



湖北工业大学
HUBEI UNIVERSITY OF TECHNOLOGY

Course Title	Text as Data: Computational Methods and Novel Data Sources
Course Code	ECON 5123
Semester	Winter 2027
Course Length	8 Weeks, 60 Contact Hours
Credits	4
Instructor	TBA
Office	TBA
Email	TBA
Prerequisite(s)	Graduate-level introductory statistics and research methods, or equivalent. Prior programming experience is helpful but not required. Students should be willing to learn reproducible computational workflows.
Antirequisite	ECON 5121 Text as Data: Computational Methods and Novel Data Sources (4 Weeks)

Course Description:

Text is generated across newspapers, political speeches, policy documents, corporate reports, online discussions, social media, digital archives, and many other settings. These materials contain valuable information about economic behaviour, institutions, markets, public opinion, and social change. However, such information is rarely available in a form that can be directly analysed using conventional quantitative methods. This course introduces students to the conceptual foundations and practical methods required to transform unstructured text into structured, research-ready data.

The course emphasizes applied, reproducible research. Students will develop workflows for collecting or accessing text, preparing and documenting corpora, measuring textual features, constructing quantitative indicators, and evaluating the validity and limitations of text-based measures. Topics include dictionary methods, supervised and unsupervised classification, topic models, embeddings, sentiment and stance analysis, and the use of large language models as research tools. These methods are examined in the context of substantive research questions rather than as stand-alone technical exercises.

A second emphasis is the expansion of empirical research through novel data sources. Students will evaluate how digital traces, administrative and archival documents, business communications, public consultation submissions, online forums, and other non-traditional sources can complement established statistics, surveys, and experimental data. Throughout the course, attention is given to

sampling, measurement error, selection bias, privacy, data governance, multilingual text, reproducibility, and responsible research practice.

Through lectures, methodological workshops, critical reading, and independent research, students will develop the ability to design and implement text-as-data studies addressing substantive questions in economics, management, public policy, and the wider social sciences.

Course Goals:

Students who successfully complete this course will demonstrate competency in the following general education core goals:

- **Critical Thinking Skills** – Students will engage in analytical thinking, demonstrating the ability to critically evaluate, synthesize, and apply knowledge to complex problems, and construct well-reasoned solutions and arguments.
- **Independent Research and Inquiry** – Students will conduct independent research, utilizing academic resources to explore relevant topics, formulating research questions, analyzing data, and presenting findings in a coherent, scholarly manner.
- **Problem-Solving and Application** – Students will apply theoretical concepts and methodologies learned in the course to real-world problems, demonstrating the ability to develop practical solutions informed by academic inquiry.
- **Global and Cultural Awareness** – Students will gain awareness of the global and cultural contexts relevant to the course, appreciating diverse perspectives and considering the implications of their studies in a broader, international context.

Student Learning Outcomes:

Upon completion of this course, students will be able to:

- Formulate theoretically motivated research questions that can be investigated using textual or other novel data sources.
- Identify, assess, and document appropriate text corpora and non-traditional data sources, including their provenance, coverage, and limitations.
- Design reproducible workflows for text acquisition, cleaning, preprocessing, storage, and documentation.
- Apply and interpret foundational text representations and computational methods, including frequency-based measures, dictionary analysis, supervised classification, clustering, and topic modelling.
- Explain the assumptions, advantages, and limitations of modern text-analysis approaches, including word embeddings and large language models.
- Evaluate text-based measures through appropriate validation strategies, benchmarks, error analysis, and robustness checks.
- Integrate text-derived variables with conventional quantitative data and interpret findings in light of the research design.
- Recognize and address potential problems involving sampling bias, construct validity, privacy, consent, copyright, fairness, and model bias.

- Produce reproducible research materials and communicate empirical findings, methodological choices, and limitations effectively to specialist audiences.

Textbooks:

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media.

Jurafsky, D., & Martin, J. H. *Speech and Language Processing*. Online draft textbook. Selected chapters on text classification, language models, and embeddings.

Course Requirements:

Reading Memos and Seminar Participation (10%)

Students will submit concise critical memos on selected readings and participate in seminar discussions. Memos should identify the research question, theoretical motivation, methodology, assumptions, empirical evidence, and limitations of the assigned work. Students are expected to engage critically with methodological choices and consider possible extensions or alternative approaches.

