**Syllabus**

**Text Data in Economics**

**Warwick QAPEC Summer School**

**Summer 2022**

[**link to this page**](https://docs.google.com/document/d/14hFLKAwllLOHD45hS_xtGA3xWirNmcV9Bt0M_PNnxhA/edit?usp=sharing)

[**GitHub Repo with Slides, Notebooks, and Problem Sets**](https://github.com/elliottash/text_econ_2022)

**Instructor**: [Elliott Ash](http://elliottash.com/), ashe@ethz.ch  
**TA**: [Claudia Marangon](https://lawecon.ethz.ch/group/scientific-team/marangon.html), [claudia.marangon@gess.ethz.ch](mailto:claudia.marangon@gess.ethz.ch)

**Course Format**

* 10 lectures on zoom, recorded
* 4 TA sessions on zoom, recorded
* In-person workshopping of student project papers

**Assignments:**

* 3 problem sets based on the example notebooks (submit to Claudia by email)
* In-class presentation of a course reading, with a partner ([sign-up sheet](https://docs.google.com/spreadsheets/d/17NIga-7iqB9VGQx-79XbUHxt0D0jVAgchb5tqUCg1Wo/edit?usp=sharing))
* Referee report on one of the course readings (submit to Claudia by email by August 1st)
* Research proposal on a text-data project ([sign up here](https://docs.google.com/spreadsheets/d/1ook5lO24yDLU7GUZMB01iZSC3vs9pTj6gSthu8eMMRk/edit?usp=sharing) or email Claudia about it by 1st Aug, 1st draft by Sept 1st, 2nd draft by Nov 1st)

**Critical presentations**

* Done in pairs
* up to 10 minutes
* present and critique the following, focusing on the text methods:
  + research question
  + text-analysis methods
  + empirical methods
  + results
  + contribution

**Research proposal (optional)**

* Write a proposal (as if for a grant) for an economics paper using text as data.
* Individually or partners
* 10-20 pages
* Contribution to Literature, Data, Text Methods (including validation approach), Empirical Methods, Interpretation of Potential Results

Lecture Schedule and Recordings Links

| June 15th | 13h (UK) | [Lecture 1](https://ethz.zoom.us/rec/share/cELC6zc9IV2mhlA4PMq235JTDEfJ3gLXvYpNXRp_3CTI2tV_ggEu9Xn9k56W-WV6.69e9e7uUlAZoSf9F?startTime=1655294496000) |
| --- | --- | --- |
| June 17th | 10h | [Lecture 2](https://ethz.zoom.us/rec/share/4n8Z3978lsyvyqPnPh9Mc5NYApaiEVeO0kgkULrRQgWHOzNjts88i7PrhjgGMd1Z.LpptUfT95hd-RuB3?startTime=1655456558000) |
|  |  |  |
| June 20th | 10h | [Lecture 3](https://ethz.zoom.us/rec/share/pIJLQGl7Lp8r_UeCrf_zum9ELBQsxh1Frl6nvvWnhKMIE59TWXfzVhJjOhxdmOrx.1f19oTwKkpFzVkL_?startTime=1655715724000) |
| June 22nd | 13h | [Lecture 4](https://ethz.zoom.us/rec/share/kzRnYK3n3-DtDQ8LiFQIqOPbDbozDvULFbMmKUHVaJNq5zF3qHz1z7SQSCkMXYWk.Xg3QFxJhx8M75XAL?startTime=1655899440000) |
|  |  |  |
| June 29th | 15h | [Lecture 5](https://ethz.zoom.us/rec/share/CN5SqRbLUbboeN444T4GtzioV-OczjhMuC06wu5CsYTykBbVgTmg7ZiEXNOTExCY.X4ZV_spmS8AThBOl?startTime=1656511295000) |
|  |  |  |
| July 4th | 10h | [Lecture 6](https://ethz.zoom.us/rec/share/o-h339nVWhYjW1iOImU8KhPUCgv0LNSrALeybQyNalrWldFAblz31LdPuKhlSon4.5s4fHQreDlpwCn1R?startTime=1656925539000), [Presentation on GST](https://www.dropbox.com/s/40fvwe1hgmdg1vg/Pre_Junxi.mp4?dl=0) |
| July 6th | 10h | [Lecture 7](https://ethz.zoom.us/rec/share/_Cgq8ZRDt6wEWJ8vhHJD5p0OG6whW48K4Xprbl9nI1wT2zp_NJWIKYd879Ck5hrg.iOUQGZwiy4a2i-RS?startTime=1657098134000) |
| July 8th | 10h | [Lecture 8](https://ethz.zoom.us/rec/share/Sr8_sL_FYQ4f1-d6qLLV_yqNNGAkTqgMkTEY02I_S7WOZ834W6l5-_B3E8nwKtUA.VfcjFmM7BJ_iKgiu?startTime=1657271258000) |
|  |  |  |
| July 11th | 10h | [Lecture 9](https://ethz.zoom.us/rec/share/nOwKxK03zcADH1-voJASLl1JCdSGCHHvlBJdMqhheecFrB7TCF0hDPr1SprZccAz.LGoy0hKgKvx1I0Ea?startTime=1657530491000) |
| July 12th | 10h | (office hours) |

