

4.1.2.2.1 Parameter Estimation for LDA

The goal of topic modelling is to infer the topics given a collection of documents. Although the documents are observed, all else is unobserved, i.e. hidden structure. The main computational problem in topic modelling is to use what is observed - documents - to discover the unobserved/hidden structure - the topics, per-document topic distributions, and the per-document per-word topic assignments. The problem can be rephrased as computing the posterior distribution in the model, which is the conditional distribution of the hidden variables given the documents.

$$P(\theta, z \mid w, \alpha, \beta) = \frac{P(\theta, z, w \mid \alpha, \beta)}{P(w \mid \alpha, \beta)} \tag{3}$$

Although this posterior distribution is quite difficult to compute, there exist a wide variety of approximate inference algorithms that can be considered for LDA, including Laplace approximation, variational approximation, and Markov chain Monte Carlo (Blei, Ng, and Jordan 2003). These algorithms, which can be organized into two major categories as sampling-based algorithms and variational algorithms, aim to approximate the posterior distribution by forming an alternative distribution over the latent topic structure that is adapted to be close to the true posterior (Bao and Datta, 2014). Sampling-based algorithms attempt to collect samples from the posterior to approximate it with an empirical distribution. Rather than approximating the posterior with samples, variational methods posit a parametrized family of distributions over the hidden structure and then solve an optimization problem to find the member of that family that is closest to the posterior. Collapsed Gibbs sampling (Griffiths and Steyvers 2004) and variational EM algorithm (Blei, Ng, and Jordan 2003) are the most commonly used sampling-based and variational methods, respectively.

4.2 Applications

Ever since the seminal paper by Blei, Ng, and Jordan (2003), topic modeling techniques have garnered interests from scholars from plethora of scientific fields. While initially topic modeling was used to understand patterns in non-scientific historical documents (e.g. newspapers, historical

records, language books), soon its application found a footprint in academia (Boyd-Graber, Hu, and Mimno 2017). Griffiths and Steyvers (2004) were among the early researchers to use automated techniques to conduct a systematic literature review of scientific fields in which they classified the fields of science based on prior published studies (Griffiths and Steyvers 2004). With the emergence of big data and availability of most of textual data online, the opportunities for using topic modeling techniques increased exponentially. Namely, textual data of blog posts, financial statements, corporate policies, health records, product descriptions, and product advertisement created an environment for experts in textual analysis to shine. The boom of social media only added to this hype and amount of available data. LDA applications were eventually used in many fields as well as the in information systems (IS) research from which we will examine a few.

In one of the earlier works in IS research, Sidorova et al. (2008) used latent semantic analysis (LSA) to capture the core of IS research over the last thirty years; using this automated textual analysis technique, they discovered that IS studies can be categorized under five distinct groups (i.e. market, organizations, groups, individuals, development) and further categorized the studies up to one hundred sub-categories (Sidorova et al. 2008). Soon after, topic modeling application expanded beyond retrospective analysis of historical documents. Automated textual analysis in IS has predominantly been used in social media analysis, e-commerce, fraud detection, and information security; in social media analyses, automated topic modeling has been used to extract firm characteristics (Shi, Lee, and Whinston 2016), individual characteristics (Lee, Qiu, and Whinston 2016), leadership characteristics (Johnson, Safadi, and Faraj 2015), and personality traits (Adamopoulos, Ghose, and Todri 2018). Studies in financial fraud detection and information security have been far and few in between. In a recent study, Dong, Liao, and Zhang (2018) developed an automated system to raise red flags for financial fraud based on social media posts of companies. Yue, Wang, and Hui (2019) analyzed how the topic discussions in Denial of Service Attack (DDoS) forums can predict actual DDoS attacks.

- Q. Wang, Li, and Singh (2018) propose a detection framework that aims to distinguish copycats apps from original apps based on both functionality and appearance. For this, they convert app descriptions and consumer reviews into bag of words. Then, they consider the unique words within these bags as the features of an app and compute the term weights of the app features using the standard term frequency-inverse document frequency (TF-IDF) scheme, which is a measure of how important a word is to a document in a collection. This methodology enables mapping each app to a vector of features. Each value within the vector represents the weighted frequencies of an app feature that appears in an app's description or reviews. To reduce the dimensionality and the independence between words, they conduct singular value decomposition (SVD). They then calculate the feature similarity between apps by taking a cosine of their feature vectors which outputs a value between zero and one that captures the probability of being identical.

- Singh, Sahoo, and Mukhopadhyay (2014) investigate the dynamics of blog reading behavior of employees in an enterprise blogosphere. In this study, the blog reading behavior of an employee is modeled using the number of posts the employee reads on different topics, which are determined using the LDA model.

- Adamopoulos, Ghose, and Todri (2018) examine whether personality traits of social media users attenuate or accentuate the effectiveness of word-of-mouth (WOM). They include latent personality traits and pairwise characteristics of users in social media platforms in an econometric model specification that measures the effect of WOM and subsequent economic outcomes as its dependent variable. In this study, LDA is one of the methods used to measure

the *recipient-sender similarity*, which is one of the explanatory variables included in the model. Specifically, the LDA model gives a measure for the similarity of interests and topics discussed in social media posts by the recipient and the sender.

