ECON 630: Using Text as Data

University of San Francisco
Meeting: TuTh 4-6pm

Summer 2023 Session 3 CRN: 30529
Location: Lo Schiavo Science 209

Instructor Information:

Professor: Konrad Posch
E-mail: kposch@usfca.edu
Mail: McLaren 101 (Economics Main Office)

Office Location: McLaren 110
Office Hours: In person, Thursdays 1-2pm,
(zoom by appointment)

Course Description

This course provides an introduction to how social scientists use text as data when seeking to answer research questions about the world. While it is a lab-based course that will center a hands-on approach in gathering, processing, and analyzing data, it will also be heavily grounded in research methods. To that end, while you will be introduced to a number of tools for using text as data, you will also be expected to understand when and how those tools are appropriate for particular research questions (and when they might lead to tidy but spurious conclusions).

While this course does ***not*** require advanced knowledge of mathematical approaches to economics, you will be expected to be gather the economics (or other) modeling knowledge you need to answer the research question you pose in your final project. Speak with the professor and your advisor if you are not sure whether you should take this course.

Canvas site: https://usfca.instructure.com/courses/1616311

Grading Breakdown & Policies

Attendance & Participation	10%
Labs (3 total, 1 per week, assigned Tu due M, start in week 2)	30%
Final Project Prospectus (due 12pm (noon), Tuesday 7/25)	5%
Prospectus Peer Reviews (in class, Tuesday 7/25)	5%
Final Project Presentation (Last day of Class)	10%
Final Project Writeup (1500 words min, 20pg with visualizations)	40%

Everyone taking the class for credit (letter grade or pass/fail) is expected to complete all assignments or they will receive an F. Conflicts and crises do arise and will be dealt with graciously, but it is your responsibility to communicate with the Professor as early as possible to sort out issues.

Grading Scale

Percentage grade will be mapped to letter grades in the following manner (percentages will be rounded up to the unit digit):

A+	A	A-	B+	B	B-	C+	C	C-	F
97-100%	93-96%	90-92%	87-89%	83-86%	80-82%	77-79%	73-76%	70-72%	0-69%

Grade Disputes

All written assignments in this class require you to take a position and defend it. As with any such endeavor, there is a certain amount of subjectivity to the grading of how successfully you are able to defend your position within the constraints of the courses and the expectations of a USF student. I therefore highly recommend that you not attempt to dispute your grades unless you believe the grade you have earned is more than 2/3 of a grade category (e.g. B- --> B+) different from the grade you believe you should have earned. If you are unsure why you earned the grade you did, speak with your instructor so you can improve on future assignments

If you do wish to pursue a grade dispute, please note: All grade disputes must be made in writing, in not less than a paragraph, not more than a page. The dispute should outline very specifically why you feel that you received a grade in error and should not contain information about what kinds of grades you usually get or the effort you put in. You must wait at least 24 hours after receiving your grade to raise the issue of a dispute with the instructor, but you must turn in your written dispute within a week of getting the grade. There are no exceptions to this policy.

Office Hours (Thursdays 1-2pm, McLaren 110, and zoom by appointment)

By default, I run my office hours as group office hours for substantive questions related to course material. I've found that this creates the maximum opportunity for students to learn from each other as well as to lower the barriers to come to office hours. Everyone is always welcome at office hours, but I do ask that if you wish to come you try to come with at least one question or topic you would like to discuss to get the conversation started. I heartily encourage you to stay as long as you are interested in the conversation as some of the greatest learning resources at USF are your fellow students, how they think, and how they ask questions. If you wish to discuss an individual matter privately, please let me know using the web form and we will excuse ourselves from the group and speak in private.