TA Session Schedule

| June 16th | 10am | OH for questions and problems with setup (optional) |
| --- | --- | --- |
| June 23rd | 10am | [TA Session 1](https://ethz.zoom.us/rec/share/LWo3hEHtJN0sXiSA9b3KHElrvJUkXSH2oSxFe0PgEUmww4sNYnPI8j6iRY-G2WVB.QtTZbM-ks58_KSMi) |
| July 5th | 10am | [TA Session 2](https://ethz.zoom.us/rec/share/FR9n1qE_n9e2L5kHlU33ZLTOwbuvQwf82foMhMz-y9bUzuoST7MRzYIKNnlOhM04.zYl9VUXYpmskqN42?startTime=1657011628000) |
| July 7th | 10am | [TA Session 3](https://ethz.zoom.us/rec/share/ntJvU9CxTy8BBRwXwKaN50zVwza7CeF6hakFu3uJ6Q90CN7UC2ckQGcpSM0ld9o.xX5-8WuWtj6aw8dL?startTime=1657184360000) |
| July 13th | 10am | [TA Session 4](https://ethz.zoom.us/rec/share/ZL2hO2XQcKfyIEyW0o2rg-Kgzz-oTGfmJh7FYCF-dGR2Ci3tZt85oAVurMkA5TM.VH5AyoU9ZNfAKr3t) |

Topics Outline and Main Economics Papers Readings

[Sign-up sheet for discussant presentations](https://docs.google.com/spreadsheets/d/17NIga-7iqB9VGQx-79XbUHxt0D0jVAgchb5tqUCg1Wo/edit?usp=sharing)

1. Overview
   1. Gentzkow, Kelly, and Taddy (2019), “[Text as Data](https://web.stanford.edu/~gentzkow/research/text-as-data.pdf).”
2. Style Features and Dictionaries
   1. Enke (2020), [Moral values and voting](https://www.journals.uchicago.edu/doi/full/10.1086/708857)
   2. Michalopoulous and Xue (2021), [Folklore](https://www.researchgate.net/profile/Melanie-Meng-Xue/publication/348914698_Folklore/links/6167e82b66e6b95f07c33916/Folklore.pdf)
3. Tokenization
   1. Gentzkow and Shapiro (2010), [What Drives Media Slant? Evidence from U.S. Daily Newspapers](https://web.stanford.edu/~gentzkow/research/biasmeas.pdf).
   2. Hassan, Hollander, Van Lent, and Tahoun (2019), [Firm-Level Political Risk: Measurement and Effects](https://academic.oup.com/qje/article-abstract/134/4/2135/5531768?redirectedFrom=fulltext)
4. Document Distance
   1. Kelly, Papanikolau, Seru, and Taddy, [Measuring technological innovation over the very long run](https://www.aeaweb.org/articles?id=10.1257/aeri.20190499).
   2. Cage, Herve, and Viaud (2019), [The production of information in an online world](https://juliacage.com/wp-content/uploads/2020/07/Cage-Herve-Viaud-2020.pdf)
   3. Bertrand, Bombardini, Fisman, Hackinen, and Trebbi (2021), [Hall of Mirrors: Corporate Philanthropy and Strategic Advocacy](https://academic.oup.com/qje/article-abstract/136/4/2413/6313294)
5. Topic Models
   1. Hansen, McMahon, and Prat (2017), [Transparency and deliberation with the FOMC: A computational linguistics approach](https://academic.oup.com/qje/article/133/2/801/4582916).
   2. Ash, Morelli, and Vannoni (2022), “[More laws, more growth? Evidence from U.S. states](https://elliottash.com/wp-content/uploads/2022/04/Ash-Morelli-Vannoni-Laws-Growth-2022-04v4.pdf)”
   3. Djourelova, Durante, and Martin (2021), [The impact of online competition on local newspapers](https://web.stanford.edu/~gjmartin/papers/Craigslist_Draft_November_2021.pdf).
6. Supervised Learning
   1. Gentzkow, Shapiro, and Taddy (2019), [Measuring group differences in high-dimensional choices: Method and application to Congressional Speech](https://scholar.harvard.edu/files/shapiro/files/politext.pdf)
   2. Widmer, Galletta, and Ash (2022), [Media Slant is Contagious](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3712218)
7. Word Embeddings
   1. Ash, Chen, and Ornaghi (2022), “[Gender attitudes in the judiciary: Evidence from U.S. Circuit Courts](https://elliottash.com/wp-content/uploads/2022/04/Ash-Chen-Ornaghi-2022-04.pdf)”
   2. Ash, Gennaro, Hangartner, and Stampi-Bombelli (2022), “Immigration and Social Distance: Evidence from Newspapers during the Age of Mass Migration”.
8. Linguistic Parsing
   1. Antoniak, Mimo, and Levy (2019), [Narrative paths and negotiation of power in birth stories](https://maria-antoniak.github.io/resources/2019_cscw_birth_stories.pdf).
   2. Ash, Gauthier, and Widmer (2022), [Relatio: Text semantics capture political and economic narratives](https://arxiv.org/abs/2108.01720)
9. Sequence Models
   1. Ash, Durante, Grebenschikova, and Schwarz (2022), [Visual Representation and Stereotypes in News Media](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3934858).
   2. [Transformers tutorial](https://github.com/chkla/NLP2CSS-Tutorial)

## Learning Materials

**Books**

* *Natural Language Processing in Python*, Third Edition (“NLTK Book”).
  + Available at [nltk.org/book](https://www.nltk.org/book/).
  + Classic treatments of traditional NLP tools.
* Aurelien Geron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2019)
  + [O’Reilly Book](https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/), should be available with an academic account using ETH email.
  + A great practical book for machine learning and deep learning in Python, but not NLP-focused. We will use material from Chapters 2-4, 7-11, 13, and 15-17.
  + The deep learning chapters use Keras + TensorFlow.
  + [Jupyter notebooks](https://github.com/ageron/handson-ml2)
* Yoav Goldberg, *Neural Network Methods for Natural Language Processing* (2017)
  + [ETH Library Online Access](https://search.library.ethz.ch/permalink/f/12hkior/sfx_hbz3710000001177406) (email me if this doesn’t work)
  + A more advanced theoretical treatment of neural networks with an NLP focus, but already somewhat dated. We will use material from Chapters 1-17 and 19.
* Jurafsky and Martin, *Speech and Language Processing* (3d Ed. 2019).
  + [Available here](https://web.stanford.edu/~jurafsky/slp3/).
  + The standard theory text on computational linguistics.

