PLSC 597: Text as Data
Fall 2021

Burt L. Monroe
Office: Sparks B002 (The Databasement) / [Pond 320]
Course website: https://burtmonroe.github.io/TextAsDataCourse and Canvas
Office Hours: Friday 2-4 & by appt. Appointments: http://burtmonroe.youcanbook.me
Contact: burtmonroe@psu.edu

Description

This course investigates the use of digitized texts – news articles, speeches, laws, treaties, press releases, party manifestos, campaign ads, interviews, transcripts, open-ended surveys, Tweets, Reddit posts, YouTube comments, Yelp reviews ... – as sources of data for social science research.

We'll begin the semester with overviews of the "text as data" field in political science and more broadly in computational social science, as well as the big picture in overlapping fields like natural language processing (NLP) and computational linguistics. Here we will discuss what sorts of social scientific questions we can address with text as data, and some of the challenges and opportunities of using text as social science data.

We'll then begin discussing the theory and mechanics of how to convert text into data. This will include topics like preprocessing text and related NLP tasks (e.g., stemming, tokenizing) and representing text as data (e.g., bag-of-words, word embeddings, measures of association), as well as discussion of the downstream consequences. We'll start the semester assuming we're working with "clean" text, but throughout the course of the semester we will also address some of the more practical data science issues when dealing with real-world "dirty" text (e.g., file encodings, file formats, data that has been poorly transcribed, thumb-typed, or OCRed) and of obtaining or sharing text in the first place (e.g., the mechanics and ethics of web scraping).

We'll then turn to the major approaches to measuring social science concepts with textual data, including rule-based methods, supervised learning from human-coded or known examples, and unsupervised methods from matrix decompositions to generative Bayesian measurement models. As we go, we will discuss particular measurement objectives like classification, scaling, topic modeling, and analysis of sentiment and stance, as well as ways of validating our models. We will also be learning about the neural network / deep learning approach that has come to dominate NLP in recent years. This will arise lightly in the first half of the semester in our discussion of language models, logistic regression, and embeddings. In "Neural November" we will dive in to this a little deeper, so to speak, and discuss concepts in the rapidly changing state-of-the-art, including transfer learning using pretrained language models and transformers like BERT.

The course will assume students have a base facility with Python or R, and some graduate level work in statistical inference, quantitative social science methodology, or machine learning.

Assignments and Grades

Grades will be based on the following:

- **Engagement in Seminar - 20%**

  - **Readings and Seminar Discussion**
    These points will be based on participation and engagement with the readings and the discussion, both in class and on the Canvas discussion boards. For the Canvas boards, you will by Monday (midnight, but earlier is better), post questions / comments to that week's discussion board. Questions / comments can be about key and/or confusing terminology or concepts, connections or disagreements between / among the authors or others, points you found insightful or you disagreed with, applications of concepts from the readings, etc. The primary objective with these is to provide your colleagues (including me) with opportunity for substantive discussion of the readings and key concepts.

    By Tuesday, 5:00 pm (earlier is better), you will need to reply to at least two of your colleagues' posts on that week's discussion board. Replies should be substantive ... at least a few sentences, not just "yes" or "no."

  - Full 20 points if you post and are present every week, demonstrate engagement with all of the assigned readings, are prepared to talk about the week's themes, and consistently contribute in ways that are productive to the discussion (good questions, thoughtful responses, etc.).

- **Exercises - 20%** It is my intention to have exercises (almost) weekly for the first 8-10 weeks. (This has historically been the most difficult of the course elements to execute fully, partly because of typically wide variation in R/Python background in the class, and there may in the event be fewer exercises than intended.) Generally, you are welcome to speak with each other about the exercises, but code must be your own or its source documented in the code. After the fact, I may in some cases share your code, or a modification thereof, as an example solution. These will be graded for "good faith effort" only.

- **Project - 40%** Research paper. This may be an application of interest to your (or some) discipline, which uses text as data, or a methodological contribution relevant to the literature on text as data. (It may not, broadly speaking, be a data collection just for the sake of data collection, full stop. It also may not be a dictionary-based sentiment analysis, full stop.) You'll need to submit a one page description of your proposed paper/project by October 8, but I encourage you to do so, or to discuss it with me, as soon as you formulate your idea. (Among other reasons, we are unlikely to fully cover all of the techniques all of you might need to execute all of your projects, especially by early October, so I will likely need to point you to relevant material.)

