Topic Modeling Reading List

Introductory & Survey Papers

- Blei, D.M. (2012). Probabilistic topic models. Communications of the ACM, 55(4): 77-84.
- Grimmer, J. and Stewart, B.M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21(3): 267-297.
- Blei, D.M., and Lafferty, J. (2009). Topic models. Text Mining: Theory and Applications.
- Roberts, M.E., Stewart, B.M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. In Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.
- Dou, W. and Liu, S. (2016). Topic- and time-oriented visual text analysis. IEEE Computer Graphics and Applications, 36(4): 8-13.

Seminal Papers

- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. H. (1990). Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6): 391-407.
- Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3: 993-1022.
- Landauer, T., and Dumais, S. (1997). A Solution to Platos Problem: The Latent Semantic Analysis of Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104(2): 211-240.
- Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis, Machine Learning, 42(1): 177-196.

Pre-Processing

- Schofield, A., & Mimno, D. (2016). Comparing Apples to Apple: The Effects of Stemmers on Topic Models. Transactions of the Association for Computational Linguistics, 4: 287-300.
- Schofield, A., Thompson, L., and Mimno, D. (2017). Quantifying the Effects of Text Duplication on Semantic Models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2727-2737.
- Denny, M.J., and Spirling, A. (2017). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. Available at SSRN: https://ssrn.com/abstract=2849145 or http://dx.doi.org/10.2139/ssrn.2849145
- Schofield, A., Magnusson, M., and Mimno, D. (2017). Pulling Out the Stops: Rethinking Stopword Removal for Topic Models. EACL.
- Schofield, A., Magnusson, M., Thompson, L., and Mimno, D. (2017). Understanding Text Pre-Processing for Latent Dirichlet Allocation. ACL Workshop for Women in NLP (WiNLP).

Evaluation & Validation

- Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D.M. (2009). Reading tea leaves: How humans interpret topic models. Neural Information Processing Systems.

- Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In Proc. of the 26th International Conference on Machine Learning.
- Wallach, H., Mimno, D., and McCallum, A. (2009). Rethinking LDA: why priors matter. In Neural Informational Processing Systems.
- Mimno, D., Wallach, H.M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing, 262-272.
- Bischof, J., and Airoldi, E.M. (2012). Summarizing topical content with word frequency and exclusivity. In ICML, 201208.

Model Extensions

- Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2004). The author-topic model for authors and documents. UAI '04 Proc. of the 20th conference on Uncertainty in Artificial Intelligence: 487-494.
- Blei, D.M., and Lafferty, J. (2006). Dynamic topic models. In Proc. of the 23rd International Conference on Machine Learning, ACM.
- Blei, D.M., & Lafferty, J. D. (2007). A correlated topic model of science. The Annals of Applied Statistics, 17-35.
- Mcauliffe, J. D., & Blei, D.M. (2008). Supervised topic models. In Advances in neural information processing systems (pp. 121-128).
- Mimno, D., & McCallum, A. (2012). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. arXiv preprint arXiv:1206.3278.

Social Science Applications

- Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., and Radev, D.R. (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Political Science, 54 (1): 209-228.
- Grimmer, J (2010). A Bayesian hierarchical topic model for political texts: measuring expressed agenda in Senate press releases, Political Analysis, 18 (1).
- Paul, M.J., and Dredze, M (2014). Discovering health topics in social media using topic models, PLoS ONE, 9(8), e103408.
- Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., and Rand, D.G. (2014). Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science, 58(4), 1064-1082.
- Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. Political Analysis, 23(2), 254-277.
- Baumer, E.P., Mimno, D., Guha, S., Quan, E., & Gay, G.K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?. Journal of the Association for Information Science and Technology, 68(6), 1397-1410.
- Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihok, G., Den Hartog, D.N. (2017). Text Mining in Organizational Science. Organizational Research Methods.

Causal Inference

- Roberts, M.E., Stewart, B.M., and Airoldi, E.M. (2016). A model of text for experimentation in the social sciences. Journal of the American Statistical Association, 111(515), 988-1003.
- Roberts, M.E. Stewart, B.M., and Tingley, D. (2016). Navigating the Local Modes of Big Data: The Case of Topic Models. In Data Analytics in Social Science, Government, and Industry. New York: Cambridge University Press.
- Fong, C.J., and Grimmer, J. (2016). Discovery of Treatments from Text Corpora, In Proceedings of the Annual Meeting of the Association for Computational Linguistics.
- Egami, N., Fong, C.J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2017). How to Make Causal Inferences Using Texts.

Word Embedding & Deep Learning

- Mikolov, T., Chai, K., Corrado, G.S., and Dean, J. (2013). Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.
- Pennington, J., Socher, R., and Manning, C.D. (2014). Glove: Global Vectors for Word Representation, In EMNLP, 14, 1532–1543.
- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model, In JMLR, 3, 1137–1155, 2003.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning, Nature, 521 (7553), 436–444.

Code/Packages

- McCallum, A. (2002). MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.
- Rehurek, R., and Sojka, P. (2010). Software framework for topic modelling with large corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45-50. https://radimrehurek.com/gensim/
- Roberts, M., Stewart, B., and Tingley, D. (2017). stm: R package for structural topic models. R package version 1.3.0. http://www.structuraltopicmodel.com/

Code Tutorials

- Silge, J., and Robinson, D. (2017). Tidy Topic Modeling. https://cran.r-project.org/web/packages/tidytext/vignettes/topic_modeling.html.
- Wesslen, R. (2017). Topic Modeling workshop with R. https://github.com/wesslen/Topic-Modeling-Workshop-with-R.