Topic model

From Wikipedia, the free encyclopedia

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. 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. 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.

Although topic models were first described and implemented in the context of natural language processing, they have applications in other fields such as bioinformatics.

Contents

History

An early topic model was described by Papadimitriou, Raghavan, Tamaki and Vempala in 1998. [1] Another one, called Probabilistic latent semantic indexing (PLSI), was created by Thomas Hofmann in 1999.[2] Latent Dirichlet allocation (LDA), perhaps the most common topic model currently in use, is a generalization of PLSI developed by David Blei, Andrew Ng, and Michael I. Jordan in 2002, allowing documents to have a mixture of topics.[3] 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.

Case studies

Templeton's survey of work on topic modeling in the humanities grouped previous work into synchronic and diachronic approaches. The synchronic approaches identify topics at a certain time, for example, Jockers used topic modelling to classify 177 bloggers writing on the 2010 'Day of Digital Humanities' and identify the topics they wrote about for that day. Meeks modeled 50 texts in the Humanities Computing/Digital Humanities genre to identify self-definitions of scholars working on digital humanities and visualize networks of researchers and topics. Drouin examined Proust to identify topics and show them as a graphical network

Diachronic approaches 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. Blevins has been topic modeling Martha Ballard's diary to identify thematic trends across the 27-year diary. 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.

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.[4] Several groups of researchers starting with Papadimitriou et al.[1] have attempted to design algorithms with probable guarantees. Assuming that the data was 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), the method of moments, and very recently an algorithm based upon non-negative matrix factorization (NMF). This last algorithm also generalizes to topic models that allow correlations among topics. [5]

See also

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

Software / Libraries

- 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/)

References

1. Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh (1998). "Latent Semantic Indexing: A probabilistic analysis" (Postscript). *Proceedings of ACM PODS*.
2. Hofmann, Thomas (1999). "Probabilistic Latent Semantic Indexing" (PDF). *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*.
3. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John (January 2003). "Latent Dirichlet allocation". *Journal of Machine Learning Research* **3**: 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.
4. Blei, David M. (April 2012). "Introduction to Probabilistic Topic Models" (PDF). *Comm. ACM* **55** (4): 77–84. doi:10.1145/2133806.2133826.
5. Sanjeev Arora; Rong Ge; Ankur Moitra (April 2012). "Learning Topic Models—Going beyond SVD". arXiv:1204.1956.

External links

- Mimno, David. "Topic modeling bibliography".
- Templeton, Clay. "Topic Modeling in the Humanities: An Overview". Maryland Institute for Technology in the Humanities.
- Brett, Megan R. "Topic Modeling: A Basic Introduction". Journal of Digital Humanities.
- Topic Models Applied to Online News and Reviews (http://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 (http://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". The Programming Historian.
- Blei, David M. "Introductory material and software" (http://www.cs.princeton.edu/~blei/topicmodeling.html)
- jLDADMM (http://sourceforge.net/projects/jldadmm/) A Java package for topic modeling on normal or short texts. jLDADMM includes implementations of LDA and the *one-topic-per-document* Dirichlet Multinomial Mixture model (i.e. mixture-of-unigrams). jLDADMM also provides an implementation for document clustering evaluation to compare topic models.

Further reading

- Steyvers, Mark; Griffiths, Tom (2007). "Probabilistic Topic Models" (PDF). In Landauer, T.; McNamara, D; Dennis, S.; et al. *Handbook of Latent Semantic Analysis* (PDF). Psychology Press. ISBN 978-0-8058-5418-3.
- Blei, D.M.; Lafferty, J.D. (2009). "Topic Models" (PDF).
- Blei, D.; Lafferty, J. (2007). "A correlated topic model of *Science*". *Annals of Applied Statistics* **1** (1): 17–35. doi:10.1214/07-AOAS114.

- Mimno, D. (April 2012). "Computational Historiography: Data Mining in a Century of Classics Journals" (PDF). *Journal on Computing and Cultural Heritag* **5** (1). doi:10.1145/2160165.2160168.
- 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
- Meeks, E. 2011 Comprehending the Digital Humanities (https://dhs.stanford.edu/comprehending-the-digital-humanities/) Digital Humanities Specialist, posted 19 February 2011
- 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". *Proceedings of the National Academy of Sciences* **101** (Suppl 1): 5228–35. doi:10.1073/pnas.0307752101. PMC 387300. PMID 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". *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" (PDF). *Journal of the American Society for Information Science and Technology* **57** (5). doi:10.1002/asi.20342.
- Blevin, C. 2010. Topic Modeling Martha Ballard’s Diary (http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/) historying. posted 1 April 2010.

Categories: Statistical natural language processing | Latent variable models