Progress Report:
Latent Dirichlet Allocation and Applications to Big Corpora

Ben Eysenbach

May 8, 2016

1 Introduction

LDA[1] is a canonical example of a mixed-membership model, and it has been applied in a number of fields. In this project, we implement LDA and use it to model a dataset of academic papers.

2 Implementing LDA

The first part of this project focused on implementing the variational inference algorithm in the original LDA paper. The goal was to produce a functional and comprehensible implementation. Learning C++ was a convenient side effect We did not optimize for performance.

2.1 Hyperparameter Optimization

Implementing the updates to hyperparameter α were challenging because the original LDA paper derived them incorrectly in section A.4.2. The log-likelihood is maximized w.r.t α using Newton's Method, which requires the second derivative of the log-likelihood w.r.t α. The correct derivatives are shown below, with differences highlighted in green.

$$L_{[\alpha]} = \sum_{d=1}^{M} \left(\log\Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \log\Gamma(\alpha_i) + \sum_{i=1}^{k} ((\alpha_i - 1)(\Psi(\gamma_{di}) - \Phi(\sum_{j=1}^{k} \gamma_{di}))) \right)$$

$$\frac{\partial L}{\partial \alpha_i} = M \left(\Psi(\sum_{j=1}^{k} \alpha_j) - \Psi(\alpha_i)) + \sum_{d=1}^{M} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj}) \right)$$

$$\frac{\partial L}{\partial \alpha_i \alpha_j} = - \delta(i,j) M \Psi'(\alpha_i) + M \, \Psi'(\sum_{j=1}^{k} \alpha_j)$$

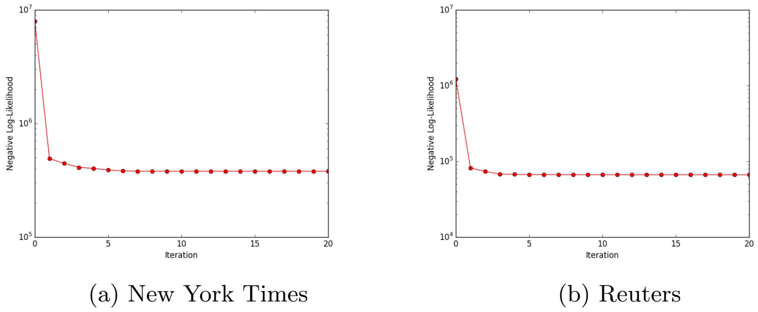
2.2 Random Initialization

Random initialization was important to our variation inference algorithm. As shown in Fig. ??, we achieved the highest log-likelihood when we randomly initialized all the variational parameters. We found this surprising given that the original LDA paper called for a fixed initialization of ϕ and γ. We hypothesize that random initialization helps break symmetry. We also found that removing stopwords was important for breaking symmetry between topics.

random initialization

Randomization	Log-likelihood on NYT	Log-likelihood on Reuters
No randomization	-380382	-66457.8
Randomized α	-380375	-66392.3
Randomized α, ϕ, γ	**-380364**	**-66389.9**

Figure 1: Log-likelihood with random initializations



(a) New York Times

(b) Reuters

Topic 13	Topic 8
net	yeutter
loss	remained
company	class
bank	case
would	cable
qtr	certificates
inc	regulatory
revs	feed
billion	television
oper	unemployment

Figure 2: Negative log-likelihood converging on the New York Times (left) and Reuters (center) datasets. Top words for two topics in Reuters (right).

2.3 Simple Experiments with LDA

Next, we tested our implementation on two small datasets. We originally planned to use the TREC AP news corpus and C Elegans abstract corpus from the original LDA paper. Unfortunately, we were unable to acquire these datasets. Instead, we used a dataset of music articles from the New York Times[1] and business news from Reuters[13].

As a sanity check, we computed the negative log-likelihood for each iteration and confirmed that they converged. This is shown in Fig. 2 (left and center). Note how quickly the algorithm converges, even when plotted in log-scaled. We also computed the top words each topic. Fig. **??** (right) shows the top worsd for topics 13 and 8. The first topic relates to business revenues, while the second relates to government regulation and agriculture. “yeutter” refers to Clayton Keith Yeutter, the Secretary of Agriculture under George H. W. Bush.