  Papers must be in a format plausible as a submission (other than anonymization) to an appropriate peer-reviewed outlet, with appropriate supplementary materials available. In practice, this means the expectation is a pithy short paper – either in computer science conference "short paper" format (e.g., four pages plus references in a dense two-column style like that of ACL or NeurIPS) or in social science "letter" format (about 3-4000 words) – but

that I am perfectly fine with a longer paper – e.g., an 8-12000 word social science journal article – if that is what you wish to do.

Papers will be submitted a month from the end of the semester. If all goes according to script, each of you will receive (and write) two (random, blind) peer reviews within two weeks. (In past iterations, health issues and similar have disrupted the math of everybody-does-two-everybody-gets-two.) At the end of the semester you will submit a revised version of your paper along with a memo to reviewers and the editor (me) discussing how (or why not) reviewer comments have been addressed.

If your scientific objective requires data collection and cleaning beyond the scale of what is possible in this time frame, modify your objective. Consider, for example, whether you can conduct a pilot study on a smaller sample or on a readily available proxy for the ultimate data of interest.

Papers may not require the violation of copyright or other legal constraints, or violate any appropriate terms of service.

- **Oct 8 (Fri) - Project proposal / description due, but encouraged earlier**
- **Nov 10 (Wed) - Initial submission deadline**
- **Nov 28 (Sun)- Receive peer reviews this day or sooner**
- **Dec 15 (Wed) - Final submission deadline**

- **Project Peer Reviews - 20%**

  You will have two weeks to provide peer reviews of two colleagues' papers, assigned randomly by me single-blind (the author will not know who acted as reviewers). Submit your reviews in plain text so that I can ensure reviewer anonymity. The sole criterion I will use to grade these is "constructiveness." A constructive review offers reasonable suggestions for improvement of the paper, is not unnecessarily snide in any criticism it offers, and is returned on time.

  - **Nov 12 (Fri) - Receive papers to review by this date.**
  - **Nov 27 (Sat) - Deadline to return reviews.** Note: if a paper is received by a reviewer later than Nov 12, that reviewer has two weeks from that date to return the review, but the final submission deadline for the paper is not extended. Note 2: That is the Saturday of the Thanksgiving break, so you will likely want to tackle this earler.

Tutorial Notebooks & Code

In a typical class session, I will spend the last half of class reviewing notebooks with code demonstrations in Python and R. Often, your take-home exercises will involve modifying the notebook code. They will mostly be available on the class github site – https://burtmonroe.github.io/TextAsDataCourse/Tutorials – and/or on Google Colab. Links to the relevant notebooks will be made available shortly before class via Canvas. (Note that the website also contains older versions of some notebooks from previous iterations of the course – some of which may no longer function due to deprecations in the codebase, and some of which contain exercises that may not be assigned this semester – as well as notebooks from a partially overlapping "advanced" course taught for Essex in Summer 2021 – some of which assume prior knowledge of material covered in the first eight weeks of this course, and some of which contain exercises that may not be assigned this semester.)

Schedule and Readings

There is no textbook for the course and most of our readings will be articles. But we will read, or I will refer you to, multiple sections of each of the following free book (drafts):

- **[SLP]** Daniel Jurafsky and James Martin. 2020. *Speech & Language Processing* 3rd ed., draft. https://web.stanford.edu/~jurafsky/slp3/
- **[NLP]** Jacob Eisenstein. 2019. *Natural Language Processing* (Preprint version) https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

August 26 - Introduction

Syllabus, administrivia, R and Python resources, Google Colab.

