EGMT 1520: Empirical & Scientific Engagement
From Language to Data

Instructor Erik Fredner (he/him)
Email fredner@virginia.edu
Office Nau Hall, Room 341

Description

People who share a language share its words. Yet the way each of us uses those words is nearly as unique as our fingerprints. Stranger still, our distinctive patterns of speech and writing do not solely depend upon rare words we may think distinguish ourselves (e.g. ballerina, foodie), but common ones we scarcely think about using like the, a, or of. Using data about such words, forensic linguists can identify the author of a document accurately enough to serve as evidence in court.

In this course, we study how and why we create data from language. We consider different types of texts, including speech, digital communication, and literature. What do we gain when we transform language into data? What do we lose? This course not only engages questions from academic fields like digital humanities, but also practical questions of everyday life in the twenty-first century. Google became powerful by transforming language into data. What does understanding that process teach us about how we live today?

Goals

Students will be able to…

- Define and delimit empirical evidence that can be derived from English-language texts
- Evaluate what we can know with respect to the digital cultural record
- Identify and articulate insights from reading academic articles
- Distinguish between frequentist and probabilistic knowledge derived from language data
- Articulate strengths and weaknesses of empirical approaches to language

Syllabus

The official course syllabus is the version of this document currently available on Collab.

Calendar

Date	*Mtg.*	*Guiding question*	*Readings*	*Assignment due*
Jan 20	1	Why transform language into data?	-	-
Jan 25	2	Why not transform language into data?	Chun[1] Zuboff[2] Haugen (22:00-33:50)[3]	My data
Jan 27	3	What is (big) data?	"What is Data?"[4] Grimmer and Stewart[5] Drucker[6] D'Ignazio and Klein[7]	Insight
Feb 1	4	What is metadata?	Irani[8] Fields and Fields[9] Krause[10]	Insight
Feb 3	5	What is information?	Tse[11] "Infinite Monkey Theorem"[12] Borges[13]	Insight
Feb 8	6	What did people who transformed language into data do with it early on?	Holmes and Kardos[14] Mendenhall[15] Weizenbaum[16]	Insight
Feb 10	7	How does this research happen?	Nguyen et al.[17] Nelson[18]	Insight
Feb 15	8	How do we transform language into data?	AI Dungeon[19] Jurafsky and Martin[20] Bode[21]	Insight
Feb 17	9	What is a corpus?	Algee-Hewitt and McGurl[22] Risam[23]	Insight
Feb 22	10	What can counting teach us?	Bailey[24] Gavin[25] Porter[26] Klein[27]	My corpus
Feb 24	11	What can we learn beyond counting?	Heuser and Le Khac[28] Blei[29] Parrish[30]	Insight
Mar 1	12	What is a model?	Yee and Chu[31] Seabrook[32] Piper[33]	Insight
Mar 3	13	-	-	Engagements Experience
Mar 15	14	-	-	Presentations

Health

Current UVA policy on coronavirus is available here. Noncompliance with coronavirus policies will be penalized as described in SEC-045.

Students who miss class due to illness will receive excused absences and alternative assignments. If you have been exposed, or believe you may have been exposed to COVID, contact Student Health.

Grading

Aspect	*Percentage of final grade*
Attendance	10
Engagements Experience	10
Final Project	30
Insights	20
My Corpus	10
My Data	10
Peer Evaluation	10

- If you have obligations that require you to miss specific classes (e.g. away games), please let me know about them as far in advance as possible.
- Students requesting an extension due to *foreseeable* circumstances must email me 48 hours prior to the deadline.

Email

I reply to email within two business days. Do not expect replies over the weekend.

Office Hours

I will schedule office hours in person or virtually at a mutually convenient time.

Engagements Experience

See the Engagements Experience assignment on Collab for instructions and form.

