

Topic model

In [machine learning](#) and [natural language processing](#), a **topic model** is a type of [statistical model](#) for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear equally in both. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words. The "topics" produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.

Topic models are also referred to as probabilistic topic models, which refers to statistical algorithms for discovering the latent semantic structures of an extensive text body. In the age of information, the amount of the written material we encounter each day is simply beyond our processing capacity. Topic models can help to organize and offer insights for us to understand large collections of unstructured text bodies. Originally developed as a text-mining tool, topic models have been used to detect instructive structures in data such as genetic information, images, and networks. They also have applications in other fields such as [bioinformatics](#).^[1]

Contents

History

Topic models for context information

Algorithms

See also

Software/libraries

References

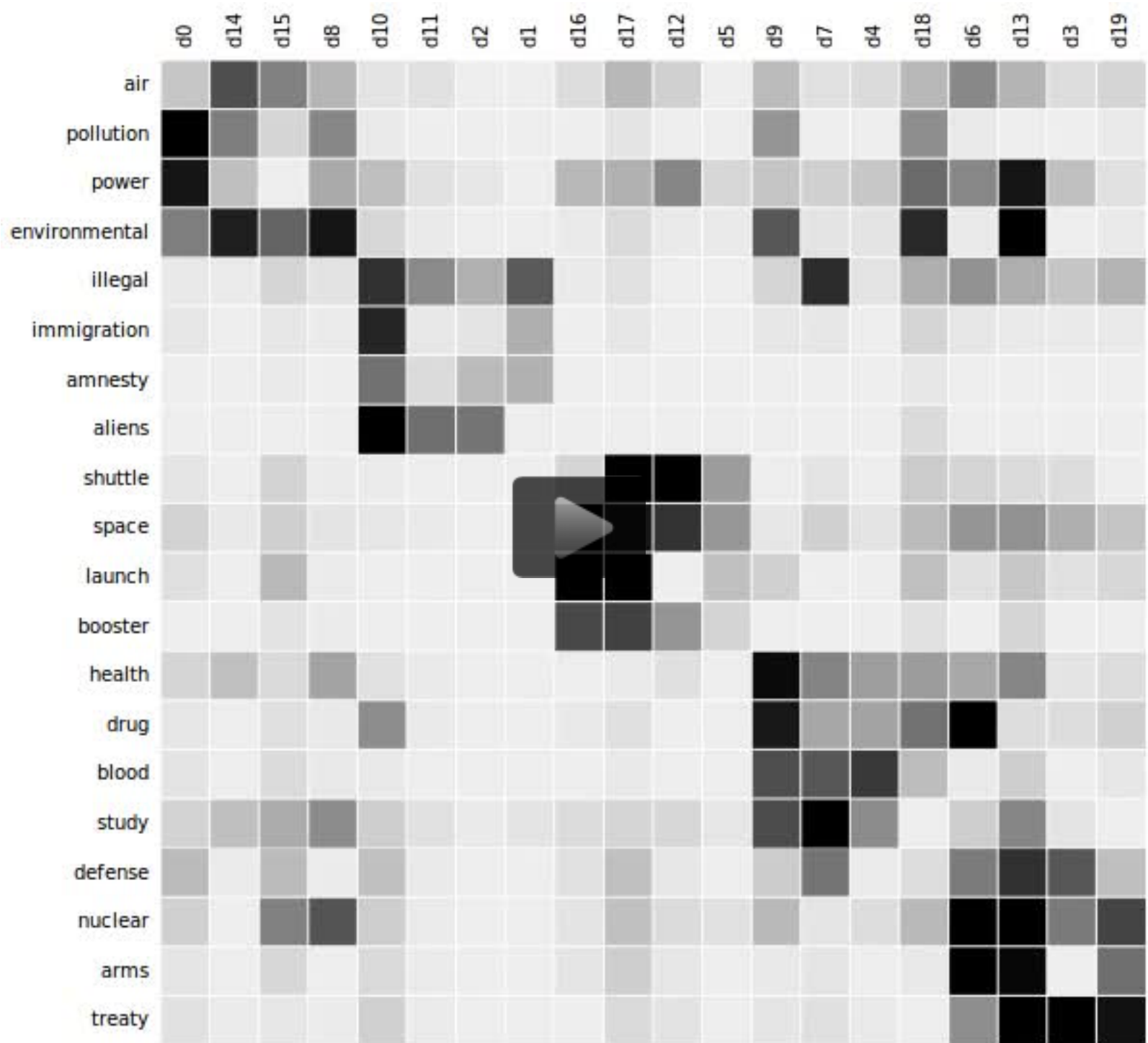
Further reading

External links

History

An early topic model was described by Papadimitriou, Raghavan, Tamaki and Vempala in 1998.^[2] Another one, called [probabilistic latent semantic analysis](#) (PLSA), was created by Thomas Hofmann in 1999.^[3] [Latent Dirichlet allocation](#) (LDA), perhaps the most common topic model currently in use, is a generalization of PLSA. Developed by [David Blei](#), [Andrew Ng](#), and [Michael I. Jordan](#) in 2002, LDA introduces sparse [Dirichlet prior distributions](#) over document-topic and topic-word distributions, encoding the intuition that documents cover a small number of topics and that topics often use a small number of words.^[4] Other topic models are generally extensions on LDA, such as [Pachinko allocation](#), which improves on LDA by modeling correlations between topics in addition to the word correlations which constitute topics.

Top mod for con info



Animation of the topic detection process in a document-word matrix. Every column corresponds to a document, every row to a word. A cell stores the frequency of a word in a document, dark cells indicate high word frequencies. Topic models group both documents, which use similar words, as well as words which occur in a similar set of documents. The resulting patterns are called "topics".^[5]