Texts & Supplies

You are required to purchase the following book (in paper or digital form) for this class:

> Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

This book will be referred to as "Grimmer, Roberts, & Stewart" or "The Book" in class and in the list below; it is available for purchase at the book store and through all major online platforms. This book, written by three icons of social science computational text methods, forms a backbone for understanding how and when to use what methods. It is thus a worthwhile investment even on dead trees because it does not contain potentially dated coding advice but instead focuses on enduring concepts of research design.

For research guidance, I suggest purchasing the following book that talks about research design and writing:

Booth, Wayne C., Gregory G. Colomb, Joseph M. Williams, Joseph Bizup, and William T. FitzGerald. 2016. *The Craft of Research, Fourth Edition*. University of Chicago Press.

I will post several vital chapters of (Booth et. al 2016), but having the entire book can be very helpful when you get stuck on a conceptual or compositional issue.

Other readings (as needed) will be posted under modules on the Canvas Site.

Technology Stack

You must have a laptop running a modern generation of Windows, Mac OS, or Linux.

This course will use R and R Studio for text analysis. You should download and install them before the second class.

Python (specifically Anaconda) and Jupyter Notebook will be needed for the class on text scraping.

You are welcome to use whatever additional analysis programs are needed for your final project (e.g. Stata, Mathematica, Latex, etc), but note that the final submission will need to be a PDF that incorporates textual analysis and explanation, code syntax, code output, and graphics. The assignments are designed to achieve this through R markdown files (Rmd) and I will support students in this implementation; if you decide to go rogue, you do so at the risk that I will not be able to explain how to use your software suite to you. I'm happy to try to help, but students working with Rmd will receive priority in office hours.

ChatGPT and Coding…

Perhaps unique among your courses so far, this class ***expects*** you to use ChatGPT (https://chat.openai.com/) to help you with your coding. Writing syntax is a constant battle of trial & error where a single missing punctuation mark can break your code or lead to misleading results.

In the bad old days, searching for advice on the internet meant hoping someone else had the exact problem you did AND that, when they finally found an answer, they would post it rather than just finish their project and go to bed.

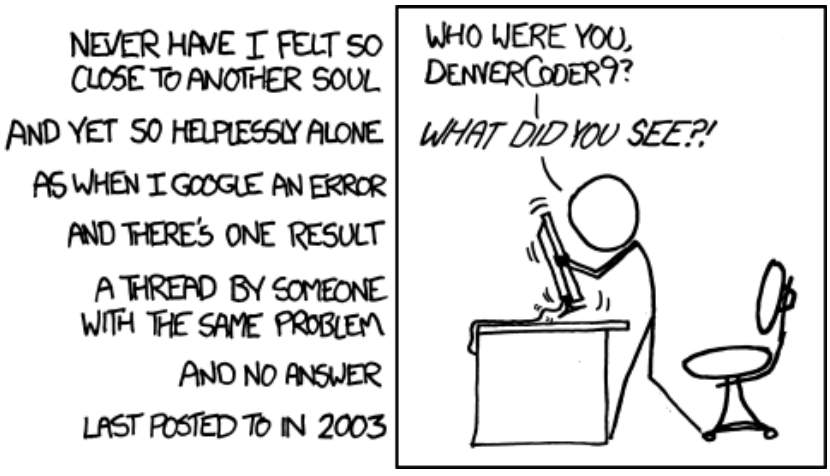

Figure 1: The Bad Old Days: Wisdom of the Ancients (https://xkcd.com/979/)

For better or worse, we live in the troubling current times when you no longer have to use search queries to find snippets of poorly formatted old code with partially incorrect and partially incomplete solutions. Now, you can ask a Large Language Model (LLM) like ChatGPT how to do something.

BE WARNED: ChatGPT can make mistakes! You are ultimately responsible for your code, the analysis you base on it, and the conclusions you draw. Assignment rubrics will require you to explain what your code is doing.

