Unstructured Data for Economics

Stephen Hansen, stephen.hansen@ucl.ac.uk

Over the past decade, the use of unstructured data has been growing steadily in economics and related disciplines. This course covers the core empirical methods used to extract economically meaningful information from such data, with a focus on natural language. We first cover methods that operate on raw words counts across documents. Next, we introduce neural language models which represent words as vectors constructed to complete word prediction tasks. Finally, we discuss how such word prediction tasks form the basis for modern large language models. During the course, we will show how these models can be used to extract information from non-textual data such as surveys and financial transactions. Time permitting, we will also discuss issues arising from using the output of such models to conduct inference in econometric models.

There is no one source that covers all of the material in the course. Gentzkow et al. (2019a) and Ash and Hansen (2023) are survey articles that provide accessible introductions to natural language processing in economics. Jurafsky and Martin (2023) (which I call JM below) is in draft form with publicly available chapters at `https://web.stanford.edu/~jurafsky/slp3/`. Below I provide readings for each theme, where readings in green are background material from the computer science and machine learning literatures.

1 Document-Term Matrix

Dictionary Methods

- Tetlock (2007)
- Loughran and Mcdonald (2011)
- Baker et al. (2016)
- Hassan et al. (2019)

Document Similarity

- Hoberg and Phillips (2010, 2016)
- Kelly et al. (2021)

Text Regression

- Taddy (2013, 2015)
- Gentzkow et al. (2019b)

Dimensionality Reduction of Doc-Term Matrix

- Deerwester et al. (1990)
- Blei et al. (2003)
- Hansen et al. (2018)
- Bandiera et al. (2020)
- Draca and Schwarz (2021)

2 Word Embeddings

Word2Vec

- Mikolov et al. (2013a,b)
- JM Chapter 6
- Kozlowski et al. (2019)
- Gennaro and Ash (2022)
- Ash et al. (2024)

Embedding Products and Firms

- Magnolfi et al. (2023)

3 Large Language Models

- JM Chapter 10
- Vaswani et al. (2017)
- Devlin et al. (2019)
- Brown et al. (2020)

4 Finetuning Large Language Models

Illustrative case study taken from Hansen et al. (2023).

Self-Supervised FT

- Hinton et al. (2015)
- Sanh et al. (2020)

Supervised FT

- Hu et al. (2021)
- Shapiro et al. (2022)
- Gorodnichenko et al. (2023)
- Bybee (2023)

Instruction FT and Reinforcement Learning with Human Feedback

- Stiennon et al. (2020)
- Ouyang et al. (2022)

5 Econometrics of Unstructured Data

- Battaglia et al. (2024)

References

Ash, E., Chen, D. L., and Ornaghi, A. (2024). Gender Attitudes in the Judiciary: Evidence from US Circuit Courts. *American Economic Journal: Applied Economics*, 16(1):314–350.

Ash, E. and Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, 15(1):659–688.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). CEO Behavior and Firm Performance. *Journal of Political Economy*, 128(4):1325–1369.

Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2024). Inference for Regression with Variables Generated from Unstructured Data. https://arxiv.org/abs/2402.15585v3.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(null):993–1022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners.

Bybee, J. L. (2023). The Ghost in the Machine: Generating Beliefs with Large Language Models. Working Paper.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Draca, M. and Schwarz, C. (2021). How Polarized are Citizens? Measuring Ideology from the Ground-Up. SSRN Scholarly Paper ID 3154431, Social Science Research Network, Rochester, NY.

Gennaro, G. and Ash, E. (2022). Emotion and Reason in Political Language. *The Economic Journal*, 132(643):1037–1059.

Gentzkow, M., Kelly, B., and Taddy, M. (2019a). Text as Data. *Journal of Economic Literature*, 57(3):535–574.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340.

Gorodnichenko, Y., Pham, T., and Talavera, O. (2023). The Voice of Monetary Policy. *American Economic Review*, 113(2):548–584.

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., and Taska, B. (2023). Remote Work across Jobs, Companies, and Space.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Hassan, T. A., Hollander, S., van Lent, L., and Tahoun, A. (2019). Firm-Level Political Risk: Measurement and Effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the Knowledge in a Neural Network.

Hoberg, G. and Phillips, G. (2010). Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. *The Review of Financial Studies*, 23(10):3773–3811.

Hoberg, G. and Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models.

Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing*. 3rd edition.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring Technological Innovation over the Long Run. *American Economic Review: Insights*, 3(3):303–320.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5):905–949.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Magnolfi, L., McClure, J., and Sorensen, A. T. (2023). Triplet Embeddings for Demand Estimation.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics*, 228(2):221–243.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. (2020). Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 3008–3021, Red Hook, NY, USA. Curran Associates Inc.

Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Taddy, M. (2015). Distributed Multinomial Regression. *The Annals of Applied Statistics*, 9(3):1394–1414.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3):1139–1168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.