Additional Resources for R and Python

- Open Source Tools for Text-as-Data/NLP in R https://burtmonroe.github.io/TextAsDataCourse/Notes/RText/
- Open Source Tools for Text-as-Data/NLP in Python https://burtmonroe.github.io/TextAsDataCourse/Notes/PythonText/
- Mango Solutions. 2018. "Python for R Users Workshop." https://github.com/MangoTheCat/python-for-r-users-workshop
- Kohei Watanabe and Stefan Müller. Quanteda Tutorials. (R) https://tutorials.quanteda.io/
- Cornelius Puschmann. 2019. "Advancing Text Mining with R and quanteda." http://cbpuschmann.net/quanteda_mzes/
- The following books are available electronically through the Penn State library:
    - Zhenya Antić. 2021. *Python Natural Language Processing Cookbook.* Packt.
    - Ajay Ohri. 2017. *Python for R Users: A Data Science Approach* Wiley.
    - Rick Scavetta. 2021. *Python and R for the Modern Data Scientist.* O'Reilly.
    - François Chollet. 2018. *Deep Learning with Python* Manning.
    - Julia Silge & David Robinson. 2017. *Text Mining with R: A Tidy Approach.* O'Reilly. (tidytext)
    - Ted Kwartler. 2017. *Text Mining in Practice with R* Wiley. (tm, openNLP)
    - Mark Hodnett. 2019. *Deep Learning with R for Beginners* Packt.

September 2 - Text-as-Data and NLP

Overview of the goals and methods of social scientific text-as-data analysis, with comparison and contrast to the goals and methods in the field of NLP (Natural Language Processing). Measurement with text as data. In-class notebooks on the basics of bag-of-words text data in R and Python (with emphasis on quanteda, tidytext, and CountVectorizer).

- Benoit, Kenneth. 2020. "Text as Data: An Introduction." In Franzese and Curini, Eds. *SAGE Handbook of Research Methods in Political Science and International Relations.* Chapter 26. PDF as published: https://kenbenoit.net/pdfs/CURINI_FRANZESE_Ch26.pdf.
- **[NLP]** Ch. 1, "Introduction," esp. Sect. 1.1, "Natural Language Processing and Its Neighbors."
- Further reference:

- Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents." *Political Analysis*. 21(3): 267-297. https://web.stanford.edu/~jgrimmer/tad2.pdf
- "NLP-Overview: Modern Deep Learning Techniques Applied to Natural Language Processing." 2021. https://nlpoverview.com/index.html.
- Paul DiMaggio. 2015. "Adapting computational text analysis to social science (and vice versa)." *Big Data & Society* 2(2). https://journals.sagepub.com/doi/10.1177/2053951715602908
- John Wilkerson and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* https://www.annualreviews.org/doi/10.1146/annurev-polisci-052615-025542
- Matthew Gentzkow, Bryan T. Kelly, and Matt Taddy. 2017. "Text as Data." https://web.stanford.edu/~gentzkow/research/text-as-data.pdf
- Brendan O'Connor, David Bamman, and Noah A. Smith. 2011. "Computational Text Analysis for Social Science: Model Assumptions and Complexity." (2011) NIPS Workshop on Computational Social Science and the Wisdom of Crowds. http://brenocon.com/oconnor+bamman+smith.nips2011css.text_analysis.pdf
- Burt Monroe and Philip Schrodt. 2008. "Introduction to the Special Issue: The Statistical Analysis of Political Text." *Political Analysis* 16, 4, 351-355. https://doi.org/10.1093/pan/mpn017
- Margaret E. Roberts. 2016. "Introduction to the Virtual Issue: Recent Innovations in Text Analysis for Social Science." *Political Analysis* https://doi.org/10.1017/S1047198700014418

September 9 - Text Processing and Regular Expressions

Introduction to wrangling text data, text "preprocessing" / normalization, regular expressions, and text data in the wild.

- **[SLP]** Ch. 2, "Regular Expressions, Text Normalization, and Edit Distance."
- Matthew Denny and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26(2): 168-89. https://doi.org/10.1017/pan.2017.44

September 16 - Representing and Comparing Texts

The vector space model and bag-of-words / one-hot representations of documents. Comparing texts, information retrieval, and cosine similarity. Co-occurrence, (positive) pointwise mutual information (PMI/PPMI), tf-idf and Fightin' Words.

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *Introduction to Information Retrieval* Cambridge University Press. http://nlp.stanford.edu/IR-book/. Chapter 6, "Scoring, Term Weighting, and the Vector Space Model." (Note: Our interest here is in the representation of documents as vectors and comparing such vectors by cosine similarity. The context of this chapter is Google-like information retrieval in which one "document" is a "query" consisting of search terms and you want a list of documents in the database ranked by similarity to the query. The first five chapters of IIR are focused on *Boolean* information retrieval, where your query is something like `(TERM_1 OR TERM_2) AND TERM_3`, and documents in the database match or don't.)
- **[SLP]** Ch. 6, "Vector Semantics."
- Burt L. Monroe, Michael Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis*. 16(4): 372-403. https://doi.org/10.1093/pan/mpn018.

