Text Analysis for Economists

Instructor: Ruben Durante (ICREA-Universitat Pompeu Fabra, IPEG, Barcelona GSE)
Day and Time: TBD
Location: online

Course objective

An ever-increasing share of human communication is recorded as digital text. Analysing and making sense of this vast amount of data is increasingly important for research in the social sciences. This course provides an accelerated introduction to the theory and practice of text analysis by surveying methods for systematically extracting quantitative information from text, from classical content analysis and dictionary-based methods, to classification methods, scaling methods, and topic models. The course introduces the theoretical foundations for text analysis but mainly takes a practical approach, illustrating the methods through state-of-the-art applications to research questions in economics, political science, and finance. Lectures will be complemented with hands-on exercises working with text data in Python.

Main readings

In addition to the course lecture slides, the course will rely heavily on the following readings:

- Gentzkow, M., Kelly, B. T., and Taddy, M., 2017, *Text as Data*. NBER Working Paper #23276.
- Grimmer, J. and Stewart, B., 2013, *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*, Political Analysis, vol. 21, n. 3, pp. 267-297.
- Manning, C. D., Raghavan, P., and Shutze, H., 2008, *An Introduction to Information Retrieval,* Cambridge University Press.
- Bengfort, B., Ojeda, T., Bilbro, R., *Applied Text Analysis with Python,* 2018, O'Reilly Media.
- Krippendorff, K., 2013, *Content Analysis: An Introduction to Its Methodology*, Sage.

Programming

For all course exercises we will use Python. Specifically we will use the Jupyter Notebook online code editor, an application available through the Anaconda platform (link).

Assessment

Class participation and practical assignments (30%): all students are expected to attend and actively participate in the lectures and to complete the practical assignments.

In-class presentation (20%): each student is expected to present once during the course a research paper of his choice, but relevant to the object of the course. Presentations can be individual or in group (max 3 people) and last no longer than 15-20 minutes.

Research proposal (50%): a research proposal of about 10-15 pages is due at the end of the course. This will be an original analysis of texts using some of the methods covered in class, and may focus on extending a published work. Ideally this should be something that the student can use for her/his thesis/own research.

Course schedule

1. From text to data
    - Introduction to documents, metadata, corpora
    - Pre-processing: tokenizing, stemming, stop-words
    - Word counts, document-feature matrix
    - Collocation and n-grams

2. Important notions
    - TF-IDF
    - Measuring text length, diversity, and complexity
    - Measuring similarity between documents

3. Machine Learning
    - Introduction to statistical learning theory
    - Supervised and unsupervised learning

4. Statistical methods
    - Dictionary-based methods
    - Penalized linear models
    - Dimension reduction and feature selection
    - Non-Linear text regression
    - Random forests

5. Generative language models: supervised methods
    - Naive Bayes
    - Support vector machines

- K-nearest neighbors

6. Generative language models: unsupervised methods
    - Latent semantic analysis
    - Topic models and Latent Dirichlet Allocation (LDA)
    - K-means clustering

7. Word Embeddings
    - Word2Vec
    - Doc2Vec

Applications

Antweiler, W. and Frank, M. 2004, *Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards*, The Journal of Finance, vol. 59, n.3, pp. 1259-1294.

Ash, E., Chen, D., and Naidu, S., 2018, *Ideas Have Consequences: The Impact of Law and Economics on American Justice,* Working Paper.

Baker, S. R., Bloom, N., and Davis, S. J., 2016, *Measuring Economic Policy Uncertainty*, Quarterly Journal of Economics, vol. 131, n.4, pp. 1593-1636.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R., 2020, *CEO Behavior and Firm Performance*, Journal of Political Economy, vol. 128, n. 4, pp. 1325-1369.

Barberá, P., Casas, A., Nagler, J., Egan, P., Bonneau, R., Jost, J., and Tucker, J., 2019, *Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data*, American Political Science Review, vol. 113, n. 4, pp. 883-901.

Benoit, K., Conway, D., Lauderdale, B., Laver, M., and Mikhaylov, S., 2016, *Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data*, American Political Science Review, vol. 110, n. 2, pp. 278-295.

Blei, D., Ng, A., and Jordan, M., 2003, *Latent Dirichlet Allocation,* Journal of Machine Learning Research, vol. 3, pp. 993–1022.

Blei, D., and Lafferty, D., 2006, *Dynamic Topic Models*, In Proceedings of the 23rd International Conference on Machine Learning.

Cage, J., Hervé, N., and Viaud, M.-L., *The Production of Information in an Online World*, forthcoming, Review of Economic Studies.

Djourelova, M., *Media persuasion through slanted language: Evidence from the coverage of immigration*, 2020, Job Market Paper

Durante, Ruben, Paolo Pinotti, and Andrea Tesei. 2019, *The Political Legacy of Entertainment TV*, American Economic Review, 109 (7): 2497-2530.

Gentzkow, M., Shapiro, J., and Taddy, M., 2016, *Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech*, NBER Working Paper.

Gentzkow, M. and Shapiro, J., 2009, *What Drives Media Slant?,* Econometrica, vol. 78, n.1, pp.35-71.

Groseclose, T. and Milyo, J., 2005, *A Measure of Media Bias*, Quarterly Journal of Economics, vol. 120, n. 4, pp. 1191-1237.

Hassan, T., S. Hollander, L. van Lent, and A. Tahoun, 2018, *Firm-Level Political Risk: Measurement and Effects*, Working Paper.

Hansen, S., McMahon, M., and Prat, A., 2018, *Transparency and Deliberation within the FOMC: a Computational Linguistics Approach*, Quarterly Journal of Economics, vol. 133, n. 2, pp. 801-870.

Kelly, Bryan T., Papanikolaou, D., Seru, A., and Taddy, M, *Measuring Technological Innovation Over the Long Run*, forthcoming, American Economic Review: Insights.

Jegadeesh, N. and Wu, D., 2013, *Word Power: A New Approach for Content Analysis*, Journal of Financial Economics, vol. 110, n. 3, pp. 712-729.

Loughran, T. and McDonald, B., 2011, *When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10K-s*, The Journal of Finance, vol. 66, n. 1, pp. 35-65.

Quinn, K., Monroe, B., Colaresi, M., Crespin, M. and Radev, D., 2010, *How to Analyze Political Attention with Minimal Assumptions and Costs*, American Journal of Political Science, vol. 54, n. 1, pp. 209-228

Stock, J., H. and Trebbi, F., 2003, *Retrospectives: Who Invented Instrumental Variable Regression?*, Journal of Economic Perspectives, vol. 17, n. 3, pp. 177-194

Tetlock, P. C., 2007, *Giving Content to Investor Sentiment: The Role of Media in the Stock Market*, Journal of Finance, vol. 62, n. 3, pp. 1139-1168.