3 DSPACE

3.1 Collecting Data

The next part of the project applied LDA to a dataset of academic papers. The MIT libraries manages DSPACE[2], a digital repository of papers written by MIT affiliates. We scraped this website to gather the following metadata for 100,906 papers: authors, title, department and abstract. We have made this metadata dataset publicly available.[3].

3.2 Algorithm

With this data, we wanted to recover the underlying topics of each paper by applying LDA to the paper abstracts. Given the size of these data (11,299,213 words) we opted to use an optimized, parallelized implemtation of Online LDA [7] provided by Gensim [6]. We used this implementation on our dataset of DSPACE abstracts to learn a topic model with 50 topics. It converged after three iterations through the dataset.

3.3 Evaluating Learnt Topics

Both the model and the learning algorithm we used are approximations. Whether they are useful depends on the task at hand. Because we are not interested in any particular task, we use “semantic meaningfullness” of the learnt topics as a proxy. We evaluate the learnt topics by computing the top words and documents for each topic and by visualizing the geometric of the topic latent space.

First, we compute the most likely words for each topic by examining the rows of the β matrix. Table 3 shows the top words for six chosen topics. Column titles such as ”Supply Chains” were not produced by the algorithm. Note that the words within each topic are specific, but the six topics themselves are different.

[1] `https://code.google.com/archive/p/topic-modeling-tool/downloads`

[2] `https://dspace.mit.edu/`

[3] `INSERTURLHERE`

Topic 2 - Supply Chains	Topic 22 - Climate Change	Topic 26 - Air Pollution	Topic 30 - Chemistry	Topic 38 - Human Diseases	Topic 43 - Neuroscience
model	climate	policy	surface	gene	cell
supply	model	water	energy	human	cells
research	temperature	emissions	phase	genetic	protein
product	global	economic	using	disease	expression
cost	atmospheric	housing	temperature	genes	proteins
system	surface	data	high	biological	dna
management	emissions	environmental	water	cell	gene
industry	changes	development	thermal	model	signaling
business	using	air	properties	expression	cellular
chain	change	carbon	experimental	data	role

Figure 3: Top words for each topic

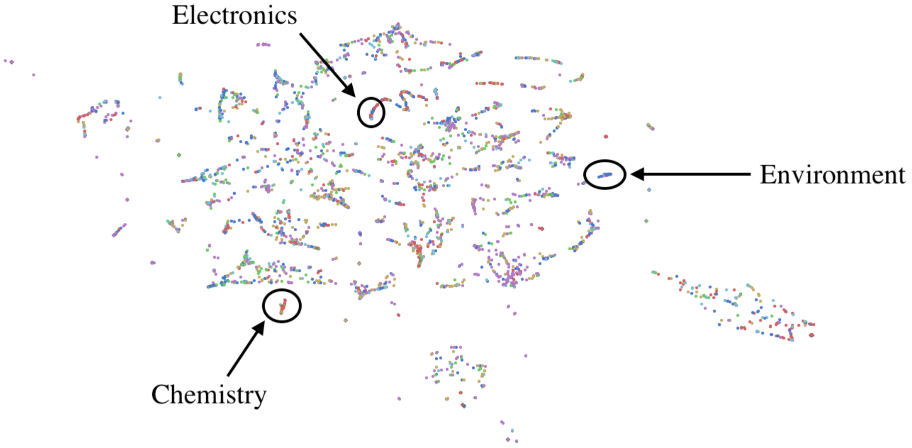


Figure 4: Embedding of papers by learnt topics

Next, we compute the top documents for each topic. We do this examining the distribution over topics assigned to each document. For some topic i, we sort the documents by the probability each belongs to topic i. Figure 4 shows the titles of these top documents. The selected topics are the same as in 3. As before, notice that each topic appears well defined yet distinct. Note that the learning algorithm did not have access to the document titles during training.

These two figures demonstrate that LDA learns meaningful topics for this dataset. However, they do not reveal relationships between topics.

3.4 Visualization

LDA posits that each document is a mixture over topics. We expect that semantically similar documents have "close" topic proportions. Unfortunately, human eyes are two dimensional while these distributions over topics are too-many[4] dimensional ($k = 50$). We explored two approaches: Johnson Lindenstrauss and TSNE.

