Probabilistic Graphical Models

Latent Dirichlet Allocation (LDA) for Probabilistic Topic Modeling

Motivation for topic models, Latent Dirichlet Allocation (LDA), parameter estimation in LDA, selection of the number of topics, application of LDA, evaluation methods of topic coherence

Mingyang Xu, mxu5@ncsu.edu
PhD Student in Dr. Samatova's research lab
Department of Computer Science
North Carolina State University

NC STATE UNIVERSITY
Department of Computer Science

Why Topic Modeling from Unstructured Text?

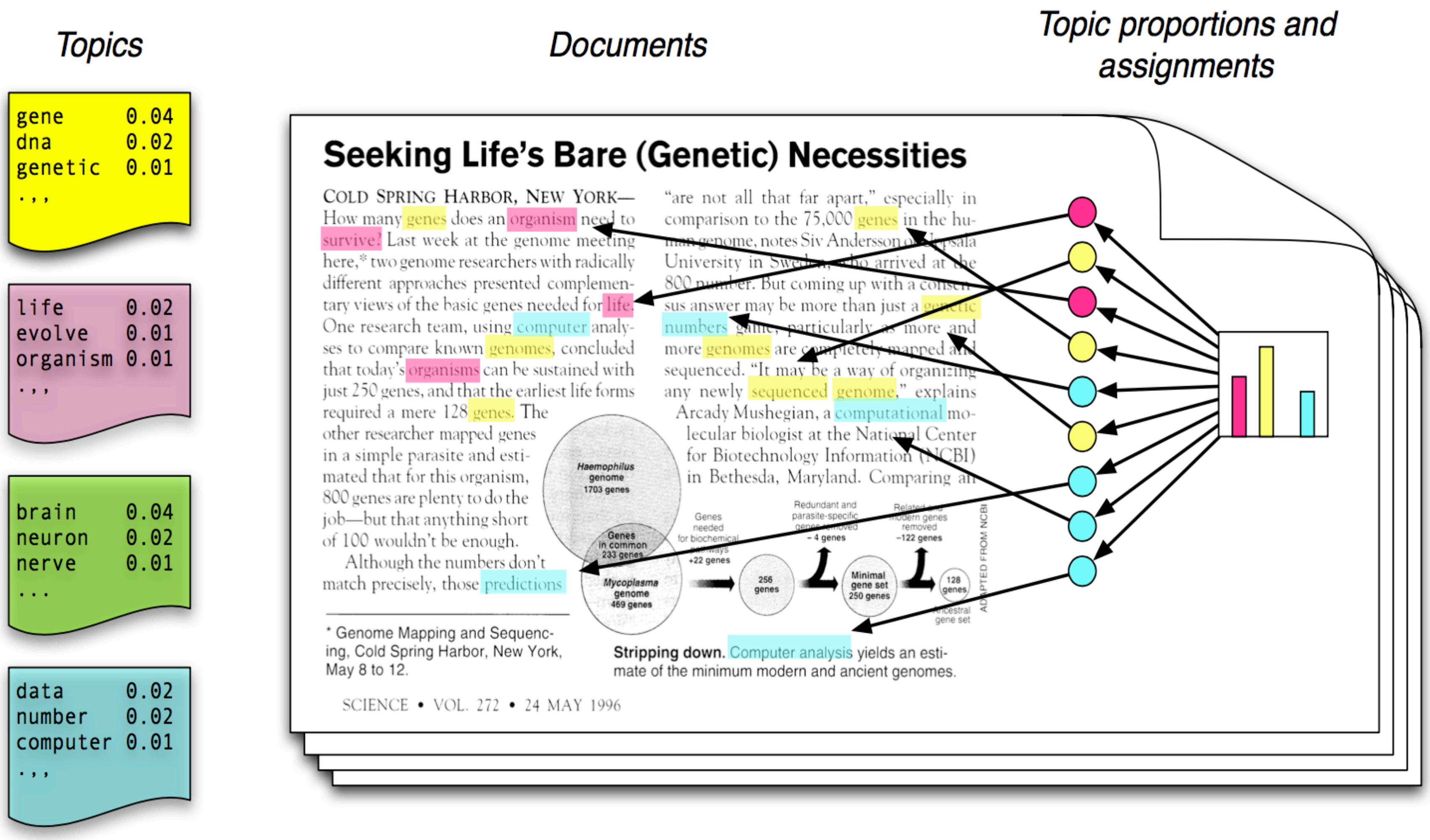
Motivation

- Unstructured text data is ubiquitous: online reviews, news, blogs, etc.
- It's difficult to find what we are looking for
- We need algorithms to help us organize and understand this vast amount of unstructured information

Capabilities of Topic Models

- Automatic organization and summarization of large electronic unstructured text corpus
  - Uncover the major themes (topics) that pervade the corpus
  - Annotate the documents according to those topics
  - Use the annotations to organize and summarize the texts

Overview of a Topic Modeling



Input:

A collection of text documents

Output:

- **A set of topics; topic is the probability distribution over the unique words in the input documents**
- **Probabilistic assignment of each word to a topic**
- **Probability distribution over topics for each document**

Src: Figure from "Probabilistic Topic Models" by David Blei, April 2012 | vol. 55 | no. 4 | Communications of the ACM

A Classic Topic Model: LDA (Latent Dirichlet Allocation)

What is LDA?

- A topic modeling method proposed by Prof. David Blei in JMLR 2003
- A generative model
  - Each document is assumed to be generated by a generative process
  - Presented as a probabilistic graphical model
- Unsupervised learning methodology
  - Only the number of topics is specified in advance

Key Assumptions of LDA

- Documents exhibit multiple topics (but not too many)
- The order of words does not matter in a document ("bag of words")
