Advanced Social Data Science 2 (ASDS2) Syllabus

Readings

The main textbook for the course is *Text Analysis in Python for Social Scientists* by Dirk Hovy (herafter referred to as Hovy), volumes 1 ('Discovery and Exploration') and 2 ('Prediction and Classification'). Both volumes are accessible via the Royal Library's online service (`rex.kb.dk`).

Some lectures will also rely on selected chapters from *Text as Data: A New Framework for Machine Learning and the Social Sciences* by Justin Grimmer, Margaret E. Roberts and Brandon Stewart (hereafter GRS). These chapters will be made available via Absalon.

Finally, the course also features article readings. Readings for each class meeting are listed below.

Teachers

The course is taught through a combination of lectures and exercises. The lectures are given by:

- Frederik Hjorth (FH)
- Clara Vandeweerdt (CV)
- Simon Polichinel van der Maase (SP)
- Mathias Rask (MR)

The exercise classes are covered by the course TAs:

- Christian Thomas Nielsen Garcia
- Jacob Aarup Dalsgaard
- Johan Ødum Lundberg
- Sofie Læbo Astrupgaard
- Vojtech Houska

Topics

The course covers six topics:

1. Getting Started (lecture 1-2, FH)
2. Discovery & Scaling (3-5, FH)
3. Classification (6-7, CV)
4. Neural networks (8-10, CV)
5. Applications in natural language (11-12, CV)
6. Non-textual Data (13-14, SP + MR)

Exam

The course exam is a 72-hour take-home exam.

Topic 1: Getting Started

April 17: Tokenization and regex

Teacher: FH

Keywords: Text as data sources, importing text data, regular expressions

Readings:

- Hovy vol. 1, sections 1 + 2.1.1 + 3 + 5-5.2.1

April 19: Preprocessing

Teacher: FH

Keywords: Stopping, stemming, lemmatization, bag-of-words

Readings:

- Hovy vol. 1, section 2
- Hovy vol. 2: section 3-3.1
- GRS: chapter 5 (Bag of Words)

Supplementary reading:

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2), 168-189.

Topic 2: Discovery & Scaling

April 24: Word discovery

Teacher: FH

Keywords: keyword selection, tf-idf

Readings:

- Hovy vol. 1, section 5.2.3

King, Gary, Patrick Lam and Margaret E. Roberts. 2017. “Computer-Assisted Keyword and Document Set Discovery from Unstructured Text.” *American Journal of Political Science* 61(4):971–988.

April 26: Topic models

Teacher: FH

Keywords: topic models, latent dirichlet allocation

Blei, David M. 2012. "Probabilistic Topic Models." Communications of the ACM 55(4):77–84.

- Hovy vol. 1, section 9

Terman, Rochelle. 2017. "Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage." *International Studies Quarterly* 613(3):489–502.

May 1: Scaling

Teacher: FH

Keywords: wordscores, wordfish

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2):311–331.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

[Note for lab to use WordFish in the lab: https://github.com/umanlp/SemScale]

Topic 3: Classification

May 3: Classification I

Teacher: CV

Keywords: logistic regression (recap), loss functions, train-test splits, performance metrics

Readings:

- Hovy vol. 2: sec. 2, 4, 5, 6 (Classification, labels, train-dev-test, performance metrics)

May 8: Classification II

Teacher: CV

Keywords: regularization (recap/advanced), cross-validation (recap), Support Vector Machines, Naive Bayes

Readings:

- Hovy vol. 2: sec. 8 (Overfitting & regulatization)
- GRS: sec. 19.1 (Naive Bayes)
- Visually Explained. Support Vector Machine (SVM) in 2 minutes.https://www.youtube.com/watch?v=_YPScrckx28

Supplementary reading:

- Visually Explained. [The Kernel Trick in Support Vector Machine (SVM).]https://www.youtube.com/watch?v=Q7vT0--5VII

Topic 4: Neural networks

May 10: Shallow neural nets

Teacher: CV

Keywords: network notation, logistic regression as a net, softmax and multi-category classification, gradient descent

Readings:

- Hovy vol. 2: sec. 13-13.3 (Shallow neural nets)
- 3Blue1Brown. But what is a neural network? https://www.youtube.com/watch?v=aircAruvnKk
- Visually Explained. Gradient Descent in 3 minutes. https://www.youtube.com/watch?v=qg4PchTECck

Supplementary reading:

- Jurafsky, D. & Martin., J. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/ Chapter 7: Neural Networks and Neural Language Models.