September 23 - NLP Annotation Pipelines

Basic NLP language modeling. Sequence labeling annotation tasks like part-of-speech tagging, lemmatization, morphological analysis, and named entity recognition. Dependency parsing. In-class notebooks will introduce popular NLP pipelines available in R and/or Python, with particular focus on spaCy, UDPipe, Stanza, OpenNLP, and NLTK.

- **[SLP]** Chapter 3, "N-gram Language Models." (Don't get too caught up in the math - I just want you to get the gist of what an n-gram language model is and what its strengths and weaknesses are.) [Additional reference, [NLP] Chapter 6. "Language Models."]
- **[SLP]** Chapter 8. "Sequence Labeling for Parts of Speech and Named Entities." (Again, I just want you to know the basic goals and limitations of things like POS tagging and NER. Don't get too caught up in *how* these work, e.g., Hidden Markov Models, the Viterbi Algorithm, and Conditional Random Fields.) [Additional reference, [NLP] Chapter 7. "Sequence Labeling."]
- **[NLP]** Chapter 8. "Applications of Sequence Labeling."
- **[SLP]** Chapter 14. "Dependency Parsing." (Don't get too caught up in the linguistics or the computation. I just want you to understand what dependency parsing *is*.)

September 30 - Dictionaries, Lexicons, & Keywords.

Dictionary-based analysis, including sentiment analysis. In-class notebooks will demonstrate dictionary-based sentiment analysis with dictionaries like Lexicoder and VADER.

- **[SLP]** Ch. 19 "Lexicons for Sentiment, Affect, and Connotation."
- Lori Young and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29(2): 205-31. https://doi-org.ezaccess.libraries.psu.edu/10.1080/10584609.2012.671234. (Lexicoder).
- Leah Cathryn Windsor, James Grayson Cupit, and Alistair James. 2019. "Automated content analysis across six languages." *PLoS ONE* https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224425. (An application of LIWC.)
- Additional items for reference:
    - C. J. Hutto and Eric Gilbert. 2014. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text." *Eighth International AAAI Conference on Weblogs and Social Media* https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399 (VADER.)
    - Sayeed Salam, Patrick Brandt, Vito D'Orazio, Jennifer Holmes, Javiar Osorio, and Latifur Khan. 2020. "An Online Structured Political Event Dataset based on CAMEO Ontology." https://files.osf.io/v1/resources/vrt4a/providers/osfstorage/5e722a270cd06c046c001ec7?action=download&direct&version=2 (An example of dictionary analysis beyond sentiment.)
    - On the closely related task of identifying "keywords" for search or other filtering to obtain a relevant subset of documents for subsequent analysis.
        - Shuai Wang, Zhiyuan Chen, Bing Liu, and Sherry Emery. 2016. "Identifying Search Keywords for Finding Relevant Social Media Posts." *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1). https://ojs.aaai.org/index.php/AAAI/article/view/10387
        - Gary King, Patrick Lam, and Margaret E. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61, no. 4 (2017): 971-988. https://scholar.harvard.edu/files/gking/files/ajps12291_final.pdf.
        - Fridolin Linder. 2017. "Improved Data Collection for Online Research Using Query Expansion and Active Learning." SSRN: https://dx.doi.org/10.2139/ssrn.3026393.

October 7 - Supervised Learning & Classification.

"Classical" (non-neural) approaches to text classification and other supervised learning problems (e.g., Naive Bayes, logistic regression, support vector machines, random forests, XGBoost).