- The order of documents does not matter ("bag of documents")
- The number of topics is specified and fixed *a priori*

Why Does LDA Work?

- LDA Trades off two goals:
  1. For each document, assigns its words to as few as topics as possible.
  2. For each topic, assigns high probability to as few terms as possible.

- However, these two goals contradict to each other:
  - Assigning each word to a single topic will make many words have equal probability in the topic.
  - Assigning a few words to each topic will make each word in each document be assigned many different topics.

- Trading off these two goals finds groups of tightly co-occurring words in the similar context, which are likely to be semantically related.

Latent Dirichlet Allocation

SELECTION OF MODEL PARAMETERS

How to Choose α and β?

- The intuition of choosing α and β:

  α represents document-topic density - with a higher alpha, documents are made up of more topics, and with lower alpha, documents contain fewer topics.

  β represents topic-word density - with a high beta, topics are made up of most of the words in the corpus and with a low beta they consist of few words.

In practice:

There is no standard for setting α and β.

A rule of thumb given by Griffiths & Steyvers(2004) is to set:

- α = 50/T, where T is the number of topics
- β = 0.1, which is a small number and can be expected to result in a fine-grained decomposition of the corpus into topics

How to Choose the Number of Topics?

- **There is no best approach or standard for choosing the number of topics.**
- **It should be selected based on different datasets.**
- **The intuition: a larger number of topics can provide more detailed information, while a smaller number of topics can provide a bigger picture of your datasets.**

The method proposed by Griffiths & Steyvers(2004):

- **The intuition: Find the number of topics that can most likely generate the observed dataset**

- **Calculate $\log(P(w|\mathrm{T}))$ with different number of topics and select the best number of topics**