May 15: Deep neural nets and sequence models

Teacher: CV

keywords: backpropagation and forward propagation, language models, recurrent neural nets, LSTMs

Readings:

- Hovy vol. 2, sec. 13.4, 14.2 (Multilayer perceptrons, RNNs and LSTMs)
- Louis Serrano. A friendly introduction to RNNs. https://www.youtube.com/watch?v=UNmqTiOnRfg
- CodeEmporium. LSTM Networks - EXPLAINED! https://www.youtube.com/watch?v=QciIcRxJvsM

Supplementary reading:

- Jurafsky, D. & Martin., J. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/ Chapter 9: RNNs and LSTMs.

May 17: Attention, transformers and large language model caveats

Teacher: CV

keywords: attention, transformers, BERT, GPT, stochastic parrots

Readings:

- Hovy vol. 2, sec. 14.3-14.5 (Attention & transformers)
- AssemblyAI. Transformers for beginners | What are they and how do they work. https://www.youtube.com/watch?v=_UVfwBqcnbM
- AssemblyAI. What is BERT and how does it work? | A Quick Review. https://www.youtube.com/watch?v=6ahxPTLZxU8
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 610–623. https://doi.org/10.1145/3442188.3445922

Supplementary reading:

- Alammar, J. The Illustrated Transformer. https://jalammar.github.io/illustrated-transformer
- Jurafsky, D. & Martin., J. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/ Chapter 10: Transformers and Pretrained Language Models.
- CodeEmporium. Transformer Neural Networks Explained. https://www.youtube.com/watch?v=TQQlZhbC5ps
- CodeEmporium. BERT Neural Network - EXPLAINED! https://www.youtube.com/watch?v=xI0HHN5XKDo

Topic 5: Applications in natural language

May 22: Word embeddings

Teacher: CV

Keywords: word2vec, FastText, BERT embeddings, bias in embeddings

Readings:

- Hovy vol. 2, sec. 3.2 (Word embeddings)
- Hovy vol. 1: sec 5.3 (Distributed representations)
- GRS: chapter 8 (Distributed representations)
- Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. The Journal of Politics, 84(1), 101-115.

May 25: Sentence classification

Teacher: CV

keywords: sentiment analysis, transfer learning, bias in annotations

Readings:

- GRS: chapter 17-18
- Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524.
- Terechshenko, Z., Linder, F., Padmakumar, V., Liu, M., Nagler, J., Tucker, J. A., & Bonneau, R. (2020). A comparison of methods in political science text classification: Transfer learning language models for politics. Available at SSRN 3724644.

May 31: No class

Topic 6: Non-textual data

June 5: Images as data

Teacher: SP

Keywords: images as data, computer vision, neural networks

Readings:

- Torres, Michelle, and Francisco Cantú 2022. "Learning to see: Convolutional neural networks for the analysis of social science data."
- Williams, Nora Webb, Andreu Casas, and John D. Wilkerson 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification.* Cambridge University Press. p. 14-44
- Zeiler and Fergus 2013, Visualizing and understanding Convolutional Networks

June 7: Audio as data

Teacher: MR

Keywords: audio as data

Readings:

- Rheault, Ludovic, and Sophie Borwein. "Audio as data." *Elgar Encyclopedia of Technology and Politics.* Edward Elgar Publishing 86-90
- Dietrich, Bryce J., Matthew Hayes, and Diana Z. O'Brien. "Pitch perfect: Vocal pitch and the emotional intensity of congressional speech." *American Political Science Review* 113.4 (2019): 941-962.
- Knox, Dean, and Christopher Lucas. "A dynamic model of speech for the social sciences." *American Political Science Review* 115.2 (2021): 649-666