Research Experience: Text-as-Data Research Methods for the Social Sciences
ECON 396
St. Olaf College
Meeting Time: M-F 1:00-3:00pm; Room: HH 201

Instructor: Colin Harris
Email: Harris14@stolaf.edu
Office Location: Holland Hall 304
Office Hours: TBD

Course Description: This course introduces students to various text-as-data methods used in social science research. Emphasizing hands-on experience, it covers a range of techniques from basic text processing to advanced computational text analysis. Students will engage with both theoretical foundations and practical applications, analyzing real-world datasets to answer social science research questions.

Topics Outline:

0. General Readings
    a. Bryan (2023), "A User's Guide to GPT and LLMs for Economic Research"
    b. Ash and Hansen (2023), "Large Language Models for Economic Research: Four Key Questions"
    c. Athey and Imbens (2019), "Machine Learning Methods That Economists Should Know About"
    d. Korinek (2023), "Generative AI for Economic Research: Use Cases and Implications for Economists"

1. Course Overview; The Research Process; Choosing a Topic

2. Economics; Choosing a Question; Crash Course in Econometrics
    a. Becker (1976), "The Economic Approach to Human Behavior"
    b. Harris, Myers, Briol, and Carlen (2022), "The Binding Force of Economics"
    c. Leeson (2020), "Economics is not Statistics (and vice versa)"
    d. Ruhm (2019), "Shackling the Identification Police?"

3. Text-as-Data
    a. Grimmer, Roberts, and Stewart (2022), *Text as Data*, Chapters 1-2
    b. Gentzkow, Kelly, and Taddy (2019), "Text as Data"
    c. Ash and Hansen (2023), "Text Algorithms in Economics"
    d. Grimmer and Stewart (2013), "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts"

4. Text Data; Data; Web Scraping
    a. Grimmer, Roberts, and Stewart (2022), *Text as Data*, Chapters 3-4

5. Preprocessing and Data Cleaning; RegEx
    a. Grimmer, Roberts, and Stewart (2022), *Text as Data*, Chapter 5, 15
    b. Denny and Spirling (2018), "Text Preprocessing for Unsupervised Learning"
    c. Python RegEx: https://www.w3schools.com/python/python_regex.asp

6. Dictionary Approaches
    a. Grimmer, Roberts, and Stewart (2022) *Text as Data*, Chapters 16
    b. Tausczik and Pennebaker (2010), "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods"
    c. Gentzkow and Shapiro (2010), "What Drives Media Slant? Evidence from U.S. Daily Newspapers."
    d. Enke (2020), "Moral Values and Voting"
    e. Michalopoulos and Xue (2021), "Folklore"
    f. Harris, Myers, and Kaiser (2023), "The Humanizing Effect of Market Interaction"
    g. Harris (2023), "Sentiments Outside the Lab"

7. Distribution, Distance, and Similarity
    a. Grimmer, Roberts, and Stewart (2022), *Text as Data*, Chapters 6-7
    b. Almelhem, Iyigun, Kennedy, and Rubin (2023), "Enlightenment Ideals and Belief in Science in the Run-up to the Industrial Revolution: A Textual Analysis"
    c. Ash and Chen (2019), "Case Vectors: Spatial Representations of the Law Using Document Embeddings"
    d. Kelly, Papanikolaou, Seru, and Taddy (2021) "Measuring Technological Innovation over the Long Run"
    e. Bertrand, Bombardini, Fisman, Hackinen, and Trebbi (2021), "Hall of Mirrors: Corporate Philanthropy and Strategic Advocacy"

8. Topic Models
    a. Grimmer, Roberts, and Stewart (2022), *Text as Data*, Chapters 10-14
    b. Hansen, McMahon, and Prat (2018) "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach"
    c. Ash, Morelli, and Vannoni (2023), "More Laws, More Growth? Evidence from U.S. States"
    d. Barron, Huang, Spang, and DeDeo (2018), "Individuals, Institutions, and Innovation in the Debates of the French Revolution"
    e. Jelveh, Kogut, and Naidu (2022), "Political Language in Economics"