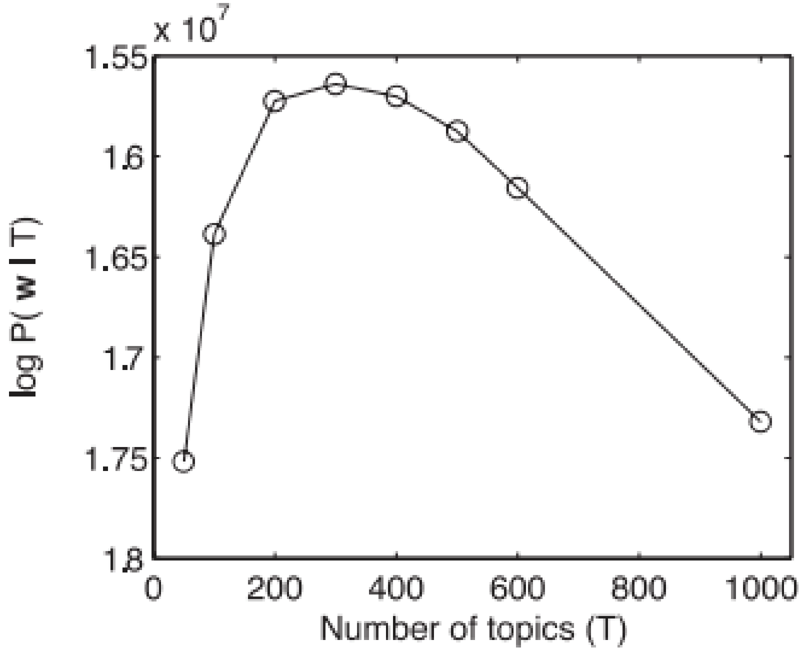


Fig. 3. Model selection results, showing the log-likelihood of the data for different settings of the number of topics, T. The estimated standard errors for each point were smaller than the plot symbols.

Evaluation of Model Performance

Perplexity (A lower perplexity score indicates better generalization performance)

Perplexity(test dataset) = $\exp\left\{-\frac{\sum_{d=1}^{M}\log p(\boldsymbol{w}_d)}{\sum_{d=1}^{M} N_d}\right\}$

- $\log p(\boldsymbol{w}_d)$: log likelihood of the data
- N_d: The number of words in document d

Topic Coherence Score (maximization score: the higher, the better)

$$\text{NPMI}(t) = \sum_{j=2}^{N}\sum_{i=1}^{j-1}\frac{\log\frac{P(w_j, w_i)}{P(w_i)P(w_j)}}{-\log P(w_j, w_i)}$$ ([6] Newman et al., 2009)

$$\text{LCP}(t) = \sum_{j=2}^{N}\sum_{i=1}^{j-1}\log\frac{P(w_j, w_i)}{P(w_i)}$$ ([4] Mimno et al.2011)

- N: The number of top words to keep in each topic
- $P(w_j, w_i)$: The frequency of documents containing both w_j and w_i
- $P(w_i)$: The frequency of documents containing w_i

Demo: Topic Modeling

LDA: LATENT DIRICHLET ALLOCATION

Topic Modeling

STATE-OF-THE-ART

State-of-the-Art: Topic Modeling

Labeled LDA – LLDA [15]

A supervised topic model, which constrains LDA by defining a one-to-one correspondence between LDA's latent topics and user tags.
Specifically, the words in a document can only be assigned the topics corresponding to the document's (observed) label set.

Partially Labeled LDA – PLDA [16]

A partially supervised topic model, which discovers the latent topics within each label, as well as unlabeled, corpus-wide latent topics. Specifically, for each document, PLDA introduces a set of latent topics within each label of the document and a set of latent topics without any labels.

Topical N-grams Model – TNG [12]

A phrase-based topic model, which considers the order of words in the model to discover phrases within each latent topic. Specifically , the model assigns a status distribution for each word to sample a status indicating whether the word should form a bigram with its previous word.

State-of-the-Art: Topic Modeling (cont.)

Sentiment-Topic Models – JST [10], ASUM [11]

Unsupervised topic models, which are able to discover the sentiment of the words, documents, and topics. Specifically, each document is assigned a sentiment distribution, which is used to determine the sentiment of each word in that document. Each discovered topic contains only the words with the same sentiment. The words in different topics can have different sentiments.

Latent Feature Topic Model – LFLDA [6]

An unsupervised topic model that extends LDA by incorporating pre-trained word embedding for discovering more coherent topics, as pre-trained word embedding contains rich semantic information of words that can help topic modeling on small datasets.