**Programming**

Python is probably the best option for NLP, used by most data scientists. All the sample code is in Python. You are welcome to use another programming language.

* New to Python?
  + [Python installation instructions](https://docs.google.com/document/d/1UkCytHT4ZF-rDoh_buH6xb9mLz4GcGjT1qIu-pEWThI/edit?usp=sharing)
  + [Codecademy Online Python Course](https://www.codecademy.com/learn/learn-python-3)
  + [numpy tutorial](https://numpy.org/devdocs/user/absolute_beginners.html)
  + [pandas tutorial](https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/)
  + [Jupyter Notebooks Tutorial](https://github.com/fastai/course-v3/blob/master/nbs/dl1/00_notebook_tutorial.ipynb)
  + [Jupyter Notebook Keyboard Shortcuts](https://www.cheatography.com/weidadeyue/cheat-sheets/jupyter-notebook/)
  + [Google Colab Tips for Power Users](https://amitness.com/2020/06/google-colaboratory-tips/)
  + [Dash Web Apps Tutorial](https://dash.plot.ly/getting-started)
  + [Other Resources](https://forums.fast.ai/t/recommended-python-learning-resources/26888)
* New to Machine Learning?
  + [Codecademy Machine Learning Course](https://www.codecademy.com/learn/machine-learning)
  + Read the Geron Book, Chapters 1-7
  + [fast.ai Practical Deep Learning for Coders Course](https://course.fast.ai/index.html)
* New to Text Mining / NLP?
  + [Codecademy Online NLP Course](https://www.codecademy.com/learn/natural-language-processing)
  + Read the [NLTK Book](https://www.nltk.org/book/), Chapters 1-5
  + [fast.ai Code-First Introduction to Natural Language Processing](https://www.fast.ai/2019/07/08/fastai-nlp/)
* [Papers with Code (NLP)](https://paperswithcode.com/area/nlp)
  + Lists of papers with replication repos.
* Want to use R instead?
  + [Quanteda](https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html) is popular for text analysis among political scientists.
* Other resources:
  + [How to use the terminal](https://course.fast.ai/terminal_tutorial.html)
  + [How to use Google Colab notebooks](https://course.fast.ai/start_colab.html)

**Python Libraries**

**pip install pandas seaborn scikit-learn tensorflow nltk gensim flair spacy transformers**

* Basics:
  + [pandas](https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python): data loading and management
  + [seaborn](https://seaborn.pydata.org/): visualization
  + [sklearn](https://scikit-learn.org/stable/): general purpose Python ML library
  + [Keras + TensorFlow](https://www.tensorflow.org/guide/keras/overview): deep learning library
* NLP Necessities:
  + [nltk](https://www.nltk.org/): standard NLP tools
  + [gensim](https://radimrehurek.com/gensim/): topic models and embeddings
  + [spaCy](https://spacy.io/): tokenization, NER, syntactic parsing, word vectors
  + [flair](https://github.com/flairNLP/flair): sentiment analysis and some other tools ([tutorials](https://github.com/flairNLP/flair/tree/master/resources/docs))
  + [huggingface transformers](https://github.com/huggingface/transformers): transformer architectures
* Specialized tools:
  + [AllenNLP](https://allennlp.org/): library of models for semantic role labeling, entailment, question answering, etc
  + [fastText](https://fasttext.cc/): library of embeddings
  + [spacy-transformers](https://github.com/explosion/spacy-transformers): interface from spaCy to huggingface

## 

## **References**

## Yellow highlighting indicates required reading Blue highlighting indicates recommended methods reading

**Reference (Overview):**

* Gentzkow, Kelly, and Taddy, “[Text as Data](https://web.stanford.edu/~gentzkow/research/text-as-data.pdf).”
* Goldberg, Ch. 1
* NLTK book, Chapters 1, 2, 4
* Grimmer and Stewart, “[Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts](https://web.stanford.edu/~jgrimmer/tad2.pdf).”
* Raschka, “[Turn your Twitter Timeline into a Word Cloud](http://sebastianraschka.com/Articles/2014_twitter_wordcloud.html)”.

**Reference (Dictionary Methods):**

* [RegExOne Regular Expressions Lessons](https://regexone.com/)

**Reference:**  **Tokenization**

* [scikit-learn text feature extraction](https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction)
* Goldberg, Ch. 6
* NLTK book, Chapter 3, 5, 7, 8
* [A deep dive into preprocessing in NLP](https://mlexplained.com/2019/11/06/a-deep-dive-into-the-wonderful-world-of-preprocessing-in-nlp/)
* Denny and Spirling, “[Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145).”

**Reference (Dimensionality Reduction):**

* Geron, Chapters 8-9
* Gilis, [The why and how of nonnegative matrix factorization](https://arxiv.org/pdf/1401.5226.pdf).

**Methods (Document Distance)**:

* Lee et al, [An empirical evaluation of models of text document similarity](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.7144&rep=rep1&type=pdf).
* Brandon Rose, “[Document clustering in python](http://brandonrose.org/clustering).”