4.3 Interpreting Results

The most typical evaluation of topic models involves measuring the performance of a a model performs when predicting unobserved documents. Specifically, when estimating the probability of unseen held-out document given a set of training documents, a "good" model should give rise to a higher probability of held-out documents (Bao and Datta 2014). The *perplexity* metric is common in language modelling for this purpose. Perplexity can be defined as the predicted number of equally likely words for a word position on average. It is a monotonically decreasing function of the log-likelihood, so a lower perplexity over a held-out document would indicate a higher log-likelihood, which in turn indicates better predictive performance. Our explanation of perplexity here is cursory; a more complete discussion of this metric can be found in Azzopardi, Girolami, and Risjbergen (2003) and Blei, Ng, and Jordan (2003).

5 References

Aral, S., Ipeirotis, P., and Taylor, S. 2011. "Content and Context: Identifying the Impact of Qualitative Information on Consumer Choice," in *Proceedings of the 32nd International Conference on Information Systems*, Shanghai, China: Association for Information Systems.

Bao, Y., and Datta, A. 2014. "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures," *Management Science* (60:6), pp. 1371-1391.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3), pp. 993-1022.

Blei, D. (2009). *Topic Models* [video file]. Retrieved from http://videolectures.net/mlss09uk_blei_tm/.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models," in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, pp. 288-296.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proc. Natl. Acad. Sci. USA* 101(Suppl 1):5228–5235.

Singh, P. V., Sahoo, N., and Mukhopadhyay, T. 2014. "How to Attract and Retain Readers in Enterprise Blogging?," *Information Systems Research* (25:1), pp. 35-52.

References2

Abbasi, Ahmed, and Hsinchun Chen. 2008. "CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication." *MIS Quarterly* 32 (4): 811–37.

doi:10.2307/25148873.

Adamopoulos, Panagiotis, Anindya Ghose, and Vilma Todri. 2018. "The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms." *Information Systems Research* 29 (3): 612–40. doi:10.1287/isre.2017.0768.

Azzopardi, Leif, Mark Girolami, and Keith van Risjbergen. 2003. "Investigating the Relationship Between Language Model Perplexity and IR Precision-Recall Measures," 2.

Bao, Yang, and Anindya Datta. 2014. "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures." *Management Science* 60 (6): 1371–91. doi:10.1287/mnsc.2014.1930.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77. doi:10.1145/2133806.2133826.

Blei, David M., Andrew Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (March): 993–1022.

Boyd-Graber, Jordan, Yuening Hu, and David Mimno. 2017. "Applications of Topic Models." *Foundations and Trends® in Information Retrieval* 11 (2-3): 143–296. doi:10.1561/1500000030.

Cummings, Jeff, and Alan R. Dennis. 2018. "Virtual First Impressions Matter: The Effect of Enterprise Social Networking Sites on Impression Formation in Virtual Teams." *MIS Quarterly* 42 (3): 697–717. doi:10.25300/MISQ/2018/13202.

Dong, Wei, Shaoyi Liao, and Zhongju Zhang. 2018. "Leveraging Financial Social Media Data for Corporate Fraud Detection." *Journal of Management Information Systems* 35 (2): 461–87. doi:10.1080/07421222.2018.1451954.

Griffiths, Tom, and Mark Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101 (Supplement 1): 5228–35. doi:10.1073/pnas.0307752101.

Hofmann, Thomas. 2001. "Unsupervised Learning by Probabilistic Latent Semantic Analysis." *Machine Learning* 42 (1-2): 177–96.

Johnson, Steven L., Hani Safadi, and Samer Faraj. 2015. "The Emergence of Online Community Leadership." *Information Systems Research* 26 (1): 165–87. doi:10.1287/isre.2014.0562.

Landauer, Thomas K, and Susan T Dumais. 1997. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." *Psychological Review* 104 (2): 211–40.

Lee, Gene Moo, Liangfei Qiu, and Andrew B. Whinston. 2016. "A Friend Like Me: Modeling Network Formation in a Location-Based Social Network." *Journal of Management Information Systems* 33 (4): 1008–33. doi:10.1080/07421222.2016.1267523.

Samtani, Sagar, Ryan Chinn, Hsinchun Chen, and Jay F. Nunamaker. 2017. "Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence." *Journal of Management Information Systems* 34 (4): 1023–53. doi:10.1080/07421222.2017.1394049.

Shi, Zhan, Gene Moo Lee, and Andrew B. Whinston. 2016. "Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence." *MIS Quarterly* 40 (4): 1035–A53. http://ezproxy.library.ubc.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=119473816&site=ehost-live&scope=site.

Sidorova, Anna, Nicholas Evangelopoulos, Joseph S. Valacich, and Thiagarajan Ramakrishnan. 2008.

"Uncovering the Intellectual Core of the Information Systems Discipline." *MIS Quarterly* 32 (3): 467–82. doi:10.2307/25148852.

Singh, Param Vir, Nachiketa Sahoo, and Tridas Mukhopadhyay. 2014. "How to Attract and Retain Readers in Enterprise Blogging?" *Information Systems Research* 25 (1): 35–52. doi:10.1287/isre.2013.0509.

Steyvers, Mark, and Tom Griffiths. 2006. "Probabilistic Topic Models." In *Latent Semantic Analysis: A Road to Meaning*, 15.

Wang, Quan, Beibei Li, and Param Vir Singh. 2018. "Copycats Vs. Original Mobile Apps: A Machine Learning Copycat-Detection Method and Empirical Analysis." *Information Systems Research* 29 (2): 273–91. doi:10.1287/isre.2017.0735.

Yue, Wei T., Qiu-Hong Wang, and Kai-Lung Hui. 2019. "See No Evil, Hear No Evil? Dissecting the Impact of Online Hacker Forums." *MIS Quarterly* 43 (1): 73–95. doi:10.25300/MISQ/2019/13042.