Text Data Acquisition and Reproducible Pipeline (15%)

Students will construct a documented workflow for acquiring or accessing an appropriate text corpus, cleaning and preprocessing the text, recording data provenance, and producing an analysis-ready dataset. The submission must include reproducible code or equivalent computational materials, a data dictionary, and a brief methodological report explaining the principal processing decisions.

Method Comparison and Validation Report (20%)

Students will apply and compare at least two computational text-analysis approaches to a clearly defined research task. The report must justify the selection of methods, describe the evaluation or validation strategy, present relevant error analysis or robustness checks, and discuss substantive interpretation and methodological limitations.

Novel Data Source Research Proposal (15%)

Students will develop an original postgraduate-level research proposal using text or another unconventional data source. The proposal should specify the research question, theoretical motivation, data access and sampling strategy, operationalization of key concepts, empirical or identification strategy where relevant, ethical considerations, and feasibility of the proposed research.

Final Text-as-Data Research Project (30%)

Students will complete an empirical research project that transforms text into quantitative evidence to investigate a substantive research question. The project must include a research paper, a reproducible analytical workflow, documentation of data and methods, validation and robustness analysis, and a critical discussion of

limitations. Projects may be completed individually or in instructor-approved teams, with individual contributions clearly documented.

Project Presentation and Peer Review (10%)

Students will present their research design, analytical methods, and findings. They will respond to methodological questions and provide structured feedback on the work of their peers. Assessment will emphasize clarity of communication, analytical depth, methodological justification, appropriate interpretation of results, and critical engagement with limitations.

Assessments: Activity	Percent Contribution
Reading Memos and Seminar Participation	10%
Text Data Acquisition and Reproducible Pipeline	15%
Method Comparison and Validation Report	20%
Novel Data Source Research Proposal	15%
Final Text-as-Data Research Project	30%
Project Presentation and Peer Review	10%

Grading:

Final grades will be based on the sum of all possible course points as noted above.

Grade	Percentage of available points
A	94-100
A-	90-93
B+	87-89
B	84-86
B-	80-83
C+	77-79
C	74-76
C-	70-73
D	64-69
D-	60-63
F	0-59

Course Schedule:

The schedule of activities is subject to change at the reasonable discretion of the instructor. Minor changes will be announced in class, major ones provided in writing.

ECON 5123 Schedule		
Lecture	Topic	Readings
L1	Introduction to Text as Data: Research Questions, Scope, and Epistemology	Grimmer, Roberts & Stewart (2022), Introduction
L2	From Substantive Theory to Text-Based Measurement	Grimmer, Roberts & Stewart (2022), selected sections on measurement and research design
L3	Novel Data Sources: Opportunities, Provenance, and Coverage	Salganik (2018), <i>Bit by Bit</i> , selected chapters on digital traces and data collection

L4	Research Design, Sampling, Selection, and Generalizability	Groves et al. (2009), <i>Survey Methodology</i> , selected sections on survey error and data quality
L5	Text Acquisition, APIs, Archives, and Web Data	Silge & Robinson (2017), selected sections on obtaining and importing text; instructor-provided API documentation
L6	Corpus Construction, Data Management, and Documentation	Benoit et al. (2018), “quanteda: An R Package for the Quantitative Analysis of Textual Data”
L7	Text Cleaning, Tokenization, and Linguistic Preprocessing	Jurafsky & Martin, <i>Speech and Language Processing</i> , selected chapters on words and tokenization
L8	Exploratory Text Analysis and Descriptive Measurement	Silge & Robinson (2017), chapters on tidy text and frequency analysis
L9	Dictionary Methods and Theory-Driven Content Analysis	Grimmer & Stewart (2013), “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”
L10	Sentiment, Emotion, Stance, and Construct Validity	Mohammad (2016), “Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text”
L11	Supervised Classification: Training Data and Model Evaluation	Jurafsky & Martin, <i>Speech and Language Processing</i> , selected sections on text classification
L12	Annotation Design, Human Coding, and Inter-Annotator Agreement	Artstein & Poesio (2008), “Inter-Coder Agreement for Computational Linguistics”
L13	Unsupervised Learning: Clustering and Document Similarity	Manning, Raghavan & Schütze (2008), <i>Introduction to Information Retrieval</i> , selected chapters on clustering and similarity
L14	Topic Models: Concepts, Estimation, and Interpretation	Blei (2012), “Probabilistic Topic Models”
L15	Evaluating Topic Models and Validating Text-Derived Measures	Chang et al. (2009), “Reading Tea Leaves: How Humans Interpret Topic Models”
L16	Word Embeddings and Semantic Representations	Mikolov et al. (2013), “Efficient Estimation of Word Representations in Vector Space”
L17	Contextual Embeddings, Transformers, and Language Models	Vaswani et al. (2017), “Attention Is All You Need”
L18	Large Language Models for Research: Prompting, Extraction, and Classification	OpenAI (2023), “GPT-4 Technical Report”; instructor-provided methodological readings
L19	Multilingual Text, Translation, and Cross-Cultural Comparability	Conneau et al. (2020), “Unsupervised Cross-lingual Representation Learning at Scale”

