Probabilistic Topic Modeling CS5340 Project

Yannis Montreuil Anand Subramanian Yijie Hu Tiana Chen Rahul Gupta

School of Computing National University of Singapore

April 13, 2023



Table of Contents

- Modeling Topics in Al Research
- 2 Latent Dirichlet allocation
- Results Demonstration
- 4 Extensions

Modeling Topics in Al Research

- Artificial intelligence (AI) has undergone substantial changes in the past few decades.
- Al research has also gained popularity.
 - Number of Al-related papers on arXiv has increased more than sixfold in the past six years.
- We aim to analyze and track the evolution of research over time, and seek to explore some of following use cases.
 - Tracking research trends
 - Identifying key researchers
 - Tracking author-specific topics
 - Identifying related fields
 - Evaluating impact of research
 - Discovering new research directions

Latent Dirichlet allocation

- Topic modeling with Latent Dirichlet allocation (LDA)
- *Topics*: k = 1, ..., K
 - Dirichlet distribution $p(\beta_k|\eta) = \text{Dir}_{\beta_k}[\eta]$, shared prior η
- Documents: d = 1, ..., D
 - Categorical distribution $p(\theta_d|\alpha) = \mathsf{Cat}_{\theta_d}[\alpha]$
 - Proportions $\theta_d = [\theta_{d,a}, \dots, \theta_{d,K}]$ is latent
 - Shared prior $\alpha = [\alpha_1, \dots, \alpha_K]^T$
- *Words*: n = 1, ..., N
 - Word *n* of document *d* assigned to (unobserved) topic $z_{d,n} \sim p(z_{d,n}|\theta_d) = \mathsf{Cat}_{z_{d,n}}[\theta_d]$
 - (observed) word $w_{d,n} \sim p(w_{d,n}|\beta_k, z_{d,n}) = p(\beta_{z_{d,n}})$

4/9

Latent Dirichlet allocation

 "Bag-of-words" model: mutually independent assumption of word-generating process

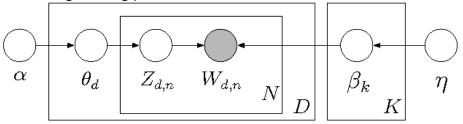


Figure 1: The graphical model for LDA, taken from [1]

- Observed words $W_{d,n}$ from latent $Z_{d,n}$ and β_k .
- $Z_{d,n}$ from latent θ_d
- Shared parameters: α and η .

Results Demonstration

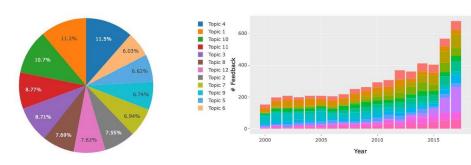
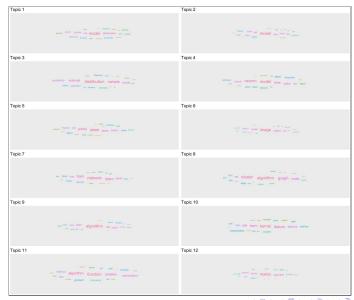


Figure 2: Overall and per year distribution of topics.

Results Demonstration



Extensions

- We extend our work to incorporate lexical priors into our base LDA model.
- Guided LDA: Provide relevant seed words to guide topic exploration.
- Select set of keywords related to the research question of interest.
- We implement this by biasing the prior assigned towards the seed keywords for its corresponding topic
- More targeted exploration of corpora
- Preliminary results are promising.

Initial Exploration of LDA with Lexical Priors

```
receptive_field learning_rule
mutual_information
basis_functions
volpp Visual_cortex
learning_rate
spike_trains_receptive_fields
gradient_descent
non_linear mit_press
hidden_layer real_time
time_series_training_set
ling_rate
advances_neural
advances_neural
```

```
international_conference semi_supervised
Object_recognition
    advances_neural
feature_vector
    more_interviewerror_rate
preprint_arxivlarge_scale
convolutional_neural
conference_computer
    supervised_learning state_art
    training_examples
pattern_recognition_ground_truth
    data_set_deep_learning
    arxiv_preprint
```

(a) Topic without CV priors. (b) Topic with CV lexical priors

Figure 3: Initial Exploration of LDA with lexical priors. In the initial case, on exploration of the topic most similar to computer vision, CV keywords were weighted very less by the topic. Explicitly introducing CV lexical priors improved the topic formation.



Blei, David M. (2010). Introduction to Probabilistic Topic Models.