Unstructured Data in Empirical Economics

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

1 Textbooks / Overview Material

Over the past decade, the use of unstructured data has been growing steadily in economics and related disciplines, with a rapid acceleration in the wake of COVID-19. This course will begin with an overview of the challenges and opportunities of working with such data with a focus on natural language. We first review relatively straightforward methods that operate on raw word counts across documents before studying machine learning algorithms for dimensionality reduction, which is a key problem in the analysis of text and unstructured data. These encompass factor models whose basic structure is similar to well-known econometric methods, as well as neural network models which form the basis of much of modern natural language processing. Finally, we show how these ideas can be applied to non-textual data such as surveys, images, and scanner data on goods purchases.

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 text mining. Manning et al. (2008) is an information retrieval textbook that is referenced below as MRS. Below I provide readings for each of the lectures, where readings in green are background material from the computer science and machine learning literatures.

2 Introduction and the Document-Term Matrix

Background

- MRS 1, 2.2, 6.1-6.3

Detecting concepts in documents

- Tetlock (2007)
- Loughran and Mcdonald (2011)
- Baker et al. (2016)
- Shapiro et al. (2020)

How concepts relate in documents

- Hassan et al. (2019)

Measuring document similarity

- Hoberg and Phillips (2010, 2016)
- Cagé et al. (2020)
- Kelly et al. (2021)

Relating text to metadata

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

3 Dimensionality Reduction of Doc-Term Matrix

- MRS 18
- Deerwester et al. (1990)
- Blei et al. (2003)
- Hansen et al. (2018)
- Mueller and Rauh (2018)
- Larsen and Thorsrud (2019)

4 Word Embedding Models

- Goldberg (2016)
- Mikolov et al. (2013a,b)
- Dieng et al. (2020)
- Kozlowski et al. (2019)
- Ash et al. (2020)
- Ruiz et al. (2020)

5 Sequence Embedding Models

- Vaswani et al. (2017)
- Devlin et al. (2019)
- Phuong and Hutter (2022)
- Hansen et al. (2023)

6 Image Data

- Chollet et al. (2022)
- James et al. (2021)
- Jean et al. (2016)
- Ash et al. (2021)

References

Ash, E., Chen, D. L., and Ornaghi, A. (2020). Gender attitudes in the judiciary : Evidence from U.S. circuit courts. https://warwick.ac.uk/fac/soc/economics/research/workingpapers/2020/twerp_1256_-_ornaghi.pdf.

Ash, E., Durante, R., Grebenshchikova, M., and Schwarz, C. (2021). Visual Representation and Stereotypes in News Media.

Ash, E. and Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15(1):null.

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

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

Cagé, J., Hervé, N., and Viaud, M.-L. (2020). The Production of Information in an Online World. *The Review of Economic Studies*, 87(5):2126–2164.

Chollet, F., Kalinowski, T., and Allaire, J. J. (2022). *Deep Learning with R.* Manning, Shelter Island, NY, second edition.

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.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

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.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57(1):345–420.

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.

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.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer, New York NY, 2nd ed. 2021 edition edition.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

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.

Larsen, V. H. and Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1):203–218.

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.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, illustrated edition edition.

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]*.

Mueller, H. and Rauh, C. (2018). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, 112(2):358–375.

Phuong, M. and Hutter, M. (2022). Formal Algorithms for Transformers.

Ruiz, F. J. R., Athey, S., and Blei, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1):1–27.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*.

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.