Text Mining for Economics and Finance

Stephen Hansen, stephen.hansen@economics.ox.ac.uk

1 Reading

Manning et al. 2009 (MRS) and Murphy 2012 (KM) contain all material relevant for the statistical ideas from the course (and much more). I will provide lecture notes, so purchasing these is not a requirement, although I do provide relevant references below. (An HTML version of MRS is available for free online). Grimmer and Stewart (2013), Bholat et al. (2015), and Gentzkow et al. (2017) provide accessible introductions to text mining and machine learning.

2 Document-Term Matrix

- MRS 1, 2.2

3 Information Retrieval

Statistical theory:

- MRS 6.1-6.3

Applications:

- Baker et al. (2016)
- Tetlock (2007)
- Loughran and Mcdonald (2011)
- Hoberg and Phillips (2010)
- Friebel and Heinz (2014)

4 Unsupervised Learning

4.1 Probability models for discrete data

Statistical theory:

- KM 2.5.4, 3.3-3.4

4.2 Finite mixture models and EM algorithm

Statistical theory:

- KM 11

4.3 Singular value decomposition

Statistical theory:

- MRS 18
- Deerwester et al. (1990)

Applications:

- Boukus and Rosenberg (2006)
- Hendry and Madeley (2010)
- Acosta (2014)

4.4 Latent Dirichlet allocation

Statistical theory:

- KM 27.1-27.3.2, 27.3.1-27.3.6; 21
- Blei et al. (2003)
- Blei and Lafferty (2009)
- Wainwright and Jordan (2008)

Applications and extensions:

- Quinn et al. (2010)
- Hansen et al. (2014)
- Hansen and McMahon (2015)
- Mueller and Rauh (2016)
- Blei and Lafferty (2006)
- Roberts et al. (2016)

5 Supervised Learning

5.1 Discriminative models

Statistical theory:

- KM 13
- Meinshausen and Bühlmann (2010)
- Belloni et al. (2014b)

Applications:

- Belloni et al. (2014a)

5.2 Generative models

Statistical theory:

- MRS 13
- Mcauliffe and Blei (2008)
- Taddy (2013)
- Taddy (2015)

Applications:

- Gentzkow and Shapiro (2010)

References

Acosta, J. M. (2014). FOMC responses to calls for transparency: Evidence from the minutes and transcripts using latent semantic analysis. Mimeograph, University of Stanford.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*. forthcoming.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on Treatment Effects after Selection among High-Dimensional Controls. *Review of Economic Studies*, 81(2):608–650.

Bholat, D., Hansen, S., Santos, P., and Schonhardt-Bailey, C. (2015). Text mining for central banks. Centre for Central Banking Studies, Handbook No. 33, Bank of England.

Blei, D. and Lafferty, J. (2006). Dynamic topic models. In *Proceedings of the 23rd Internaional Conference on Machine Learning*, pages 113–120.

Blei, D. and Lafferty, J. (2009). Topic models. In Srivastava, A. and Sahami, M., editors, *Text Mining: Classification, Clustering, and Applications*. Taylor & Francis, London, England.

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

Boukus, E. and Rosenberg, J. V. (2006). The information content of FOMC minutes. Mimeograph, Federal Reserve Bank of New York.

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.

Friebel, G. and Heinz, M. (2014). Media slant against foreign owners: Downsizing. *Journal of Public Economics*, 120(C):97–106.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as Data. NBER Working Papers 23276, National Bureau of Economic Research, Inc.

Gentzkow, M. and Shapiro, J. M. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71.

Grimmer, J. and Stewart, B. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, pages 1–31.

Hansen, S. and McMahon, M. (2015). Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication. *Journal of International Economics*. forthcoming.

Hansen, S., McMahon, M., and Prat, A. (2014). Transparency and Deliberation within the FOMC: a Computational Linguistics Approach. CEPR Discussion Papers 9994, C.E.P.R. Discussion Papers.

Hendry, S. and Madeley, A. (2010). Text mining and the information content of bank of canada communications. Working Paper 2010-31, Bank of Canada.

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

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

Manning, C. D., Raghavan, P., and Shütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

Mcauliffe, J. D. and Blei, D. M. (2008). Supervised Topic Models. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Inc.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B*, 72(4):417–473.

Mueller, H. and Rauh, C. (2016). Reading between the lines: Prediction of political violence using newspaper text.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. MIT Press.

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.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A Model of Text for Experiments in the Social Sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108.

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. *Journal of Finance*, 62(3):1139–1168.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.