Machine Learning Techniques for Social Science Texts and Datasets

Third-term workshop, May 26-27, 2017, 09:00-18:00, Seminar room 2, Badia Fiesolana

Guest instructor: Dr Slava Mikhaylov (UCL)

Preliminary programme

The workshop focuses on the application of supervised and unsupervised machine learning techniques to political texts and dataset. The goal of the workshop is to enable participants to apply machine learning to data collections of their own interest (e.g. parliamentary speeches, newspaper data or party manifestoes) and discuss the challenges they face while doing so. The focus of the workshop is practical, thus, more time is devoted to application than a general introduction. Furthermore, the discussion of students' research designs and first results is an important part of the workshop and participants are strongly encouraged to bring their own designs and data. The workshop discusses both unsupervised and supervised methods of machine learning, e.g. text scaling, topic models, and word embeddings.

All participants should install the latest version of R and RStudio, and try out an R Notebook project (that will install some necessary packages as well). The main text analytics package to install is "quanteda". Closer to the date, there may be a list of additional packages to install.

A general reading for those with less experience with machine learning is

- James et al. (2013) An Introduction to Statistical Learning: With applications in R, Springer. The book is available from the authors' page: http://www-bcf.usc.edu/~gareth/ISL/

Specialised readings for the sessions are indicated in the respective sessions

May 26th (9am - 5pm)

Introduction: : <~0.5 hour>

- When does which method/technique make most sense and how can we figure that out?

Pre-reading:

- Grimmer, J, and B M Stewart (2013), "Text as Data: the Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Political Analysis.

Block 1: <~0.5 hour>

- Replicability in social science (R Markdown, R Notebooks, GitHub)

Pre-reading:

- R Markdown: http://rmarkdown.rstudio.com

- R Notebooks: http://rmarkdown.rstudio.com/r_notebooks.html

- GitHub and RStudio: Hadley Wickham book “R packages”, chapter “Git and GitHub”: http://r-pkgs.had.co.nz/git.html

Block 2: <~2 hours>

- Text scaling models (Naive Bayes, PCA, CA)

- Coffee break

- Applications of text scaling models

Pre-reading:

- Laver, Michael, Kenneth Benoit and John Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” American Political Science Review 97: 311-331.
- Lowe, William. 2008. “Understanding Wordscores.” Political Analysis 16(4): 356-371.
- Greenacre, M. (2007). Correspondence Analysis in Practice, 2nd edition. Appendix A & B.

- Spirling, A. (2012), “U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784-1911.” American Journal of Political Science, 56: 84–97.
- Herzog, A. and K. Benoit (2015), “The most unkindest cuts: Speaker selection and expressed government dissent during economic crisis.” Journal of Politics, 77(4):1157–1175.

Block 3: <~3 hours>

- Topic models (LDA, CTM, STM)

- Lunch break (45-60min)

- Applications of topic models

- Coffee break

Pre-reading:

- David Blei (2012). “Probabilistic topic models.” Communications of the ACM, 55(4): 77-84.
- Blei, David, Andrew Y. Ng, and Michael I. Jordan (2003). “Latent dirichlet allocation.” Journal of Machine Learning Research 3: 993-1022.

- Blei, David (2014) “Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models.” Annual Review of Statistics and Its Application, 1: 203-232.
- Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson, and Rand (2014). “Structural topic models for open-ended survey responses.” American Journal of Political Science, 58(4): 1064-1082.

- Blei, D. and J. Lafferty “Topic Models.” In Text Mining: Classification, clustering, and applications, A. Srivastava and M. Sahami (eds.), pp 71-94, 2009. Chapter available here: http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf.

Block 4: <~2 hours>

- Word embeddings (word2vec, text2vec)

- Applications of word embeddings

- Coffee break

- Textual patterns and quality assessment (keyness, similarity, distance, readability)

- Applications of patterns and quality measures

Pre-reading:

- Mikolov, Tomas et al. “Efficient Estimation of Word Representations in Vector Space.”
- Goldberg, Yoav and Omer Levy “word2vec Explained: Deriving Mikolov et al.’s Negative-Sampling Word-Embedding Method.”

- Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S.; Dean, Jeff (2013). “Distributed representations of words and phrases and their compositionality.” Advances in Neural Information Processing Systems.
- Levy, Omer; Goldberg, Yoav; Dagan, Ido (2015). “Improving Distributional Similarity with Lessons Learned from Word Embeddings.” Transactions of the Association for Computational Linguistics.
- Pennington et al. “GloVe: Global Vectors for Word Representation.”
- Huang et al. “Improving Word Representations via Global Context and Multiple Word Prototypes.”

<<Evening exercise: apply text analytics to a research project (or toy data)>>

May 27 (10am - 12pm)

Block 5: <~2 hours>

- Prepare a presentation on participant's project with text analytics (from R Markdown). Can also be group work.

- Presentation and discussion of each project