- **[SLP]** Chapters 4 (Naive Bayes and Sentiment Classification) and 5 (Logistic Regression).
- Pablo Barberá, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29, no. 1: 19–42. doi:10.1017/pan.2020.8.
- **[NLP]** Ch. 4 "Linguistic applications of classification." (For additional reference: Chapter 2 - "Linear text classification," and Ch. 3 "Nonlinear classification.")
- Samuel E. Bestvater and Burt L. Monroe. 2020. "Sentiment is not Stance: Target-Aware Opinion Classification for Political Text." (R&R *Political Analysis*).
- Additional applications for reference:
    - D'Orazio, V., Landis, S. T., Palmer, G. and Schrodt, P. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines," *Political Analysis*, 22(2), pp. 224–242. doi: 10.1093/pan/mpt030.
    - Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., and Mikhaylov, S. 2016. "Crowd-sourced text analysis: reproducible and agile production of political data." *American Political Science Review*, 110(2), 278-295. https://kenbenoit.net/pdfs/Crowd_sourced_data_coding_APSR.pdf
    - Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A. and Parnet, O. 2016. "A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates." *Journal of Communication*, 66: 1007–1031. https://doi.org/10.1111/jcom.12259
    - (More "regression" than classification.) Nicholas Beauchamp. 2017. "Predicting and interpolating state level polls using Twitter textual data." *American Journal of Political Science* https://doi.org/10.1111/ajps.12274

Background note: If you haven't seen supervised learning / classification before - in say, IST or STAT 557: (a) we're in the same methodological neighborhood as "logistic regression" aka "logit," and modern alternatives, and (b) For a fairly gentle introduction to supervised learning and classification generally (not just in text), with lots of R code and examples, see James et al. 2014 *Introduction to Statistical Learning*. Chapter 4 (Classification), Chapter 8 (Tree-based Methods), Chapter 9 (Support Vector Machines) https://www.statlearning.com/.

October 14 - Unsupervised Learning: Topic and Scaling Models

Latent variable and dimension reduction techniques for text data. Clustering, matrix factorization (e.g., SVD, LSI, LSA, NMF), generative topic modeling (e.g., LDA, CTM), topic modeling with structure (e.g., STM), and scaling (e.g., WordFish). "Choosing K." Validation.

- David M. Blei . 2012. "Probabilistic Topic Models." http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf
- Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science*. 54(1): 209–28. http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2009.00427.x/full (Don't get caught up in the math of the topic model introduced here, as it's unlikely to be appropriate for your particular application. Focus on (a) the comparison to supervised learning and (b) the approach to validation.)

- Margaret Roberts, Brandon Stewart, Dustin Tingley, and Edoardo Airoldi. 2013. "The Structural Topic Model and Applied Social Science." https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf (This is a short overview paper. I encourage you to seek out more complete treatments here: https://www.structuraltopicmodel.com, including R package vignette, and more well known papers Roberts et al 2014 *American Journal of Political Science* and Lucas et al., 2015, *Political Analysis*.)
- Jonathan Slapin and Sven-Oliver Proksch. 2008. "A scaling model for estimating time series party positions from texts." *American Journal of Political Science*. (WordFish) [See also Burt Monroe and Ko Maeda. 2004. "Talk's cheap: Text-based estimation of rhetorical ideal-points." Covered the same ground but wasn't sent out for publication because the technique does not, in my view, actually work.]
- Benjamin Lauderdale and Alex Herzog. 2016. "Measuring political positions from legislative speech." *Political Analysis*. (WordShoal) https://alexherzog.net/files/Lauderdale_Herzog_PA_2016.pdf
- Additional References:
    - **[NLP]** Chapter 5 "Learning without supervision."
    - David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(Jan):993–1022 (Canonical LDA cite.) https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
    - Karl Rohe, Muzhe Zeng. 2020. "Vintage Factor Analysis with Varimax Performs Statistical Inference." https://arxiv.org/abs/2004.05387.
    - Chris Ding, Tao Li, and Wei Peng. 2008. "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing." *Computational Statistics & Data Analysis*, 52(8):3913–3927. https://ranger.uta.edu/~chqding/papers/NMFpLSIequiv.pdf
    - David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing semantic coherence in topic models." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 262–272. Association for Computational Linguistics. http://dirichlet.net/pdf/mimno11optimizing.pdf
    - David M. Blei, and John D. Lafferty. 2005. "Correlated Topic Models" *NeurIPs* https://proceedings.neurips.cc/paper/2005/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf
    - Gallagher, R. J., Reing, K., Kale, D., & Ver Steeg, G. (2017). Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. Transactions of the Association for Computational Linguistics (TACL), 5, 529-542. (CorEx topic model.) https://ryanjgallagher.github.io/publications/gallagher2017anchored
    - Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. 2020. "Keyword assisted topic models." https://arxiv.org/pdf/2004.05964.pdf (keyATM topic model.)
    - Will Lowe, Ken Benoit, Slava Mikhalov, and Michael Laver. 2011. "Scaling policy preferences from coded political texts." *Legislative Studies Quarterly*. https://doi.org/10.1111/j.1939-9162.2010.00006.x
    - Michael Laver, Kenneth Benoit, and John Garry (2003) "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2). (Wordscores) https://doi.org/10.1017/S0003055403000698
    - Will Lowe 2008. "Understanding Wordscores." *Political Analysis*, 16(4), 356-371. https://doi.org/10.1093/pan/mpn004
    - William R. Hobbs. 2019. "Text Scaling for Open-Ended Survey Responses and Social Media Posts." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3044864
    - Claire Kelling and Burt L. Monroe. 2021. "Analyzing Community Reaction to Refugees through Text Analysis of Social Media Data." Forthcoming, *Journal of Ethnic and Migration Studies*. (STM application.)