Approaches for temporal information include Block and Newman's determination the temporal dynamics of topics in the *Pennsylvania Gazette* during 1728–1800. Griffiths & Steyvers use topic modeling on abstract from the journal *PNAS* to identify topics that rose or fell in popularity from 1991 to 2001. Nelson has been analyzing change in topics over time in the *Richmond Times-Dispatch* to understand social and political changes and continuities in Richmond during the *American Civil War*. Yang, Torget and Mihalcea applied topic modeling methods to newspapers from 1829–2008. Mimno used topic modelling with 24 journals on classical philology and archaeology spanning 150 years to look at how topics in the journals change over time and how the journals become more different or similar over time.

Yin et al.^[6] introduced a topic model for geographically distributed documents, where document positions are explained by latent regions which are detected during inference.

Chang and Blei^[7] included network information between linked documents in the relational topic model, which allows to model links between websites.

The author-topic model by Rosen-Zvi et al.^[8] models the topics associated with authors of documents to improve the topic detection for documents with authorship information.

Algorithms

In practice researchers attempt to fit appropriate model parameters to the data corpus using one of several heuristics for maximum likelihood fit. A recent survey by Blei describes this suite of algorithms.^[9] Several groups of researchers starting with Papadimitriou et al.^[2] have attempted to design algorithms with probable guarantees. Assuming that the data were actually generated by the model in question, they try to design algorithms that probably find the model that was used to create the data. Techniques used here include singular value decomposition (SVD) and the method of moments. In 2012 an algorithm based upon non-negative matrix factorization (NMF) was introduced that also generalizes to topic models with correlations among topics.^[10]

See also


- Explicit semantic analysis
- Latent semantic analysis
- Latent Dirichlet allocation
- Hierarchical Dirichlet process
- Non-negative matrix factorization

Software/libraries


- BigARTM (<https://github.com/bigartm/bigartm>)
- Mallet (software project) (<http://mallet.cs.umass.edu/>)
- Stanford Topic Modeling Toolkit (<http://nlp.stanford.edu/software/tmt/tmt-0.4/>)
- Gensim – Topic Modeling for Humans (<http://radimrehurek.com/gensim/>)
- topicmodels R package (<https://cran.r-project.org/package=topicmodels>)
- jLDADMM (<https://github.com/datquocnguyen/jLDADMM>) A Java package for topic modeling on normal or short texts. jLDADMM includes implementations of the LDA topic model and the *one-topic-per-document* Dirichlet Multinomial Mixture model. jLDADMM also provides an implementation for document clustering evaluation to compare topic models.
- TopicModelsVB.jl Julia package (<https://github.com/ericproffitt/TopicModelsVB.jl>)