**Methods (Topic Models)**

* Prabhakaran, [Topic Modeling with Gensim](https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/)
* Quinn, et al, [How to analyze political attention with minimal assumptions and costs](http://faculty.washington.edu/jwilker/559/Quinn.pdf).
* Roberts et al, “[Structural Topic Models for Open-Ended Survey Responses](https://scholar.princeton.edu/sites/default/files/bstewart/files/topicmodelsopenendedexperiments_0.pdf)”.
* Christian Fong and Justin Grimmer, “[Discovery of treatments from text corpora](https://stanford.edu/~jgrimmer/SE_Short.pdf).”

**Reference (Machine Learning):**

* Goldberg Ch. 2, 7
* [Google Developers Text Classification Guide](https://developers.google.com/machine-learning/guides/text-classification/)
* NLTK book, chapter 6
* Geron, Chapters 2-4, 7

**Overview (Deep Learning for NLP)**

* Sebastian Ruder, [Deep Learning for NLP, Best Practices](https://ruder.io/deep-learning-nlp-best-practices/index.html)

**Reference (Neural Nets):**

* [Text classification from raw text (Google Colab)](https://colab.research.google.com/drive/1XcMJqKcTLjduIqjysRd1jGdRaKdetc3X)
* Goldberg, Ch. 3-5
* Geron, Chapters 10-11
* Leslie Smith, [A disciplined approach to neural network hyper-parameters](https://arxiv.org/abs/1803.09820)
* Chris Olah, [Backpropagation](https://colah.github.io/posts/2015-08-Backprop/)
* Baldi and Sadowski, [Understanding Dropout](https://papers.nips.cc/paper/4878-understanding-dropout.pdf)

**Reference (Embedding Layers):**

* Goldberg, Ch. 8
* [Bag of tricks for efficient text classification](https://arxiv.org/abs/1607.01759)

**References (RNNs):**

* Geron Ch. 15-17
* Goldberg, Ch. 14-17
* Sutskever, Vinyals, and Le, [Sequence to sequence learning with neural networks](https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf)
* Michael Nguyen, [Illustrated Guide to LSTMs and GRUs](https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21)
* Andrej Karpathy, [The unreasonable effectiveness of recurrent neural networks](http://karpathy.github.io/2015/05/21/rnn-effectiveness/)
* Chang and Masteron, [Using word order in political text classification with long short-term memory models](https://www.cambridge.org/core/journals/political-analysis/article/using-word-order-in-political-text-classification-with-long-shortterm-memory-models/D556D0AE5B1270124BC72EEB74244A80).

**Reference (Model Interpretation)**

* Ribeiro, Singh, and Guestrub, [Local interpretable model-agnostic explanations (LIME): An introduction](https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime).
* [Python Notebook with Model Interpretation Examples](http://savvastjortjoglou.com/intrepretable-machine-learning-nfl-combine.html#Permutation-Importance)

**Applications (MLP):**

* Vamossy, [Investor Emotions and Earnings Announcements](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3626025)
* Meursault, [The language of earnings announcements](https://finance.business.uconn.edu/wp-content/uploads/sites/723/2020/01/meursault_language_earnings_announcements.pdf)

**Applications (RNN):**

* [short] Iyyer et al, [Political ideology detection using recursive neural networks](https://www.aclweb.org/anthology/P14-1105.pdf).
* Ash et al, [In-Group Bias in the Indian Judiciary](http://www.devdatalab.org/judicial-bias)

**Reference (Word Embeddings):**

* Spirling and Rodriguez, [Word embeddings: What works, what doesn’t, and how to tell the difference for applied research](http://arthurspirling.org/documents/embed.pdf).
* Goldberg, Ch. 10-11
* Chapter Yoav Goldberg and Omer Levy, “[Word2Vec explained: Deriving Mikolov et al's Negative Sampling Word Embedding Method](https://arxiv.org/pdf/1402.3722.pdf)”.
* Piero Molino, “[Word embeddings: Past, present, and future](http://w4nderlu.st/teaching/word-embeddings)”.
* Matt Kusner, Yu Sun, Nicholas Kolkin, and Killian Weinberger, “[From word embeddings to document distances](https://mkusner.github.io/publications/WMD.pdf)”.
* Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, Andrej Risteski, [Linear Algebraic Structure of Word Senses, with Applications to Polysemy](https://arxiv.org/abs/1601.03764)
* Allen and Hospedales, [Analogies Explained: Towards Understanding Word Embeddings](https://arxiv.org/pdf/1901.09813.pdf)
* Ruder, [Approximating the Softmax](https://ruder.io/word-embeddings-softmax/)
* Peters, Ruder, and Smith, [To tune or not to tune: Adapting pretrained representations to diverse tasks](https://arxiv.org/abs/1903.05987).
* [ConceptNet NumberBatch](https://github.com/commonsense/conceptnet-numberbatch)
* Bojanowski et al, [Enriching word vectors with subword information](https://arxiv.org/pdf/1607.04606.pdf).
* Antoniak and Mimno, [Evaluating the stability of embedding-based word similarities](https://mimno.infosci.cornell.edu/info3350/readings/antoniak.pdf).
* Ash, Chen, and Ornaghi, [Gender attitudes in the judiciary: Evidence from U.S. Circuit Courts](https://www.dropbox.com/s/q0t56fhuhoffqzl/200625_Ash-Chen-Ornaghi_FINAL.pdf?raw=1)
* Hamilton, Clark, Leskovec, and Jurafsky, 2016, [Inducing domain-specific sentiment lexicons from unlabeled corpora](https://arxiv.org/pdf/1606.02820.pdf).