9. (Un)Supervised Learning
    a. Grimmer, Roberts, and Stewart (2022), *Text as Data*, Chapters 17-20
    b. Kelly, Manela, and Moreira (2021), "Text Selection"
    c. Gentzkow, Shapiro, and Taddy (2019), "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech"
    d. Widmer, Ash, and Galletta, (2022), "Media Slant is Contagious"
    e. Vafa, Naidu, and Blei (2020), "Text-Based Ideal Points"

10. Word Embeddings
    a. Grimmer, Roberts, and Stewart (2022), *Text as Data*, Chapter 8
    b. Rodriguez and Spirling (2022), "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research"
    c. Gennaro and Ash (2022), "Emotion and Reason in Political Language"
    d. Kozlowski, Taddy, and Evans (2019) "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings"
    e. Rheault and Cochrane (2019) "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora"

11. Linguistic Parsing
    a. Grimmer, Roberts, and Stewart (2022), *Text as Data*, Chapter 9
    b. Ash, Gauthier, and Widmer (2023), "Relatio: Text Semantics Capture Political and Economic Narratives"
    c. Roos and Reccius (2023) "Narratives in Economics"
    d. Ash, Chen, and Naidu (2023), "Ideas Have Consequences: The Impact of law and Economics on American Justice"
    e. Antoniak, Mire, Sap, Ash, and Piper (2023), "Where Do People Tell Stories Online? Story Detection Across Online Communities"
    f. Gehring, Adema, and Poutvaara (2022), "Immigrant Narratives"

12. Causal Inference with Text Data
    a. Grimmer, Roberts, and Stewart. *Text as Data*, Chapters 22-27
    b. Keith, Jensen, and O'Connor (2020), "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates"
    c. Zeng, Gensheimer, Rubin, Athey, and Shachter (2022), "Uncovering Interpretable Potential Confounders in Electronic Medical Records"
    d. Egami, Fong, Grimmer, Roberts, and Stewart (2022), "How to Make Causal Inferences Using Texts"
    e. Rodriguez, Spirling, and Stewart (2023), "Embedding Regression: Models for Context-Specific Description and Inference"

13. More Applied Readings
    a. Baker, Bloom, and Davis (2016), "Measuring Economic Policy Uncertainty"
    b. Baker, Davis, and Levy (2022), "State-Level Economic Policy Uncertainty"
    c. Baker, Bloom, Davis, and Sammon (2022) "What Triggers Stock Market Jumps?"
    d. Ban, Grimmer, Kaslovsky, and West (2022) "How Does the Rising Number of Women in the U.S. Congress Change Deliberation?"
    e. Benoit, Munger, and Spirling (2019) "Measuring and Explaining Political Sophistication through Textual Complexity"
    f. Blaydes, Grimmer, and McQueen, (2018) "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds"
    g. Bloom, Hassan, Kalyani, Lerner, and Tahoun (2020), "The Geography of New Technologies"
    h. Boyer, Delemotte, Gauthier, Rollet, and Schmutz (2023), "Social Media and the Dynamics of Protests"