- Burt L. Monroe. 2019. "The Meanings of 'Meaning' in Social Scientific Text Analysis." *Sociological Methodology.* (Comment on LDA application in Goldenstein and Poschmann (2019).) https://doi.org/10.1177%2F0081175019865231

October 21 - Word Embeddings

What word embeddings are, what they're useful for, and how to estimate them.

- **[SLP]** Chapter 6, "Vector Semantics and Embeddings."
- Pedro Rodriguez and Arthur Spirling. 2021. "Word embeddings: What works, what doesn't, and how to tell the difference for applied research." *Journal of Politics.* https://github.com/ArthurSpirling/EmbeddingsPaper/blob/master/Paper/Embeddings_SpirlingRodriguez.pdf
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human Biases." *Science.* https://arxiv.org/abs/1608.07187.
- Additional references (social science applications)
  - Mitchell Goist and Burt L. Monroe. 2020. "Scaling the Tower of Babel: Common-Space Analysis of Political Text in Multiple Languages."
  - Pedro Rodriguez, Arthur Spirling, and Brandon Stewart. 2021. "Embedding Regression: Models for Context-Specific Description and Inference." https://arthurspirling.org/documents/embedregression.pdf
  - Moritz Osnabrügge, Sara B. Hobolt, and Toni Rodon. "Playing to the Gallery: Emotive Rhetoric in Parliaments." *American Political Science Review* 115, no. 3 (2021): 885–99. doi:10.1017/S0003055421000356.
  - Austin C. Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84, no. 5: 905–49. https://doi.org/10.1177/0003122419877135.
- Additional references (the canonical papers)
  - Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *Proceedings of Workshop at ICLR.* (word2vec) https://arxiv.org/abs/1301.3781
  - Pennington, Socher, and Manning. 2014. "GloVe: Global Vectors for Word Representation." https://nlp.stanford.edu/projects/glove/.
  - Levy and Goldberg. 2014. "Neural Word Embeddings as Implicit Matrix Factorization.." https://levyomer.files.wordpress.com/2014/09/neural-word-embeddings-as-implicit-matrix-factorization.pdf.
  - Quoc Le and Tomas Mikalov. 2014. "Distributed Representations of Sentences and Documents." *JMLR.* (doc2vec). http://proceedings.mlr.press/v32/le14.html

October 28 - No class

I will be attending the New Directions in Analyzing Text as Data (TADA) conference.

Neural November

November 4 - Introduction to Neural NLP and Deep Learning

The perceptron and shallow networks for logistic regression. Multilayer, deep, feedforward networks. Activation functions, loss functions, backpropagation, stochastic gradient descent. In-class notebooks will demonstrate how to build and estimate a basic deep feedforward network for text classification, using Keras and Tensorflow in R and Python.