**Tools (Word Embeddings)**

* [Word embeddings in Flair](https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_4_ELMO_BERT_FLAIR_EMBEDDING.md)

**Contextualized Word Embeddings:**

* Peters et al, [Deep contextualized word representations](https://arxiv.org/abs/1802.05365).
* [ELMo embeddings with Flair](https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/ELMO_EMBEDDINGS.md)

**Reference (Syntactic Parsing):**

* NLTK Book, [Chapter 8 Analyzing Sentence Structure](https://www.nltk.org/book/ch08.html)
* [Jurafsky and Martin](https://web.stanford.edu/~jurafsky/slp3), Chapters 12-15, 20
* [ClearNLP Dependency Labels](https://github.com/clir/clearnlp-guidelines/blob/master/md/specifications/dependency_labels.md)

**Reference (Semantic Role Labeling):**

* Jurafsky and Martin, [Ch. 19: Semantic Role Labeling](https://web.stanford.edu/~jurafsky/slp3/19.pdf)
* [English PropBank Annotation Guidelines](https://verbs.colorado.edu/~mpalmer/projects/ace/EPB-annotation-guidelines.pdf)

**Tools (Syntactic Parsing):**

* [spaCy 101](https://spacy.io/usage/spacy-101)

**Tools (Semantic Role Labeling):**

* [Google Syntactic N-Grams Corpus](https://docs.google.com/document/d/14PWeoTkrnKk9H8_7CfVbdvuoFZ7jYivNTkBX2Hj7qLw/edit#)

**Tools for Document Embeddings**

* Ruder, [Deep Learning for NLP Best Practices](https://ruder.io/deep-learning-nlp-best-practices/index.html)
* [huggingface transformers](https://github.com/huggingface/pytorch-transformers)
* [spaCy interface to transformers](https://explosion.ai/blog/spacy-pytorch-transformers)

**References (Document Embeddings):**

* Arora, Liang, and Ma, “[A simple but tough-to-beat baseline for sentence embeddings](https://openreview.net/pdf?id=SyK00v5xx).”
* Doc2Vec:
  + Le and Mikolov, “[Distributed representations of sentences and documents](https://arxiv.org/abs/1405.4053).”
  + [A gentle introduction to Doc2Vec](https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e)
  + [Doc2vec implementation in Keras](https://github.com/samueljamesbell/doc2vec)
  + [Explanation of Doc2Vec Infer Vector](https://datascience.stackexchange.com/questions/37488/doc2vec-how-does-the-inference-step-work-in-pv-dbow)
* Wu et al, [Starspace: Embed all the things!](https://arxiv.org/abs/1709.03856)
* Bhatia, Lau, and Baldwin, “[Automatic labeling of topics with neural embeddings](https://arxiv.org/abs/1612.05340)”
* InferSent
* USE:
  + Cer et al, [Universal Sentence Encoder](https://arxiv.org/abs/1803.11175), ([code](https://tfhub.dev/google/universal-sentence-encoder/4))
  + Yang et al, [Multilingual universal sentence encoder for semantic retrieval](https://arxiv.org/abs/1907.04307).
* Clark, Celikyilmaz, and Smith, [Sentence mover’s similarity: Automatic evaluation for multi-sentence texts](https://www.aclweb.org/anthology/P19-1264.pdf).

**References (Attention / Transformers)**

* Bloem, [Transformers from scratch](http://peterbloem.nl/blog/transformers)
* Ruder, [NLP’s ImageNet moment has arrived](https://ruder.io/nlp-imagenet/)
* [The transformer explained](https://nostalgebraist.tumblr.com/post/185326092369/the-transformer-explained)
* Geron, Chapter 16
* Goldberg, Ch. 17
* Vaswani et al, [Attention is all you need](https://arxiv.org/abs/1706.03762)
* BERT:
  + Devlin et al, [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](https://arxiv.org/abs/1810.04805).
  + [The illustrated BERT](https://jalammar.github.io/illustrated-bert/)
  + [A primer in BERTology](https://arxiv.org/abs/2002.12327)
  + Reimers and Guryevich, [Sentence-BERT](https://arxiv.org/pdf/1908.10084.pdf)
  + Nie et al, [DisSent: Learning Sentence Representations from Explicit Discourse Relations](https://www.aclweb.org/anthology/P19-1442.pdf)
  + Hassan et al, [BERT, ELMo, USE, and InferSent Sentence Encoders](http://ceur-ws.org/Vol-2431/paper2.pdf)

Reference (Language Models):

* Goldberg, Ch. 9, 17

Reference (Transformers):

* [Huggingface Transformers Summary of Models](https://huggingface.co/transformers/model_summary.html)

**Reference (Autoregressive Models)**

* Shree, [The journey of OpenAI GPT Models](https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2)
* Radford et al, [Language models are unsupervised multitask learners](https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf).
  + [GPT-2 Demo](https://colab.research.google.com/github/ak9250/gpt-2-colab/blob/master/GPT_2.ipynb)
* [XLNet: Generalized autoregressive pretraining for language understanding](https://arxiv.org/abs/1906.08237).
* [Transformer-XL: Attentive language models beyond a fixed-length context](https://github.com/kimiyoung/transformer-xl).
* Brown et al, [Language models are few-shot learners](https://arxiv.org/abs/2005.14165) (GPT-3)
* [GPT-Neo](https://github.com/EleutherAI/gpt-neo/) (open-sourced GPT-3)

**Reference: Conditioned Text Generation**

* [Grover: A state-of-the-art defense against neural fake news](https://arxiv.org/abs/1905.12616)
* Dathathri et al, [Controlling text generation with plug and play language models](https://eng.uber.com/pplm/)
* CTRL

**Reference (Sequence-to-Sequence Transformers)**

* [EasyNMT](https://github.com/UKPLab/EasyNMT)
* Lewis et al, [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](https://arxiv.org/abs/1910.13461)
* Raffel et al, [Exploring the limits of transfer learning with a unified text-to-text transformer](https://arxiv.org/abs/1910.10683)
* Narang et al, [WT5?! Training Text-to-Text Models to Explain their Predictions](https://arxiv.org/pdf/2004.14546.pdf)

