2024 SIDE Summer School of Econometrics

"Econometrics meets Natural Language Processing: from Topic Analysis to Large Language Models"

José Luis Montiel Olea (Cornell University) and Jordan Lee Boyd-Graber (University of Maryland)

Venue: SADIBA Center, Perugia (Italy)

Dates: July 22^{n1d} – July 26th 2024

References:

Freyaldenhoven, S., Ke, S., Li, D., & Olea, J. L. M. (2023). On the testability of the anchor words assumption in topic models. Technical report, working paper, Cornell University.

Ke, S., Olea, J. L. M., & Nesbit, J. Robust Machine Learning Algorithms for Text Analysis. Working paper, 2022

Blei, D. M. (2012), 'Probabilistic topic models', Communications of the ACM 55(4), 77–84.

Blei, D. M. & Lafferty, J. D. (2009), 'Topic models', Text mining: classification, clustering, and applications 10(71), 71–89.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent Dirichlet allocation', Journal of Machine Learning research 3, 993–1022.

Boyd-Graber, J., Hu, Y., Mimno, D. et al. (2017), 'Applications of topic models', Foundations and Trends® in Information Retrieval 11(2-3), 143–296.

Bybee, L., Kelly, B. T., Manela, A. & Xiu, D. (2021), Business news and business cycles, Technical report, National Bureau of Economic Research.

Bybee, L., Kelly, B. T. & Su, Y. (2022), 'Narrative asset pricing: Interpretable systematic risk factors from news text', Johns Hopkins Carey Business School Research Paper (21-09).

Chen, Y., He, S., Yang, Y. & Liang, F. (2022), 'Learning topic models: Identifiability and finite-sample analysis', Journal of the American Statistical Association pp. 1–16.

Ke, Z. T. & Wang, M. (2022), 'Using svd for topic modeling', Journal of the American Statistical Association pp. 1–16.

Wang, X., Zhu, W. & Wang, W. Y. (2023), 'Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning', arXiv preprint arXiv:2301.11916 .

Ishani Mondal, Shwetha S, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Lee Boyd-Graber. Presentations by the People, for the People: Harnessing LLMs for Generating Persona-Aware Slides from Documents. European Association for Computational Linguistics, 2024

Zongxia Li, Andrew Mao, Daniel Kofi Stephens, Pranav Goel, Emily Walpole, Juan Francisco Fung, Alden Dima, and Jordan Lee Boyd-Graber. TENOR: Topic Enabled Neural Organization and Recommendation: Evaluating Topic Models in Task Based Settings—European Association for Computational Linguistics, 2024.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence. Neural Information Processing Systems, 2021.

Francesco Saverio Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. ClimaText: A Dataset for Climate Change Topic Detection. NeurIPS Workshop on Tackling Climate Change with Machine Learning, 2020.

Diggelmann, Thomas, Boyd-Graber, Jordan, Bulian, Jannis, Ciaramita, Massimiliano, and Leippold, Markus. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. NIPS Workshop on Tackling Climate Change with Machine Learning, 2020.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. Applications of Topic Models. 2017.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad (Jude) Alkhouja. Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce. ACM International Conference on World Wide Web, 2012.

Gillis, Nicolas, and Robert Luce. "Checking the Sufficiently Scattered Condition using a Global Non-Convex Optimization Software." arXiv preprint arXiv:2402.06019 (2024).

Chen, Yinyin, et al. "Learning topic models: Identifiability and finite-sample analysis." Journal of the American Statistical Association 118.544 (2023): 2860-2875.

Bing, Xin, Florentina Bunea, and Marten Wegkamp. "Optimal estimation of sparse topic models." Journal of machine learning research 21.177 (2020): 1-45.

Bing, Xin, et al. "Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations." *The Annals of Statistics* 50.6 (2022): 3307-3333.

Wang, X., Zhu, W., Saxon, M., Steyvers, M., & Wang, W. Y. (2024). Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, *36*.

Zhang, Liyi, R. Thomas McCoy, Theodore R. Sumers, Jian-Qiao Zhu, and Thomas L. Griffiths. "Deep de Finetti: Recovering Topic Distributions from Large Language Models." *arXiv preprint arXiv:2312.14226* (2023).