i. Butler, Kousser, and Oklobdzija (2023), "Do Male and Female Legislators Have Different Twitter Communication Styles?"
j. Bybee, Kelly, Manela, and Xiu (2023), "Business News and Business Cycles"
k. Cao, Lindo, and Zhong (2023) "Can Social Media Rhetoric Incite Hate Incidents? Evidence from Trump's 'Chinese Virus' Tweets"
l. Carter and Carter (2023), *Propaganda in Autocracies*
m. Djourelova (2023), "Persuasion through Slanted Language: Evidence from the Media Coverage of Immigration"
n. Giavazzi, Iglhaut, Lemoli, and Rubera (2023), "Terrorist Attacks, Cultural Incidents, and the Vote for Radical Parties: Analyzing Text from Twitter"
o. Gorodnichenko, Pham, and Talavera (2023), "The Voice of Monetary Policy"
p. Hansen, Lambert, Bloom, Davis, Sadun, and Taska (2023), "Remote Work across Jobs, Companies, and Space"
q. Hassan, Hollander, Van Lent, and Tahoun (2023), "The Global Impact of Brexit Uncertainty"
r. Hopkins, Lelkes, and Wolken (2023), "The Rise of and Demand for Identity-Oriented Media Coverage"
s. Kaslovsky and Kistner (2023), "Responsive Rhetoric: Evidence from Congressional Redistricting"
t. Larsen and Thorsrud (2019), "Business Cycle Narratives"
u. Lucas, Nielsen, Roberts, Stewart, Storer, and Tingley (2017), "Computer-Assisted Text Analysis for Comparative Politics"
v. Magness and Makovi (2023), "The Mainstreaming of Marx"
w. Moreno-Medina (2023), "Local Crime News Bias: Extent, Causes, and Consequences"
x. Moreno-Medina, Ouss, Bayer, and Ba (2023), "Officer-Involved: The Media Language of Police Killings"
y. Nyman, Kapadia, and Tuckett (2021), "News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment"
z. Rheault, Beelen, Cochrane, and Hirst (2016), "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis"
aa. Schneider-Strawczynski and Valette (2022), "Media Coverage of Immigration and the Polarization of Attitudes"
ab. Siege, Nikitin, Barbera, Sterling, Pullen, Bonneau, Nagler, and Tucker (2021), "Trumping Hate on Twitter? Online Hate Speech in the 2016 U.S. Election Campaign and its Aftermath"
ac. Spirling (2016), "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915"
ad. Thrall (2023), "Informational Lobbying and Commercial Diplomacy"

Text Data:
We will explore available text data in section 4. Beyond what we find, I have the following available to use:

1. *New York Times* articles from 1981-2021
2. *Wall Street Journal* articles from 1984-1990; 1998-2021

3. *Washington Post* articles from 1978-2021
4. State-level newspaper articles from 1780-1960
5. Fox News, MSNBC, CNN, Fox Business, PBS, NBC, CNBC, CBS, ABC, Bloomberg, BBC, Al Jazeera, Russia Today, CSPAN, CSPAN2, CSPAN3 transcripts from 2009-2023.
6. Presidential Documents (speeches, executive orders, etc.) going back to G. Washington.
7. Campaign Documents (speeches, press releases, etc.) going back to 1995.
8. Congressional speeches
9. House of Commons speeches
10. Supreme Court Opinions
11. Case Law from all fifty states.
12. Municipality Zoning Laws
13. The Federal Code
14. Local Government meeting transcripts
15. Titles, authors, and print location of books published between 1450-1700
16. Alex Jones transcripts
17. And more!

Programming:
I will be using Python to illustrate the various methods. You are welcome to use another programming language. I will also primarily "code" using ChatGPT or GitHub's Copilot, which you are welcome to do as well. Nevertheless, it will be useful to familiarize yourself with the basics of whichever language you choose to work with. There are countless tutorials and instructional material online for free.

Download and install Visual Studio Code: https://code.visualstudio.com/
Register for GitHub Copilot: https://docs.github.com/en/copilot/quickstart

Grading: Your grade for this course will be based on self-assessment, peer assessment, and assessment from me.

Your final grade will be a calculated as followed:

Self-Assessment	70%
Peer Assessment:	10%
My Assessment:	20%
Final Grade:	100%

Grading Scale:

A+: 100-98	A: 97-93	A-: 92-90	B+: 89-88	B: 87-83	B-: 82-80
C+: 79-78	C: 77-73	C-: 72-70	D: 69-60	F: <60	