**Reference (Coreference Resolution):**

* Jurafsky and Martin, [Ch. 21: Coreference Resolution](https://web.stanford.edu/~jurafsky/slp3/21.pdf)

**Reference (Textual Entailment)**

**Reference (Siamese Neural Nets)**

* [Few shot learning in NLP with USE and siamese networks](https://medium.com/swlh/few-shot-learning-in-nlp-use-siamese-networks-189de22459d0)

**Reference (Discourse):**

* Jurafsky and Martin, [Ch. 23: Discourse Coherence](https://web.stanford.edu/~jurafsky/slp3/23.pdf)

**Reference (Dialogue):**

* Li et al, [Adversarial learning for neural dialogue generation](https://arxiv.org/abs/1701.06547).
* [Toward a conversational agent that can chat about anything](https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html)
  + [A state-of-the-art open source chatbot](https://ai.facebook.com/blog/state-of-the-art-open-source-chatbot/)
* Luo et al, [Detecting stance in media on global warming](https://arxiv.org/pdf/2010.15149.pdf)

**Big Bird**

* [Understanding BigBird’s Block Sparse Attention](https://huggingface.co/blog/big-bird)

**Reference (Information Extraction):**

* NLTK Book, [Chapter 7: Extracting Information](https://www.nltk.org/book/ch07.html)
* Jurafsky and Martin, [Ch. 17 Information Extraction](https://web.stanford.edu/~jurafsky/slp3/17.pdf)
* Angeli et al, [Leveraging linguistic structure for open domain information extraction](https://nlp.stanford.edu/pubs/2015angeli-openie.pdf).
* [Stanford OpenIE Python Wrapper](https://pypi.org/project/stanford-openie/)
* Qin et al, [ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning](https://arxiv.org/pdf/2012.15022.pdf)

**Reference (Knowledge Graphs)**

* [Scikit-kge](https://github.com/mnick/scikit-kge)
* Nickel, [Knowledge Graph Embeddings](https://mnick.github.io/project/knowledge-graph-embeddings/)
* Joulin et al, [Fast linear model for knowledge graph embeddings](https://arxiv.org/pdf/1710.10881.pdf)
* Lehmann et al, [DBPedia -- A large-scale, multilingual knowledge base extracted from Wikipedia](https://www.researchgate.net/profile/Christian_Bizer/publication/259828897_DBpedia_-_A_Large-scale_Multilingual_Knowledge_Base_Extracted_from_Wikipedia/links/0deec52e78a6e95b73000000/DBpedia-A-Large-scale-Multilingual-Knowledge-Base-Extracted-from-Wikipedia.pdf).
* [Dense Representations for Entity Retrieval](https://arxiv.org/pdf/1909.10506.pdf) ([code](https://github.com/google-research/google-research/tree/master/dense_representations_for_entity_retrieval/))
* [Connecting the Dots: Document level neural relation extraction with edge-oriented graphs](https://www.aclweb.org/anthology/D19-1498.pdf)
* Yao et al, [KG-BERT: BERT for knowledge graph completion](https://arxiv.org/pdf/1909.03193.pdf)

**Reference (Summarization):**

* Gabriel et al, [Cooperative Generator-Discriminator Networks for Abstractive Summarization with Narrative Flow](https://arxiv.org/abs/1907.01272)
* See et al, [Get to the point: Summarization with pointer-generator networks](https://arxiv.org/abs/1704.04368)
* [TLDR: Extreme Summarization of Scientific Documents](https://arxiv.org/abs/2004.15011v1)
* Stiennon et al, [Learning to summarize from human feedback](https://arxiv.org/abs/2009.01325)

**Reference (Question Answering):**

* Jurafsky and Martin, [Ch. 25](https://web.stanford.edu/~jurafsky/slp3/25.pdf)
* [Question answering with huggingface transformers](https://medium.com/@patonw/question-answering-with-pytorch-transformers-part-1-8736196bf20e)
* [NLP Progress: Question Answering](http://nlpprogress.com/english/question_answering.html)

**Reference (Claim Checking):**

* Vlachos, e-Fever

**Language Model Interpretation:**

**Reference (Legal AI)**

* Zhong et al, [A summary of legal artificial intelligence](https://arxiv.org/abs/2004.12158)

**Methods (Causal Inference with Text):**

* Keith et al, [Text and causal inference: A review of using text to remove confounding from causal estimates](https://arxiv.org/pdf/2005.00649v1.pdf)
* Wood-Doughty et al, [Challenges of Using Text Classifiers for Causal Inference](https://arxiv.org/pdf/1810.00956.pdf)
* Egami, Fong, Grimmer, Roberts, and Stewart, [How to Make Causal Inferences Using Texts](https://arxiv.org/pdf/1802.02163.pdf)

**Additional Applications**

**Complexity in Text):**

* Katz and Bommarito, “[Measuring the complexity of the law: The United States Code](https://link.springer.com/article/10.1007/s10506-014-9160-8).”
* Katz et al, [Complex societies and the growth of the law](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3602098)
* Benoit, Munger, and Spirling (2017), “[Measuring and Explaining Political Sophistication Through Textual Complexity](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3062061)”.
* [short] Louis and Nenkova (2013), [What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain](https://www.aclweb.org/anthology/Q13-1028.pdf)