- **[SLP]** Chapter 7. "Neural Networks and Neural Language Models."
- Jay Alammar. 2016. "A Visual and Interactive Guide to the Basics of Neural Networks." https://jalammar.github.io/visual-interactive-guide-basics-neural-networks/
- Jay Alammar. 2016. "A Visual and Interactive Look at Basic Neural Network Math." https://jalammar.github.io/feedforward-neural-networks-visual-interactive/
- Additional items for reference:
    - Kakia Chatsiou and Slava Jankin Mikhaylov. 2020. "Deep Learning for Political Science." In Robert Franzese and Luigi Curini, eds. SAGE Handbook of Research Methods in Political Science and International Relations. https://sk.sagepub.com/reference/the-sage-handbook-of-research-methods-in-political-science-and-ir
    - Goodfellow, Bengio, and Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.
    - Howard, Jeremy. Video series: "Practical Deep Learning for Coders." https://www.youtube.com/playlist?list=PLfYUBJiXbdtSIJb-Qd3pw0cqCbkGeS0xn.
    - MIT Introduction to Deep Learning Class: http://introtodeeplearning.com/

November 11 - Deeper Deep Learning

Hyperparameters. Dealing with overfitting through regularization of weights, early stopping, dropout, data augmentation. Optimization techniques. Embedding layers. In-class notebooks will elaborate on the previous week's notebooks.

No reading, since papers are due this week.

November 18 - From RNNs to Transformers

Recurrent neural nets. Convolutional neural nets. The attention mechanism. Self-attention and transformers.

- **[SLP]** Chapter 9, "Deep Learning Architectures for Sequence Processing."
- Jay Alammar. 2018. "Visualizing a Neural Machine Translation Model (Mechanics of Seq2seq Models with Attention)." https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/
- Jay Alammar. 2018. "The Illustrated Transformer." https://jalammar.github.io/illustrated-transformer/
- Canonical references:
    - Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. "Sequence to Sequence Learning with Neural Networks." https://papers.nips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf (seq2seq / encoder-decoder)

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All You Need." (Transformers) https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

November 25 - No Class (Thanskgiving)

December 2 - Transfer Learning

Contextual embeddings, pretrained language models, and transfer learning. BERT, ELMo, and friends.

- Noah Smith. 2019. "Contextual Word Representations: A Contextual Introduction." https://arxiv.org/abs/1902.06006
- Jay Alammar. 2018. "The Illustrated BERT, ELMo and Co. (How NLP Cracked Transfer Learning)." https://jalammar.github.io/illustrated-bert/
- Jay Alammar. 2019. "A Visual Guide to Using BERT for the First Time." https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/
- Zhanna Terechskenko, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2020. "A Comparison of Methods in Political Science Text Classification: Transfer Learning Language Models for Politics." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3724644
- Canonical references:
  - Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. 2018. "Deep Contextualized Word Representations." https://arxiv.org/abs/1802.05365 (AllenNLP/ELMo).
  - Jeremy Howard and Sebastian Ruder. 2018. https://arxiv.org/pdf/1801.06146.pdf (ULMFiT - introduction of fine-tuning)
  - Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. 2018. "Improving Language Understanding by Generative Pre-Training." https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (OpenAI/GPT)
  - Jacob Devlin, Ming-wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." https://arxiv.org/abs/1810.04805 (Google/BERT)
  - The arms race in pretrained models: GPT-2 (OpenAI, 2019), ERNIE (Baidu, 2019), XLNet (CMU/Google, 2019), RoBERTa (Facebook, 2019), ALBERT (Google, 2019), StructBERT (Alibaba, 2019), MegatronLM (Nvidia, 2019), T5 (Google, 2019), Reformer (Google, 2020), Meena (Google, 2020), Turing NLG (Microsoft, 2020), GPT-3 (OpenAI, 2020), ELECTRA (Stanford, 2020), DeBERTa (Microsoft, 2020), OmniNET (Google, 2021).

December 9 - What do NLP models learn?

Natural language understanding (NLU), interpretability / BERTology, fairness and bias.