NOTE: ChatGPT *does* give you explanations to go with the syntax. You are **absolutely welcome to use the syntax** (although you might want to note that ChatGPT helped you with it). You **MAY NOT copy ChatGPT's explanations word-for-word without attribution** (direct quotes, citation) and an explanation in your own words of what is going on and why this code solves your *research* or *analytic* problem. Copying the explanations without attribution is a form of academic dishonesty and violates USF academic integrity.

ChatGPT is an essential resource for analytic coding in 2023; it's just not worth pretending otherwise so we won't. But just like algebra back in high school and calculus in college, even though a computer can do it doesn't mean you don't have to think about it. You MUST learn what the computer can and cannot do as well as how to tell the difference. We'll work on developing that intuition together.

Course Learning Outcomes

By the end of this course, students will:

1) Understand when (and when not) to use various computational text analysis methods to answer specific research questions
2) Understand how to use R to perform textual data cleaning and analysis
3) Understand how to use Python to gather textual data

Program Learning Outcomes

This course is a component of the MS Applied Economics. It contributes to the following program learning outcomes:

PLO-1: Economic Data Manipulation (PLO1) : Students will be able to effectively use modern programming languages to clean, organize, query, summarize, visualize, and model large volumes and varieties of data.

PLO-2: Economic Data Analytics (PLO2): Students will possess a theoretical and applied understanding of the use of econometrics and statistics for descriptive and causal inference.

PLO-5: Economic Communication: Students will be able to communicate their research approach and findings at an excellent level, both in writing and verbally.

Attendance Policy

Participation (10% of overall grade)

Participation in this course will be evaluated by your instructor based on your active engagement in the class. Participation requires showing up and being involved in the course.

This means lecture Q&A, discussion, and office hours. Please read the assignment descriptions on the Canvas site and talk to your instructor well in advance of the end of the session if you are concerned with your participation. If you would like to earn a higher participation grade, participate more actively whether in lecture Q&A, discussion, or office hours.

You are **required to attend all class sessions**. Please contact your instructor in advance using the Excused Absence Request Form if you need be excused from a class session.

A **makeup assignment** consisting of a 500 word (approx. 2 page, double spaced) response to the assigned reading for the class period must be turned in to the professor **no later than one week after an absence for the absence to be excused**. This response need not be particularly formal, but it should do more than summarize the chapter; it should reflect on the meaning of the chapter to you, perhaps with respect to your final project.

No more than one absence may be excused during the term due to the very short term of summer courses.

Note: No one may be absent on the last day of class without serious harm to their grade as that is when we will be doing in-class Project Presentations (10% of your grade for the course).

Lab Assignments

As a hands on methods course, you will be responsible for completing three labs through the middle of the course (weeks 2,3 and 4). Each lab will consist of a series of coding problems which you must resolve and then explain how and why your solution works. Equal weight will be given to the code solution and the quality of your explanation, so do not skip the explanations!

You are welcome to collaborate on this with your classmates (and with ChatGPT), but **the final explanations of your syntax must be your own**. I encourage you to work together and talk over any difficulties, but do NOT split the assignment up and then create identical final submissions.

All labs will be completed in Rmd and submitted as a PDF file through Canvas.

Labs will be circulated after lecture on Tuesdays and will be due the following Monday by 12pm (noon).

The Final Project (Total of 60% of Grade)

The final project will require you to perform an original analysis of textual data using one or more of the methods we learned in this course to answer a research question of interest to you. You are encouraged to think about what you would like to do early and speak with me about it long before it is due. The assignments below will provide incremental progress as well as assess the final outcome.

Final Project Prospectus (5% of overall grade) (Due by 12pm (noon) Tuesday, 7/25)

A 500 word (approx. 2 pages, double spaced) explanation of the research question, textual method(s), the data source(s), and the specific potential challenges you expect to encounter (such as data access, analytic complexity, etc). If you have not fully settled on one research question by this point, you may outline two, but you must then concretely explain the practical pros and cons of each so that, in peer evaluations, we can help you settle on the most feasible one.