**Dictionary Methods:**

* Michalopoulous and Xue (2019), [Folklore](https://www.nber.org/papers/w25430.pdf)
* Baker, Bloom, and Davis (2016), [Measuring economic policy uncertainty](http://www.policyuncertainty.com/media/EPU_BBD_Mar2016.pdf)
* Djourelova, [Media persuasion through slanted language: Evidence from the coverage of immigration](https://www.barcelonagse.eu/sites/default/files/working_paper_pdfs/1177.pdf)
* Enke 2020, [Moral values and voting](https://www.journals.uchicago.edu/doi/full/10.1086/708857)
* Cao et al, [How to talk when machines are listening: Corporate disclosure in the age of AI](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3683802)
* Truffa and Wong (2021), [Undergraduate Gender Diversity and Direction of Scientific Research](https://www.dropbox.com/s/qpz64fh8cs6dyg3/coed_draft.pdf?raw=1)
* Advani, Ash, Cai, and Rasul (2021), [Race-Related Research in Economics and other Social Sciences](https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/484648/CLE_WP_2021_06.pdf?sequence=1&isAllowed=y)

**Tokens/N-grams:**

* Gentzkow and Shapiro (2010), [What Drives Media Slant? Evidence from U.S. Daily Newspapers](https://web.stanford.edu/~gentzkow/research/biasmeas.pdf).
* Ash, Morelli, and Van Weelden, [Elections and divisiveness: Theory and evidence](https://www.journals.uchicago.edu/doi/full/10.1086/692587).

**Document Distance:**

* Kelly, Papanikolau, Seru, and Taddy, [Measuring technological innovation over the very long run](https://www.nber.org/papers/w25266).
* Hoberg and Phillips, [Text-based network industries and endogenous product differentiation](http://faculty.tuck.dartmouth.edu/images/uploads/faculty/gordon-phillips/text-based-network-industries-endogenous-product-differentiation.pdf).

**Topic Models**

* Barron, Huang, Spang, and DeDeo, [Individuals, institutions, and innovation in the debates of the French Revolution](https://www.pnas.org/content/115/18/4607). [has appendix]
* Hansen, McMahon, and Prat, [Transparency and deliberation with the FOMC: A computational linguistics approach](https://academic.oup.com/qje/article/133/2/801/4582916).
* Ash, Morelli, and Vannoni, “[More laws, more growth? Evidence from U.S. states](https://www.dropbox.com/s/kmzvks5ir8rtu0d/Ash-Morelli-Vannoni-Laws-Growth-Current.pdf?raw=1)”

**Text Classification:**

* Osnabrugge, Ash, and Morelli, [Cross-domain topic classification for political texts](https://elliottash.com/wp-content/uploads/2021/05/Osnabruegge-Ash-Morelli-Merged.pdf)
* Widmer, Galletta, and Ash, [Media slant is contagious](https://www.dropbox.com/s/akxaigqdqao50f0/Widmer-Ash-Galletta-current.pdf?raw=1)
* Gentzkow, Shapiro, and Taddy (2019), “[Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech](https://www.brown.edu/Research/Shapiro/pdfs/politext.pdf).”
* Kelly et al, [Text selection](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3491942)
* Peterson and Spirling, [Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems](https://www.cambridge.org/core/journals/political-analysis/article/classification-accuracy-as-a-substantive-quantity-of-interest-measuring-polarization-in-westminster-systems/45746D999CFCD1CB43E362392D7B2FB4)

**Word Embeddings:**

* Ash, Chen, and Ornaghi, [Gender attitudes in the judiciary: Evidence from U.S. Circuit Courts](https://www.dropbox.com/s/q0t56fhuhoffqzl/200625_Ash-Chen-Ornaghi_FINAL.pdf?raw=1)
* Gennaro and Ash, [Emotion and Reason in Political Language](http://bit.ly/gennaro-ash-emotions)
* Caliskan et al, “[Semantics derived automatically from language corpora contain human-like biases](http://opus.bath.ac.uk/55288/4/CaliskanEtAl_authors_full.pdf)”
* Bolukbasi et al, [Man is to computer programmer as woman is to homemaker: Debiasing word embeddings](http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf).
* Kozlowski, Taddy, and Evans 2019, [The geometry of culture: Analyzing the meanings of class through word embeddings](https://www.dropbox.com/s/3pc66x5t8pwcqwr/The_Geometry_of_Culture.pdf?dl=0)
* Stoltz and Taylor, [Concept Mover's Distance: Measuring Concept Engagement in Texts via Word Embeddings](http://scholar.google.com/scholar_url?url=https://osf.io/5hc4z/download/%3Fformat%3Dpdf&hl=en&sa=X&d=15059126327028699503&scisig=AAGBfm1Iqk1Qlxgms3En0EfCxKT7U0kUPQ&nossl=1&oi=scholaralrt&hist=o5uDfHMAAAAJ:3575187946198448895:AAGBfm0j-py_zEiokxtBExrZAD91IPAdKQ)
* [short] Gillani and Levy, [Simple dynamic word embeddings for perceptions in the public sphere](https://arxiv.org/pdf/1904.03352.pdf).
* Garg et al 2018, [Word embeddings quantify 100 years of gender and ethnic stereotypes](http://www.pnas.org/content/early/2018/03/30/1720347115/tab-figures-data) [includes appendix]
* [short] Lucy et al, [Content analysis of textbooks via natural language processing](https://journals.sagepub.com/doi/pdf/10.1177/2332858420940312)
* Thompson et al, [Cultural influences on word meanings revealed through large-scale semantic alignment](https://www.nature.com/articles/s41562-020-0924-8) [includes appendix]
* Rheault and Cochrane, [Word embeddings for the analysis of ideological placement in parliamentary corpora](https://www.cambridge.org/core/journals/political-analysis/article/word-embeddings-for-the-analysis-of-ideological-placement-in-parliamentary-corpora/017F0CEA9B3DB6E1B94AC36A509A8A7B).
* Nyarko and Sanga, [A statistical test for legal interpretation](https://juliannyarko.com/wp-content/uploads/other/nyarko_sanga_legal_interpretation.pdf).