References

1. Blei, David (April 2012). "Probabilistic Topic Models". *Communications of the ACM*. **55** (4): 77–84. doi:10.1145/2133806.2133826 (<https://doi.org/10.1145/2133806.2133826>).
2. Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh (1998). "Latent Semantic Indexing: A probabilistic analysis" (<http://www.cs.berkeley.edu/~christos/ir.ps>) (Postscript). *Proceedings of ACM PODS*.
3. Hofmann, Thomas (1999). "Probabilistic Latent Semantic Indexing" (<https://web.archive.org/web/20101214074049/http://www.cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf>) (PDF). *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*. Archived from the original (<http://www.cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf>) (PDF) on 2010-12-14.
4. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John (January 2003). "Latent Dirichlet allocation" (<http://jmlr.csail.mit.edu/papers/v3/blei03a.html>). *Journal of Machine Learning Research*. **3**: 993–1022. doi:10.1162/jmlr.2003.3.4-5.993 (<https://doi.org/10.1162/jmlr.2003.3.4-5.993>).
5. http://topicmodels.west.uni-koblenz.de/ckling/tmt/svd_ap.html

6. Yin, Zhijun (2011). "Geographical topic discovery and comparison". *Proceedings of the 20th international conference on World wide web*: 247–256.
7. Chang, Jonathan (2009). "Relational Topic Models for Document Networks". *AIStats*. **9**: 81–88.
8. Rosen-Zvi, Michal (2004). "The author-topic model for authors and documents". *Proceedings of the 20th conference on Uncertainty in artificial intelligence*: 487–494.
9. Blei, David M. (April 2012). "Introduction to Probabilistic Topic Models" (<https://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext>) (PDF). *Comm. ACM*. **55** (4): 77–84. doi:10.1145/2133806.2133826 (<http://doi.org/10.1145/2133806.2133826>).
10. Sanjeev Arora; Rong Ge; Ankur Moitra (April 2012). "Learning Topic Models—Going beyond SVD". arXiv:1204.1956 (<https://arxiv.org/abs/1204.1956>).