Prospectus Peer Reviews (5% of overall grade) (In class on Tuesday 7/25)

Given the small size of our course this term, all students will be expected to read and comment on all other students prospectuses during class time on Tuesday, 7/25. Further details will be given on the day of, but the goal of this review is to help your colleagues define a project that is (in order of priority):

1) Achievable during this term
2) Appropriately selected methods to answer research question
3) Interesting to the student for articulable reasons

Final Project Presentation (10% of overall grade)

The final class meeting (Thursday, August 10th) will be a mock research conference where you will each present your final projects in the format of a scholarly conference.

You will each have 15 minutes of class time to present your research. You are required to have a powerpoint slide deck (see the templates USF provides here) which you will submit in advance of your presentation (uploaded to Canvas).

Your presentation should include:

1) Your research question
2) An explanation of your method
3) A justification of why this method in (2) fits your question in (1)
4) Appropriate visualizations, tables, and charts based on your methods
5) Your research answer
6) A concluding statement about why we should (a) believe and (b) care about your findings based on your methodological choices in (2)

A rubric will be provided on Canvas before the presentations so you know how much each element contributes to your grade. You will be evaluated on how well you provide the six pieces of key information above as well as how well you present the material both on your slides as well as in person.

The presentation is also your last chance to get feedback to incorporate into your Final Project Writeup (explained below).

As with the final writeup below, I will not watch/read entire presentations in advance of the due date, but if you want to bring a work in progress to office hours I'm happy to workshop any questions or issues you have to help you succeed.

Final Project Writeup (40% of overall grade)

Due Monday August 14, 2023, the Monday AFTER the end of the class

Write a research report in the style of an online appendix that includes an annotated set of analyses, visualizations, and explanations. You can see an old online appendix of mine for context (yours will likely be much shorter).

You should expect to include all code syntax that is necessary for data gathering, processing, analysis, and visualization.

However, it is also at least as important that you include prose explanations of exactly what you did, why, and how it worked. You should expect to include **a minimum of 1500 words of explanations** NOT INCLUDING any code output or syntax.

Combined with the code and visualizations, your final output will likely be 15-20 pages.

This project must be submitted as a PDF that is most easily generated from an R markdown (Rmd) file in R studio.

I am always happy to talk with you in office hours as you develop your paper. While I will not read and comment on full drafts prior to submission, I'm happy to workshop paragraphs and pages with you during office hours.

Note: You are strongly encouraged to discuss your paper topic with the instructor well in advance of the due date.

Note as well: The Comparative Paper is **due the Monday AFTER the end of summer classes**. This is intended to give you some time to incorporate presentation feedback. If you plan to travel immediately after the term, you will want to submit your paper before you depart as the grade submission deadline means I cannot offer any extensions on this due date.

USF Policies & Legal Declarations

Students with Disabilities

The University of San Francisco is committed to providing equal access to students with disabilities. If you are a student with a disability, or if you think you may have a disability, please contact Student Disability Services (SDS) at sds@usfca.edu or 415 422-2613, to speak with a disability specialist. (All communication with SDS is private and confidential.) If you are eligible for accommodations, please request that your accommodation letter be sent to me as soon as possible; students are encouraged to contact SDS at the beginning of the semester, as accommodations are not retroactive. Once I have been notified by SDS of your accommodations we can discuss your accommodations and ensure your access to this class or clinical setting. For more information please visit the SDS website.

Behavioral Expectations

All students are expected to behave in accordance with the Student Conduct Code and other University policies (see http://www.usfca.edu/fogcutter/). Students whose behavior is

disruptive or who fail to comply with the instructor may be dismissed from the class for the remainder of the class period and may need to meet with the instructor or Dean prior to returning to the next class period. If necessary, referrals may also be made to the Student Conduct process for violations of the Student Conduct Code.