Notice of Non-Discrimination

The University of Virginia does not discriminate on the basis of age, color, disability, gender identity or expression, marital status, military status (which includes active duty service members, reserve service members, and dependents), national or ethnic origin, political affiliation, pregnancy (including childbirth and related conditions), race, religion, sex, sexual orientation, veteran status, and family medical or genetic information, in its programs and activities as required by Title IX of the Education Amendments of 1972, Americans with Disabilities Act of 1990, as amended, Section 504 of the Rehabilitation Act of 1973, Titles VI and VII of the Civil Rights Act of 1964, Age Discrimination Act of 1975, Governor's Executive Order Number One (2018), and other applicable statutes and University policies. UVA prohibits sexual and gender-based harassment, including sexual assault, and other forms of interpersonal violence.

Reporting Discriminatory Conduct

Per UVA policy HRM-040, I am a Responsible Employee. If you mention prohibited conduct to me, I am required to report it. Prohibited conduct includes, but is not limited to, sexual and gender-based harassment and violence, bias and discrimination/harassment, hazing, interference with speech rights, threats or acts of violence. If you wish to report prohibited conduct online, use UVA's Just Report It.

Accommodations

In accordance with the ADA, as amended, and Section 504 of the Rehabilitation Act, the University of Virginia offers an array of individualized accommodations and services to qualified students with disabilities. Accommodations are determined using an interactive process between the student and Student Disability Access Center staff. Contact the Student Disability Access Center with questions.

Religious Accommodations

Students who wish to request academic accommodation for a religious observance should submit their request to me by email as far in advance as possible. If you have questions or concerns about your request, you can contact the University's Office for Equal Opportunity and Civil Rights at UVAEOCR@virginia.edu or 434-924-3200. Accommodations do not relieve you of the responsibility for completion of any part of the coursework you miss as the result of a religious observance. More information about religious accommodations is available here, and frequently asked questions are here.

Campus Resources

- If you are struggling with mental health, UVA Counseling and Psychological Services can help. The TimelyCare app also provides mental health support.
- The UVA Writing Center will advise you on any stage of writing.
- If you are struggling to manage money, a Peer Financial Counselor can help.
- UVA helps students experiencing food insecurity get free nutritious food.
- Charlottesville's Sexual Assault Resource Agency responds to sexual and/or gender-based violence: 434-977-7273
- Students who wish to improve their English may be interested in the Sundberg International Center's programming.
- Students who wish to improve study skills like time management may benefit from UVA's resources for Academic Success.

Acknowledgments

This syllabus has benefitted from syllabi by Dan Sinykin, David Bamman, David Mimno, Laura McGrath, Lauren Klein, Mark Algee-Hewitt, Melanie Walsh, Ryan Heuser, and Ted Underwood.

---

[1] Wendy Hui Kyong Chun, *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition* (Cambridge, Massachusetts: The MIT Press, 2021).

[2] Shoshana Zuboff, "You Are the Object of a Secret Extraction Operation," *The New York Times*, November 12, 2021, sec. Opinion, https://www.nytimes.com/2021/11/12/opinion/facebook-privacy.html.