Knowledge-based Topic Models – AMC [5], LTM [7]

Unsupervised topic models that incorporates the automatically mined word correlation knowledge, such as must-links (pairs of related words) and cannot-links (pairs of unrelated words), into their sampling procedures to discover more coherent topics.

Topic Modeling Software

Model	Link	Language
LDA	lda scikit-learn.lda gensim.lda	Python Packages
	topicmodels	R Package
	Mallet.topicmodel	Java Package
Labeled LDA	JGibbsLDA	Java
	Stanford.TMT	Scala
Partially Labeled LDA	Stanford.TMT	Scala
Topical N-grams Model	Mallet.topicmodel	Java
JST	jst	C++
ASUM	asum	Java
LFLDA	lflda	Java
AMC	amc	Java
LTM	ltm	Java

Limitations of Traditional Topic Models

Example of coherent topics

Topic 0	Topic 1	Topic 2
iphone	TV	macbook
apple	4K	lenovo
samsung	HD	thinkpad
nexus	sony	windows
android	curve	macair

Example of incoherent topics

Topic 0	Topic 1	Topic 2
iphone	TV	macbook
apple	parking	water
mall	door	tire
closet	sony	windows
android	curve	macair

Limitations:

- Traditional topic models, such as LDA [1], generate topics based on higher-order word co-occurrences.
- Typically require a large number of documents, e.g., thousands of documents, for generating coherent topics

Challenge:

- **How to generate coherent topics when there are limited word co-occurrences in the given corpus?**

References

[1] Blei, D. M. et al. “Latent dirichlet allocation”. the Journal of machine Learning research 3 (2003), pp. 993–1022.

[2] D. L. Silver, Q. Yang, and L. Li. Lifelong Machine, Learning Systems: Beyond Learning Algorithms. In AAAI Spring Symposium: Lifelong Machine Learning, 2013.

[3] S. Thrun. Lifelong Learning Algorithms. In S. Thrun and L. Pratt, editors, Learning To Learn. Kluwer Academic Publishers, 1998.

[4] Mimno, D. et al. “Optimizing semantic coherence in topic models”. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2011, pp. 262–272.

[5] Chen, Z. & Liu, B. “Mining topics in documents: standing on the shoulders of big data”. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2014, pp. 1116–1125.

[6] Nguyen, D. Q. et al. “Improving Topic Models with Latent Feature Word Representations”. Transactions of the Association for Computational Linguistics 3 (2015), pp. 299–313.

[7] Chen, Z. et al. “Discovering coherent topics using general knowledge”. Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM. 2013, pp. 209–218.

[8] Baccianella, S. et al. “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.” LREC. Vol. 10. 2010, pp. 2200–2204.

[9] Hu, M. & Liu, B. “Mining and summarizing customer reviews”. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2004, pp. 168–177.

References

[10] Lin, C. & He, Y. “Joint sentiment/topic model for sentiment analysis”. Proceedings of the 18th ACM conference on Information and knowledge management. ACM. 2009, pp. 375–384.

[11] Jo, Y. & Oh, A. H. “Aspect and sentiment unification model for online review analysis”. Proceedings of the fourth ACM international conference on Web search and data mining. ACM. 2011, pp. 815–824.

[12] Wang, X. et al. “Topical n-grams: Phrase and topic discovery, with an application to information retrieval”. Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE. 2007, pp. 697–702.

[13] D. Chen and C. D. Manning, “A fast and accurate dependency parser using neural networks.” in *EMNLP*, 2014, pp. 740–750.

[14] Church, K. W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. In ACL-89, Vancouver, B.C., pp. 76–83.

[15] Ramage, Daniel, et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009.

[16] Ramage, Daniel, Christopher D. Manning, and Susan Dumais. "Partially labeled topic models for interpretable text mining." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.

[17] Mingyang Xu, Ruixin Yang, Steven Harenberg, Nagiza F. Samatova, A Lifelong Learning Topic Model Structured Using Latent Embedding, in Proceedings of IEEE 11th International Conference on Semantic Computing, 2017