Academic Integrity

As a Jesuit institution committed to *cura personalis* -- the care and education of the whole person -- USF has an obligation to embody and foster the values of honesty and integrity. USF upholds the standards of honesty and integrity from all members of the academic community. All students are expected to know and adhere to the University's Honor Code. You can find the full text of the code online at http://myusf.usfca.edu/academic-integrity/. The policy covers:

- Plagiarism — intentionally or unintentionally representing the words or ideas of another person as your own; failure to properly cite references; manufacturing references.
- Working with another person when independent work is required.
- Submission of the same paper in more than one course without the specific permission of each instructor.
- Submitting a paper written by another person or obtained from the Internet (this includes papers generated by a large language model such as ChatGPT).

Counseling and Psychological Services (Caps)

Counseling and Psychological Services (CAPS) is a great source of support for issues of sadness, anxiety, loneliness, college adjustment, relationship struggles, and others not requiring medical intervention. CAPS offers online workshop series open to all students; consultations and referrals; and extensive **website resources**. In addition, CAPS All Hours "warmline" can be contacted by calling (855) 531-0761 or students can use the peer-led Crisis Textline by texting HOME to 741741. CAPS also offers remote individual and group teletherapy to students residing within California. (State regulations prevent provision of therapy across state lines.) The student may choose to talk either by video or telephone and can engage in Single Session Therapy (SST), brief ongoing therapy, or group therapy. There are no fees for services. Please call 415.422.6352 to make an appointment. Visit **CAPS** for more details. Students seeking off campus mental health services can also receive information and support from **Case Management** (part of the Office of the Dean of Students).

Confidentiality, Mandatory Reporting, and Sexual Assault

As an employee of USF, one of my responsibilities is to help create a safe learning and working community at USF. I have mandatory reporting responsibilities related to my role as a faculty member. I am required to share any disclosure or notice of information regarding sexual misconduct (including sexual harassment, sexual assault, dating or domestic violence, and stalking). In the event I become aware of any of these behaviors I will share this information, including any names, with the Title IX Office in order to connect our students to

the best resources and information about how the University can support you. Further information and resources may be found on the Title IX page. In addition:

- To report any sexual misconduct, students may contact the Title IX coordinator at jvarga@usfca.edu or (415) 422-4563) or use the Online Reporting Form.
- Students may speak to someone confidentially or report a sexual assault confidentially by contacting Counseling and Psychological Services at (415) 422-6352. Speaking with a licensed clinician at CAPS does not generate a report to the Title IX or any other university office.
- For an off-campus resource, contact San Francisco Women Against Rape (SFWAR) at (415) 647-7273.

Communication

All course communications, like all other USF communications, will be sent to your USF official email address. You are therefore strongly encouraged to monitor that email account.

Please put "ECON 630" in the subject line of your emails so that your instructor can respond to them in a timely manner. Emails which are received on a weekday will generally receive a response within **1 business day** while those received on a weekend will receive a response by the end of the next business day. Emergencies happen, when in doubt email to document the time when you had a problem and your instructor will endeavor to reply quickly if needed. **However, you should endeavor to never leave a matter so late as to need an answer faster than these response times.**

Generally, questions about procedures & assignments might be handled in email. Substantive questions about class material are always best handled in office hours or during lecture Q&A periods.

Announcements will be made through Canvas. Be sure to check your notification settings (including frequency and email address) to ensure that you receive them. They will always also be archived on the Canvas site for this course.

Course Schedule

Reading Expectations

The required readings for this course all come from the required text, Grimmer, Roberts, & Stewart. I may post some recent articles that use computational text methods on relevant days, but these are merely for inspiration and not required.

While different students will respond to different ways of presenting information, I would recommend that if you are struggling with an overly mathematical or technical representation of information in Grimmer, Roberts, & Stewart, you move past the math and focus on the conceptual explanation. In general, they also try to keep the math as approachable and non-vital as possible but sometimes math is unavoidable.