[3] The Information Society Project at Yale Law School, *The Facebook Files: What Next? Panel 1: The Activists*, 2021, https://www.youtube.com/watch?v=YUubanGIZc0.
[4] University of Guelph Library, *What Is Data?*, 2019, https://www.youtube.com/watch?v=pg12U1BAnoA.
[5] Justin Grimmer and Brandon M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21, no. 3 (2013): 267–97, https://doi.org/10.1093/pan/mps028.
[6] Johanna Drucker, "Humanities Approaches to Graphical Display," *Digital Humanities Quarterly* 005, no. 1 (March 10, 2011).
[7] Lauren F. Klein and Catherine D'Ignazio, "Numbers Don't Speak For Themselves," in *Data Feminism* (Cambridge, Massachusetts: The MIT Press, 2020).
[8] Lilly Irani, "Justice for Data Janitors," in *Think in Public: A Public Books Reader*, ed. Sharon Marcus and Caitlin Zaloom, Public Books Series (New York: Columbia University Press, 2019), 23–39.
[9] Karen E. Fields and Barbara Jeanne Fields, *Racecraft: The Soul of Inequality in American Life* (Verso Books, 2014).
[10] Heather Krause, "Data Biographies: Getting to Know Your Data," Global Investigative Journalism Network, March 27, 2017, https://gijn.org/2017/03/27/data-biographies-getting-to-know-your-data/.
[11] David Tse, "How Claude Shannon Invented the Future," Quanta Magazine, December 22, 2020, https://www.quantamagazine.org/how-claude-shannons-information-theory-invented-the-future-20201222/.
[12] "Infinite Monkey Theorem," in *Wikipedia*, December 28, 2021, https://en.wikipedia.org/w/index.php?title=Infinite_monkey_theorem&oldid=1062391503.
[13] Jorge Luis Borges, "The Library of Babel," in *Collected Fictions*, trans. Andrew Hurley, Penguin Classics Deluxe Edition (New York, NY: Penguin Books, 1998).
[14] David I. Holmes and Judit Kardos, "Who Was the Author? An Introduction to Stylometry," *Chance* 16, no. 2 (2003): 5–8.
[15] Thomas Corwin Mendenhall, "The Characteristic Curves of Composition," *Science* 9, no. 214 (1887): 237–49.
[16] Joseph Weizenbaum, "ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine," *Communications of the ACM* 9, no. 1 (1966): 36–45.
[17] Dong Nguyen et al., "How We Do Things with Words: Analyzing Text as Social and Cultural Data," *Frontiers in Artificial Intelligence* 3 (2020): 62.
[18] Laura K. Nelson, "Computational Grounded Theory: A Methodological Framework," *Sociological Methods & Research* 49, no. 1 (February 2020): 3–42, https://doi.org/10.1177/0049124117729703.
[19] "About AI Dungeon," AI Dungeon, January 14, 2022, https://play.aidungeon.io/main/about.
[20] Dan Jurafsky and James Martin, "Regular Expressions, Text Normalization, Edit Distance," in *Speech and Language Processing*, 3rd (Draft), 2019, https://web.stanford.edu/~jurafsky/slp3/edbook_oct162019.pdf.
[21] Katherine Bode, "The Equivalence of 'Close' and 'Distant' Reading; or, toward a New Object for Data-Rich Literary History," *Modern Language Quarterly* 78, no. 1 (2017): 77–106.
[22] Mark Algee-Hewitt and Mark McGurl, "Between Canon and Corpus: Six Perspectives on 20th-Century Novels," *Stanford Literary Lab Pamphlets*, no. 8 (January 2015), https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.
[23] Roopika Risam, *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy* (Northwestern University Press, 2018).
[24] Moya Bailey, *Misogynoir Transformed: Black Women's Digital Resistance*, Intersections : Transdisciplinary Perspectives on Genders and Sexualities (New York: New York University Press, 2021).
[25] Michael Gavin, "Is There a Text in My Data? (Part 1): On Counting Words," *Journal of Cultural Analytics* 1, no. 1 (2020): 11830.
[26] J.D. Porter, "Popularity/Prestige," *Pamphlets of the Stanford Literary Lab* 17 (September 2018), https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf.
[27] Lauren F. Klein, "The Image of Absence: Archival Silence, Data Visualization, and James Hemings," *American Literature* 85, no. 4 (2013): 661–88.
[28] Ryan Heuser and Long Le-Khac, "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method," *Pamphlets of the Stanford Literary Lab*, no. 4 (May 2012), https://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.
[29] David M. Blei, "Probabilistic Topic Models," *Communications of the ACM* 55, no. 4 (April 1, 2012): 77, https://doi.org/10.1145/2133806.2133826.
[30] Allison Parrish, "NLP Concepts with SpaCy," Gist, August 11, 2017, https://gist.github.com/aparrish/f21f6abbf2367e8eb23438558207e1c3.
[31] Stephanie Yee and Tony Chu, "A Visual Introduction to Machine Learning," accessed January 11, 2022, http://www.r2d3.us/visual-intro-to-machine-learning-part-1/.

[32] Joshua Seabrook, "The Next Word: Where Will Predictive Text Take Us?," The New Yorker, October 14, 2019, https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker.
[33] Andrew Piper, *Enumerations: Data and Literary Study* (University of Chicago Press, 2018).