Johnson Lindenstrauss[9] is an approach to reducing the dimension of data while preserving pairwise distances. The proof of correctness is usually stated using L2 norms. We attempted to adopt the approach to the symmetrized KL norm. After many failed attempts, we found a paper proving that it is impossible to use Johnson Lindenstrauss with the symmetrized KL norm to achieve a low-distortion embedding.[5]

TSNE[4] is another method for computing a low dimensional embedding while preserving pairwise distances. Unlike Johnson Lindenstrauss, the embedding is nonlinear and does not have the same correctness properties. Nonetheless, it has been used successfully in a number of settings (e.g. [10], [11], [12]). We applied TSNE to a random sample of 5,000 papers, as shown in Fig. [5]. In that figure each circle corresponds to a paper, and the color of the circle indicates the department in which paper was

[4] Pun intended.

published. Recall that the learning algorithm did not have access to these department labels during training. Fig. [5] shows department-specific clusters, three of which are circled and labeled. Titles of documents from these three clusters are shown in Fig. 6. Similar to the topics shown in 4, these clusters appear well defined and distinct from one another.

We had hoped to show that documents published in the same department would be clustered in this embedding. While Fig 5 shows some clustering by department, the departments are more mixed than hoped. This result is not entirely surprising given the number of interdepartmental faculty and labs at MIT. We tested both the symmetrized KL and L2 distance in TSNE. Surprisingly, the symmetrized KL failed to form meaningful clusters. Fig 5 uses L2.

3.5 Clustering Authors

The DSPACE dataset we collected can be used not only for discovering document topics, but also for clustering authors. We propose (but do not implement) two methods for this task.

The first method takes a frequentist approach. Let $C(i)$ denote the corpus of papers written by author i and $t(d)$ be the distribution over topics assigned by LDA to document d. Then define the distance between authors i and j as

$$d(i,j) = \frac{1}{|C(i)||C(j)|} \sum_{d_i \in C(i)} \sum_{d_j \in C(j)} KL_{symm}(t(d_i), t(d_j))$$

Applying this distance metric to each pair of authors defines a graph over authors. We can then model the graph using techniques such as Mixed Membership Stochastic Block Models [8].

The second method takes a fully Bayesian approach. We extend the LDA model to include author-specific distributions over topics Θ_i in addition to a global distribution over topics Θ. Specifically, we define the following generative model:

1. Choose $N \sim Poisson(\xi)$
2. Choose $\Theta \sim Dir(\alpha)$
3. For each author i, choose $\Theta_i \sim Dir(\Theta)$.
4. For each of N words w_n:
   (a) Choose $z_n \sim Multinomial(\Theta_i)$
   (b) Choose $w_n \sim P(w_n|z_n, \beta)$

The first method is convenient given a model already fit to your corpus. The second method is more computationally intensive, but will better capture the variance of cluster assignments. While the methods described above are aimed at assigning authors to topics, they can be applied directly to other document attributes, such as department or year of publication.

4 Conclusion

What did we do What did we not do Code is available here: `INSERTURLHERE`

Additional Figures

References

[1] Blei, David M and Ng, Andrew Y and Jordan, Michael I, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, Vol 3, 2003. (pg 993 - 1022)

[2] Hoffman, Blei, Wang, and Paisley, *Stochastic Variational Inference*, Journal of Machine Learning Research, Vol 14, 2013. (pg 1303 - 1347)

[3] Andrew Kachites McCallum, *MALLET: A Machine Learning for Language Toolkit*, `http://mallet.cs.umass.edu`, 2002.