Part 1: Basics & Theory

Thursday (7/6): Introduction

- No Assigned Readings

Tuesday (7/11): Text Analysis & Social Science

- Grimmer, Roberts, & Stewart "Chapter 1: Introduction"
- Grimmer, Roberts, & Stewart "Chapter 2: Social Science Research and Text Analysis"

Thursday (7/13): Selecting a Corpus (Coding Techniques: Web Scraping & APIs)

Required

- Grimmer, Roberts, & Stewart "Part II: Selection & Representation" (p33-34)
- Grimmer, Roberts, & Stewart "Chapter 3: Principles of Selection & Representation"
- Grimmer, Roberts, & Stewart "Chapter 4: Selecting Documents"
- Grimmer, Roberts, & Stewart "Chapter 5: Bag of Words"

Optional (useful depending on final project topics):

- Grimmer, Roberts, & Stewart Chapters 6, 7, 8, & 9

Part 2: Discovery

Monday (7/17, 12pm noon): Lab 1 due thorough bCourses

Tuesday (7/18): Principles of Discovery & Characterizing Known Groups (Coding Techniques: Discriminating)

- Grimmer, Roberts, & Stewart "Part III: Discovery" (p99-102)
- Grimmer, Roberts, & Stewart "Chapter 10: Principles of Discovery"
- Grimmer, Roberts, & Stewart "Chapter 11: Discriminating Words"

Thursday (7/20): Bucketing: Discovering Groups from a Corpus (Coding Techniques: Clustering & Topic Modelling)

Required:

- Grimmer, Roberts, & Stewart "Chapter 12: Clustering"
- Grimmer, Roberts, & Stewart "Chapter 13: Topic Models"

Optional (useful depending on final project topics):

- Grimmer, Roberts, & Stewart "Chapter 14: Low-Dimensional Document Embeddings"

Part 3: Measurement

Monday (7/24, 12pm noon): Lab 2 due thorough Canvas

Tuesday (7/25): Principles of Measurement & Word Counting (Coding Technique: Word Counting)

- Grimmer, Roberts, & Stewart "Part IV: Measurement" (p171-172)
- Grimmer, Roberts, & Stewart "Chapter 15: Principles of Measurement"
- Grimmer, Roberts, & Stewart "Chapter 16: Word Counting"

Prospectus Due 7/25, Prospectus Peer Reviews in Class on 7/25

Thursday (7/27): Classification (Coding Technique: Classification)

Required:

- Grimmer, Roberts, & Stewart "Chapter 17: An Overview of Supervised Classification"
- Grimmer, Roberts, & Stewart "Chapter 18: Coding a Training Set"
- Grimmer, Roberts, & Stewart "Chapter 19: Classifiying Documents with Supervised Learning"

Optional (useful depending on final project topics):

- Grimmer, Roberts, & Stewart "Chapter 20: Checking Performance"
- Grimmer, Roberts, & Stewart "Chapter 21: Repurposing Discovery Methods"

Part 4: Inference

Monday (7/31, 12pm noon): Lab 3 due thorough Canvas

Tuesday (8/1): Principles of Inference

- Grimmer, Roberts, & Stewart "Part V: Inference" (p231-232)
- Grimmer, Roberts, & Stewart "Chapter 22: Principles of Inference"

Thursday (8/3): Prediction

- Grimmer, Roberts, & Stewart "Chapter 23: Prediction"

Tuesday (8/8): Causal Inference

Required

- Grimmer, Roberts, & Stewart "Chapter 24: Causal Inference"

Optional (useful depending on final project topics):

- Grimmer, Roberts, & Stewart "Chapter 25: Text as Outcome"
- Grimmer, Roberts, & Stewart "Chapter 25: Text as Treatment"
- Grimmer, Roberts, & Stewart "Chapter 25: Text as Confounder"

Thursday (8/10): Final Project Presentations

Monday (8/14, 11:59pm): Final Project Writeup Due through Canvas