**Syntactic Parsing:**

* Hoyle et al, [Unsupervised discovery of gendered language through latent-variable modeling](https://arxiv.org/pdf/1906.04760.pdf).
* [short] Ash, Jacobs, MacLeod, Naidu, and Stammbach, [Unsupervised extraction of workplace rights and duties from collective bargaining agreements](https://www.dropbox.com/s/2nsnt86a66u8r1i/Ash-Jacobs-MacLeod-Naidu-Stammbach.pdf?raw=1).
* Vannoni, Ash, and Morelli, [Measuring Discretion and Delegation in Legislative Texts](http://elliottash.com/wp-content/uploads/2019/08/Vannoni-Ash-Morelli-Merged.pdf)
* Michael Webb, [The impact of artificial intelligence on the labor market](https://web.stanford.edu/~mww/webb_jmp.pdf)

**Semantic Role Labeling:**

* Ash, Gauthier, and Widmer, Mining narratives from large text corpora
* Fetzer, [Can workfare programs moderate conflict? Evidence from India](https://academic.oup.com/jeea/article/18/6/3337/5727516?login=true)

**Information Extraction:**

* [short] Surdeanu et al 2011, [Customizing an information extraction system for a new domain](https://nlp.stanford.edu/pubs/relms2011.pdf).
* [short] Jurafsky and Chambers, [Unsupervised learning of narrative schemas and their participants](http://www.aclweb.org/anthology/P09-1068).
* [short] Clark, Ji, and Smith, [Neural text generation in stories using entity representations as context](https://www.aclweb.org/anthology/N18-1204.pdf).
* [short] Bamman and Smith, [Open extraction of fine-grained political statements](https://www.aclweb.org/anthology/D15-1008.pdf).
* [short] Wyner and Peters, [On rule extraction from regulations](http://wyner.info/research/Papers/WynerPetersJURIX2011.pdf).
* [short] Xia and Ding, [Emotion-Cause Pair Extraction](https://arxiv.org/abs/1906.01267)

**Document Embeddings:**

* [short] Demzsky et al, 2019, [Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings](https://arxiv.org/abs/1904.01596)
* [short] Dai, Olah, and Le, [Document Embedding with Paragraph Vectors](https://arxiv.org/pdf/1507.07998.pdf).
* Ash and Chen 2018, [Mapping the geometry of law using document embeddings](http://elliottash.com/wp-content/uploads/2018/09/Mapping_the_geometry_of_law_using_document_embeddings.pdf).
* Galletta, Ash, and Chen, [“Causal Eﬀects of Judicial Sentiment: Methods and Application to U.S. Circuit Courts”](http://ssrn.com/abstract=3415393)
* Ash, Chen, and Naidu (2020), “[Ideas have consequences: The effect of law and economics on American justice](http://elliottash.com/wp-content/uploads/2019/03/ash-chen-naidu-2019-03-20.pdf).”
* [short] Tong et al, [Low-skilled jobs face the highest re-skilling pressure](https://arxiv.org/pdf/2101.11505.pdf).

**Transformer Classification:**

* Bingler et al, [Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3796152)
* Pei and Jurgens, [Quantifying intimacy in language](https://arxiv.org/abs/2011.03020)

**Language Models:**

* [short] Peric, Mijic, Stammbach, and Ash, [Legal language modeling with transformers](http://ceur-ws.org/Vol-2764/paper2.pdf)
* Kreps, McCain, and Brundage, [All the news that’s fit to fabricate](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3525002).
* [short] Peng et al, [Fine-tuning a transformer-based language model to avoid generating non-normative text](https://arxiv.org/abs/2001.08764)
* [short] Adeem, Bethky, Reddy, [StereoSet: Measuring stereotypical bias in pre-trained language models](https://arxiv.org/abs/2004.09456)

**Local Semantics**

* Ross et al, [short] [Explaining NLP models via minimal contrastive editing](https://arxiv.org/pdf/2012.13985.pdf)
* Prabhakaran et al, [How metaphors impact political discourse](https://arxiv.org/pdf/2104.03928v1.pdf).

**Global Semantics**

* [short] Stammbach and Ash, [e-FEVER: Explanations and summaries for automated fact checking](https://truthandtrustonline.com/wp-content/uploads/2020/10/TTO04.pdf)
* Chen et al, [Opinion aware knowledge graph for political ideology detection](https://www.ijcai.org/proceedings/2017/0510.pdf)
* Vold and Conrad, [Using transformers to improve answer retrieval for legal questions](https://www.conradweb.org/~jackg/pubs/ICAIL21_Vold_Conrad.pdf).

**Causal Inference with Text:**

* Margaret Roberts, Brandon Stewart, and Richard Nielsen, “[Matching Methods for High-Dimensional Data with Applications to Text](http://www.margaretroberts.net/wp-content/uploads/2015/07/textmatching.pdf)”
* [short] Veitch et al, [Using text embeddings for causal inference](https://arxiv.org/abs/1905.12741)
* All the papers in Table 1 [here](https://arxiv.org/pdf/2005.00649v1.pdf).
* Zeng et al, [Uncovering interpretable potential confounders in electronic medical records](https://www.medrxiv.org/content/10.1101/2021.02.03.21251034v1).

**Argument Mining**

* [short] Subramanian et al, [Target Based Speech Act Classification in Political Campaign Text](https://arxiv.org/abs/1905.07856)

**Quote Extraction**

* Newell et al, [Quote extraction and analysis for news](https://research.signal-ai.com/assets/RnD_at_the_BBC__and_quotes.pdf)