- Emily M. Bender and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." https://aclanthology.org/2020.acl-main.463.pdf
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. "A Primer in BERTology: What We Know About How BERT Works." Transactions of the Association for Computational Linguistics, vol. 8, pp. 842–866. https://doi.org/10.1162/tacla00349

- Aylin Caliskan. 2021. "Detecting and mitigating bias in natural language processing." Brookings. https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/
- Additional References:
  - https://github.com/uclanlp/awesome-fairness-papers
  - Kai-Wei Chang, Vicente Ordonez, Margaret Mitchell, Vinodkumar Prabhakaran. 2019. "Bias and Fairness in NLP." Tutorial delivered at EMNLP. Slides are in three parts here: http://web.cs.ucla.edu/~kwchang/documents/slides/emnlp19-fairNLP-part1.pdf, http://web.cs.ucla.edu/~kwchang/documents/slides/emnlp19-fairNLP-part2.pdf, and http://web.cs.ucla.edu/~kwchang/documents/slides/emnlp19-fairNLP-part3.pdf. Video of the tutorials (Three hours in total) here: http://web.cs.ucla.edu/~kwchang/talks/emnlp19-fairnlp/
  - Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." https://arxiv.org/pdf/1908.09635.pdf
  - "NLP-Overview: Modern Deep Learning Techniques Applied to Natural Language Processing." https://nlpoverview.com/index.html.
  - "NLP-Progress: Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks." https://nlpprogress.com.
  - "Awesome-NLP: A curated list of resources dedicated to Natural Language Processing (NLP)." https://github.com/keon/awesome-nlp

Policy Statements

COVID-19

Penn State University requires everyone to wear a face mask in all university buildings, including classrooms, regardless of vaccination status. ALL STUDENTS MUST wear a mask appropriately (i.e., covering both your mouth and nose) while you are indoors on campus. This is to protect your health and safety as well as the health and safety of your classmates, instructor, and the university community. Anyone attending class without a mask will be asked to put one on or leave. Instructors may end class if anyone present refuses to appropriately wear a mask for the duration of class. Students who refuse to wear masks appropriately may face disciplinary action for Code of Conduct violations. If you feel you cannot wear a mask during class, please speak with your adviser immediately about your options for altering your schedule.

Academic Integrity

Academic integrity is the pursuit of scholarly activity in an open, honest and responsible manner. Academic integrity is a basic guiding principle for all academic activity at The Pennsylvania State University, and all members of the University community are expected to act in accordance with this principle. Consistent with this expectation, the University's Code of Conduct states that all students should act with personal integrity, respect other students' dignity, rights and property, and help create and maintain an environment in which all can succeed through the fruits of their efforts.

Academic integrity includes a commitment by all members of the University community not to engage in or tolerate acts of falsification, misrepresentation or deception. Such acts of dishonesty violate the fundamental ethical principles of the University community and compromise the worth of work completed by others.

Disability Accommodation

Penn State welcomes students with disabilities into the University's educational programs. Every Penn State campus has an office for students with disabilities. Student Disability Resources (SDR) website provides contact information for every Penn State campus (http://equity.psu.edu/sdr/disability-coordinator). For further information, please visit the Student Disability Resources website (http://equity.psu.edu/sdr/).

In order to receive consideration for reasonable accommodations, you must contact the appropriate disability services office at the campus where you are officially enrolled, participate in an intake interview, and provide documentation: See documentation guidelines at (http://equity.psu.edu/sdr/guidelines). If the documentation supports your request for reasonable accommodations, your campus disability services office will provide you with an accommodation letter. Please share this letter with your instructors and discuss the accommodations with them as early as possible. You must follow this process for every semester that you request accommodations.

Psychological Services

Many students at Penn State face personal challenges or have psychological needs that may interfere with their academic progress, social development, or emotional wellbeing. The university offers a variety of confidential services to help you through difficult times, including individual and group counseling, crisis intervention, consultations, online chats, and mental health screenings. These services are provided by staff who welcome all students and embrace a philosophy respectful of clients' cultural and religious backgrounds, and sensitive to differences in race, ability, gender identity and sexual orientation.

- Counseling and Psychological Services at University Park (CAPS) (http://studentaffairs.psu.edu/counseling/): 814-863-0395
- Penn State Crisis Line (24 hours/7 days/week): 877-229-6400
- Crisis Text Line (24 hours/7 days/week): Text LIONS to 741741

Educational Equity

Penn State takes great pride to foster a diverse and inclusive environment for students, faculty, and staff. Consistent with University Policy AD29, students who believe they have experienced or observed a hate crime, an act of intolerance, discrimination, or harassment that occurs at Penn State are urged to report these incidents as outlined on the University's Report Bias webpage (http://equity.psu.edu/reportbias/).