Topic 2 - Supply Chains	Topic 22 - Climate Change	Topic 26 - Air Pollution	Topic 30 - Chemistry	Topic 38 - Human Diseases	Topic 43 - Neuroscience
An exploration of supply chain management practices in the aerospace industry and in Rolls-Royce	A comparison of the behavior of different AOGCMs in transient climate change experiments	Food security and sustainable resource management	Temperature-dependent thermal conductivity in silicon nanostructured materials studied by the Boltzmann transport equation	Manipulating the Selection Forces during Affinity Maturation to Generate Cross-Reactive HIV Antibodies	An Anterior-to-Posterior Shift in Midline Cortical Activity in Schizophrenia During Self-Reflection
Using and extended enterprise model to increase responsiveness	Global warming projections : sensitivity to deep ocean mixing	Economic and policy implications of urban air pollution in the United States, 1970 to 2000	High-strain actuation of lead-free perovskites : compositional effects, phenomenology and mechanism	Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities	Sound temporal envelope and time-patterns of activity in the human auditory pathway : an fMRI study
Re-architecting the failure analysis supply chain	Consequences of Considering Carbon/Nitrogen Interactions on the Feedbacks between Climate and the Terrestrial Carbon Cycle	What does stabilizing greenhouse gas concentrations mean?	Experimental studies of the thermoelectric properties of microstructured and nanostructured lead salts	Hepatitis C Virus Network Based Classification of Hepatocellular Cirrhosis and Carcinoma	Reversal of TMS-induced motor twitch by training is associated with a reduction in excitability of the antagonist muscle.
Sales & operations planning in a global business	Sensitivity of Climate Change Projections to Uncertainties in the Estimates of Observed Changes in Deep-Ocean Heat Content	Global health and economic impacts of future ozone pollution	Orientation of MgO thin films on Si(001) prepared by pulsed laser deposition	Genetic association with overall survival of taxane-treated lung cancer patients - a genome-wide association study in human lymphoblastoid cell lines followed by a clinical association study	Laminar differences in gamma and alpha coherence in the ventral stream
Multi-echelon inventory management for a fresh produce retail supply chain	Tropical Cyclone Activity Downscaled from NOAA-CIRES Reanalysis, 1908-1958	Climate Co-benefits of Tighter SO2 and NOx Regulations in China	Heat transfer during film condensation of potassium vapor	Effects of thymic selection of the T cell repertoire on HLA-class I associated control of HIV infection	Unconscious pop-out: attentional capture by unseen feature singletons only when top-down attention is available
Inventory optimization in high volume aerospace supply chains	Formation of a localized acceleration potential during magnetic reconnection with a guide field	Consumption-Based Adjustment of China's Emissions-Intensity Targets: An Analysis of its Potential Economic Effects	Superoleophobic Surfaces through Control of Sprayed-on Stochastic Topography	Differential Virulence of Clinical and Bovine-Biased Enterohemorrhagic Escherichia coli O157:H7 Genotypes in Piglet and Dutch Belted Rabbit Models	Dissociable Influences of Auditory Object vs. Spatial Attention on Visual System Oscillatory Activity
Improving supply chain responsiveness for diesel engine remanufacturing	Sensitivity of tropical precipitation extremes to climate change	The Current Water and Agriculture Context, Challenges, and Policies	(Invited) Role of Chemical Heterogeneities on Oxygen Reduction Kinetics on the Surface of Thin Film Cathodes	A multidimensional platform for the purification of non-coding RNA species	Two Critical and Functionally Distinct Stages of Face and Body Perception
Product development risk management and the role of transparency	Historical and idealized climate model experiments: an intercomparison of Earth system models of intermediate complexity	Carbon emissions in China: How far can new efforts bend the curve?	Unified Model for Contact Angle Hysteresis on Heterogeneous and Superhydrophobic Surfaces	Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes	Neuroimaging investigation of the motor control disorder, dystonia with special emphasis on laryngeal dystonia
Modeling the impact of complexity on transportation	Electron temperature fluctuations associated with the weakly coherent mode in the edge of I-mode plasmas	Multiple metrics for quantifying the intensity of water consumption of energy production	Electrostatic charging of jumping droplets	SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia	Multivoxel Pattern Analysis Reveals Auditory Motion Information in MT+ of Both Congenitally Blind and Sighted Individuals
Emergence of strategic direction, organizational structure and employee integration : a framework for the Dialectic Organization	Time-Varying Climate Sensitivity from Regional Feedbacks	Future of oil and gas development in the western Amazon	Theory of Raman enhancement by two-dimensional materials: Applications for graphene-enhanced Raman spectroscopy	Use of a conservation-of-linkage strategy to identify a candidate for the rat Lymphopenia gene	Attention Drives Synchronization of Alpha and Beta Rhythms between Right Inferior Frontal and Primary Sensory Neocortex

Figure 5: Titles of papers which had the highest likelihood of belonging to each topic