Further reading

- Steyvers, Mark; Griffiths, Tom (2007). "Probabilistic Topic Models" (<https://web.archive.org/web/20130624013706/http://www.psypress.com/books/details/9780805854183/>). In Landauer, T.; McNamara, D; Dennis, S.; et al. *Handbook of Latent Semantic Analysis* (<http://www.psypress.com/books/details/9780805854183/>) (PDF). Psychology Press. ISBN 978-0-8058-5418-3. Archived from the original (<http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf>) (PDF) on 2013-06-24.
- Blei, D.M.; Lafferty, J.D. (2009). "Topic Models" (<http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf>) (PDF).
- Blei, D.; Lafferty, J. (2007). "A correlated topic model of *Science*". *Annals of Applied Statistics*. **1** (1): 17–35. arXiv:0708.3601 (<https://arxiv.org/abs/0708.3601>). doi:10.1214/07-AOAS114 (<https://doi.org/10.1214/07-AOAS114>).
- Mimno, D. (April 2012). "Computational Historiography: Data Mining in a Century of Classics Journals" (<http://www.perseus.tufts.edu/~amahoney/02-jocch-mimno.pdf>) (PDF). *Journal on Computing and Cultural Heritage*. **5** (1). doi:10.1145/2160165.2160168 (<https://doi.org/10.1145/2160165.2160168>).
- Marwick, Ben (2013). "Discovery of Emergent Issues and Controversies in Anthropology Using Text Mining, Topic Modeling, and Social Network Analysis of Microblog Content" (https://www.academia.edu/5508141/Discovery_of_Emergent_Issues_and_Controversies_in_Anthropology_Using_Text_Mining_Topic_Modeling_and_Social_Network_Analysis_of_Microblog_Content). In Yanchang, Zhao; Yonghua, Cen. *Data Mining Applications with R*. Elsevier. pp. 63–93.
- Jockers, M. 2010 Who's your DH Blog Mate: Match-Making the Day of DH Bloggers with Topic Modeling (<http://www.matthewjockers.net/2010/03/19/whos-your-dh-blog-mate-match-making-the-day-of-dh-bloggers-with-topic-modeling/>) Matthew L. Jockers, posted 19 March 2010
- Drouin, J. 2011 Foray Into Topic Modeling (<http://www.proustarchive.org/wp-trackback.php?p=60>) Ecclesiastical Proust Archive. posted 17 March 2011
- Templeton, C. 2011 Topic Modeling in the Humanities: An Overview (<http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/>) Maryland Institute for Technology in the Humanities Blog. posted 1 August 2011
- Griffiths, T.; Steyvers, M. (2004). "Finding scientific topics" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC387300/>). *Proceedings of the National Academy of Sciences*. **101** (Suppl 1): 5228–35. Bibcode:2004PNAS..101.5228G (<http://adsabs.harvard.edu/abs/2004PNAS..101.5228G>). doi:10.1073/pnas.0307752101 (<https://doi.org/10.1073/pnas.0307752101>). PMC 387300 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC387300/>). PMID 14872004 (<http://www.ncbi.nlm.nih.gov/pubmed/14872004>).
- Yang, T., A Torget and R. Mihalcea (2011) Topic Modeling on Historical Newspapers. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (<http://www.aclweb.org/anthology/W/W11/W11-15.pdf#page=108>). The Association for Computational Linguistics, Madison, WI. pages 96–104.
- Block, S. (January 2006). "Doing More with Digitization: An introduction to topic modeling of early American sources" (<http://www.common-place.org/vol-06/no-02/tales/>). *Common-place The Interactive Journal of Early American Life*. **6** (2).
- Newman, D.; Block, S. (March 2006). "Probabilistic Topic Decomposition of an Eighteenth-Century Newspaper" (http://www.ics.uci.edu/~newman/pubs/JASIST_Newman.pdf) (PDF). *Journal of the American Society for Information Science and Technology*. **57** (5). doi:10.1002/asi.20342 (<https://doi.org/10.1002/asi.20342>).

External links

- Mimno, David. "Topic modeling bibliography" (<http://mimno.infosci.cornell.edu/topics.html>).
- Brett, Megan R. "Topic Modeling: A Basic Introduction" (<http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/>). Journal of Digital Humanities.
- Topic Models Applied to Online News and Reviews (<https://www.youtube.com/watch?v=1wcX4fEdNUo>) Video of a Google Tech Talk presentation by Alice Oh on topic modeling with LDA
- Modeling Science: Dynamic Topic Models of Scholarly Research (<https://www.youtube.com/watch?v=8nBE5Qm8y6I>) Video of a Google Tech Talk presentation by David M. Blei
- Automated Topic Models in Political Science (<http://vimeo.com/13597441>) Video of a presentation by Brandon Stewart at the Tools for Text Workshop (<http://toolsfortext.wordpress.com/>), 14 June 2010
- Shawn Graham, Ian Milligan, and Scott Weingart "Getting Started with Topic Modeling and MALLET" (<http://programminghistorian.org/lessons/topic-modeling-and-mallet/>). The Programming Historian.
- Blei, David M. "Introductory material and software" (<https://web.archive.org/web/20121002061418/http://www.cs.princeton.edu/~blei/topicmodeling.html>)
- code (https://github.com/AmazaspShumik/sklearn-bayes/blob/master/skbayes/decomposition_models/gibbs_lda_cython.pyx), demo (https://github.com/AmazaspShumik/sklearn-bayes/blob/master/ipynotebooks_tutorials/decomposition_models/example_lda.ipynb) - example of using LDA for topic modelling
- demo (<http://elcid.demon.nl/form.html>) - Hierarchical topic extraction using compressor-based similarity measures

Retrieved from "https://en.wikipedia.org/w/index.php?title=Topic_model&oldid=851983471"

This page was last edited on 25 July 2018, at 20:59 (UTC).

Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. By using this site, you agree to the [Terms of Use](#) and [Privacy Policy](#). Wikipedia® is a registered trademark of the [Wikimedia Foundation, Inc.](#), a non-profit organization.