L20	Integrating Text with Surveys, Administrative, and Economic Data	Gentzkow, Kelly & Taddy (2019), "Text as Data"
L21	Causal Inference, Prediction, and the Limits of Text-Based Claims	Athey & Imbens (2019), "Machine Learning Methods That Economists Should Know About"
L22	Bias, Fairness, Privacy, Copyright, and Responsible Data Governance	Bender et al. (2021), "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"
L23	Reproducibility, Robustness, and Research Transparency	Nosek et al. (2015), "Promoting an Open Research Culture"
L24	Research Workshop: Project Clinics and Peer Review	Students' selected project-specific literature; Grimmer, Roberts & Stewart (2022), relevant chapters
L25	Student Presentations, Course Synthesis, and Future Directions	Review of core readings; Grimmer, Roberts & Stewart (2022), conclusion

Accommodation Statement:

Academic accommodations may be made for any student who notifies the instructor of the need for an accommodation. It is imperative that you take the initiative to bring such needs to the instructor's attention, as he/she is not legally permitted to inquire. Students who may require assistance in emergency evacuations should contact the instructor as to the most appropriate procedures to follow.

Academic Integrity Statement

Each student is expected to maintain the highest standards of honesty and integrity in academic and professional matters. The University reserves the right to take disciplinary action, up to and including dismissal, against any student who is found guilty of academic dishonesty or otherwise fails to meet the standards. Any student judged to have engaged in academic dishonesty in coursework may receive a reduced or failing grade for the work in question and/or for the course.

Academic dishonesty includes, but is not limited to, dishonesty in quizzes, tests, or assignments; claiming credit for work not done or done by others; hindering the academic work of other students; misrepresenting academic or professional qualifications within or without the University; and nondisclosure or misrepresentation in filling out applications or other University records.

Other Items:

Attendance and Expectations

All students are required to attend every class, except in cases of illness, serious family concerns, or other major problems. We expect that students will arrive on time, be prepared to listen and participate as appropriate, and stay for the duration of a meeting rather than drift in or out casually. In short, we anticipate that students will show professors and fellow students maximum consideration by minimizing the disturbances that cause interruptions in the learning process. This means that punctuality is a must, that cellular phones be turned off, and that courtesy is the guiding principle in all exchanges among students and faculty. You will be responsible for the materials and ideas presented in the lecture.

Assignment Due Dates

All written assignments must be turned in at the time specified. Late assignments will not be accepted unless prior information has been obtained from the instructor. If you believe you have extenuating circumstances, please contact the instructor as soon as possible.

Make-Up Work

The instructor will not provide students with class information or make-up assignments/quizzes/exams missed due to an unexcused absence. Absences will be excused and assignments/quizzes/exams may be made up only with written documentation of an authorized absence. Every effort should be made to avoid scheduling appointments during class. An excused student is responsible for requesting any missed information from the instructor and setting up any necessary appointments outside of class.

Access, Special Needs, and Disabilities

Please notify the instructor at the start of the semester if you have any documented disabilities, a medical issue, or any special circumstances that require attention, and the school will be happy to assist.