Environment	Chemistry	Electronics
The impact of detailed urban-scale processing on the composition, distribution, and radiative forcing of anthropogenic aerosols	Copper-catalyzed arylation of 1,2-amino alcohols. Synthesis of N-terminal, peptide helix initiators, and characterization of highly helical, capped polyalanine peptides	An aligner for X-ray nanolithography
Protection of Coastal Infrastructure under Rising Flood Risk	Halogenated 1'-methyl-1,2'-bipyrroles (MBPs) in the Norwestern Atlantic	An algorithm for rate allocation in a packet-switching network with feedback
Land conversion in Amazonia and Northern South America : influences on regional hydrology and ecosystem response	Synthesis of Marine Polycyclic Polyethers via Endo-Selective Epoxide-Opening Cascades	Propagation and scattering of electromagnetic waves in complex environments
Climate change impacts on freshwater recreational fishing in the United States	Three dimensional molecular architectures for the synthesis and improved properties of high performance polymers	A method for system performance analysis of the SuperSPARc microprocessor
Coupling of a regional atmospheric model (RegCM3) and a regional oceanic model (FVCOM) over the maritime continent	Computational Explorations of Mechanisms and Ligand-Directed Selectivities of Copper-Catalyzed Ullmann-Type Reactions	Performance prediction of an image management and communication system for cardiac ultrasound
An analysis of the carbon balance of the Arctic Basin from 1997 to 2006	The design and synthesis of polymeric assemblies for materials applications : chemosensing, liquid crystal alignment and block copolymers	An intelligent automobile diagnostic system
Effects of oceanic and atmospheric phenomena on precipitation and flooding in the Manafwa River Basin	Protein Thioester Synthesis Enabled by Sortase	Marginal cost congestion pricing under approximate equilibrium conditions
Investigating the role of Trichodesmium spp. in the oceanic nitrogen cycle through observations and models	Towards incorporation of catalytic function into small folded peptide scaffolds	Modeling poly-silicon gate depletion in submicron MOS devices
Ionospheric Backscatter Observations at Millstone Hill	Development of novel polymeric architectures for applications in drug delivery and studies towards the synthesis of perfect polymers by iterative exponential growth "Plus" (IEG+)	A methodology for sizing components in a dual-voltage automotive electrical system
Heightened hurricane surge risk in northwest Florida revealed from climatological-hydrodynamic modeling and paleorecord reconstruction	Rapid prototyping of carbon-based chemiresistive gas sensors on paper	Characterization of a wideband monopulse piezoelectric direction finder

Figure 6: Titles of papers in each of the three circled regions in Fig. 5

[4] Van der Maaten, Laurens and Hinton, Geoffrey, *Visualizing data using t-SNE*, Journal of Machine Learning Research, Vol 9, 2008. (pg 2579 - 2605)

[5] Bhattacharya, Arnab and Kar, Purushottam and Pal, Manjish, *On Low Distortion Embeddings of Statistical Distance Measures into Low Dimensional Spaces*, Database and Expert Systems Applications, 2009. (pg 164 - 172)

[6] Radim Řehůřek and Petr Sojka *Software Framework for Topic Modelling with Large Corpora*, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, `https://radimrehurek.com/gensim/index.html`, 2010. (pg 45 - 50)

[7] Hoffman, Matthew and Bach, Francis R and Blei, David M, *Online Learning for Latent Dirichlet Allocation*, Advances in Neural Information Processing Systems, 2010. (pg 856 - 864)

[8] Airoldi, Edo M and Blei, David M and Fienberg, Stephen E and Xing, Eric P, *Mixed Membership Stochastic Blockmodels*, Advances in Neural Information Processing Systems, 2009. (pg 33 - 40)

[9] Johnson, William B and Lindenstrauss, Joram, *Extensions of Lipschitz mappings into a Hilbert space*, Contemporary Mathematics, Vol 26, 1984. (pg 189 - 206)

[10] Bengio, Yoshua, *Learning Deep Architectures for AI*, Foundations and Trends® in Machine Learning, Vol 2, 2009. (pg 1 - 127)

[11] Mohamed, Abdel-rahman and Hinton, Geoffrey and Penn, Gerald, *Understanding how Deep Belief Networks Perform Acoustic Modelling*, Acoustics, Speech and Signal Processing (ICASSP), 2012. (pg 4273 - 4276)

[12] Shen, Fumin and Shen, Chunhua and Shi, Qinfeng and Hengel, Anton and Tang, Zhenmin, *Inductive hashing on manifolds*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013. (pg 1562 - 1569)

[13] Lewis, D. D. and Yang, Y. and Rose, T. and Li, F. *RCV1: A New Benchmark Collection for Text Categorization Research* Journal of Machine Learning Research, `http://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf`, Vol 5, 2004. (pg 361 - 397)