SOSC 314

Computational Social Science: Tools to Collect & Analyze Human Behavior

Session 3, 2023-24

Course meeting time: TuTh 2:45pm – 5:15pm
Course meeting location: LIB 2121
Course format: Lecture + Discussion + Lab
Academic credit: 4

Instructor's information

Markus Neumann
Assistant Professor of Political Science and Computational Social Science
Email: markus.neumann@dukekunshan.edu.cn
Office: WDR 1104
Office hours: Monday & Wednesday 2-4pm or by appointment
Professional website: https://markusneumann.github.io/

I am an Assistant Professor of Political Science and Computational Social Science at DKU. My research revolves around the application of statistics and machine learning methods to social science data, especially text, images and audio. The substantive focus of my research is political advertising.

Getting in touch with me
Feel free to send me an email about any questions you may have. I will try to respond within 24 hours. If you don't receive a response in that time, feel free to email me again. You can also come to my office hours, which are held on the same day of the week on which the homework assignments are due.

What is this course about?

This course explores the interdisciplinary field of computational social science, drawing from sociology, political science, computer science, and related disciplines. You will obtain skills to automate the collection of social science data from new sources (i.e., text-as-data), to classify unstructured data into discrete variables, and to analyze them using a combination of techniques that includes natural language processing and machine learning approaches. Complex ethical and legal issues arise when working with these novel sources of data (e.g., privacy and security

concerns, possibility of server overload, etc). You will be able to develop your imagination about new questions that can be asked with these new data sources. We will also read and discuss exemplary studies produced by computational social scientists.

What background knowledge do I need before taking this course?

Prerequisite: (MATH 101 or 105, and STATS 101) or (MATH 205) or Consent of the Instructor
** For students interested in the Computation and Design major, they should follow the sequence and take MATH 101/105 before taking this course. **

- Some background in applied/theoretical statistics and/or computer science is recommended.
- Prior knowledge in programming is not required but would be beneficial.
  - Students with no or little previous experience in Python: No problem, we will cover some basic Python programming in class.
- The most desirable prerequisite is a willingness to learn and apply unfamiliar course materials.
- In practice, I will likely give students who have taken STATS 101, but not MATH 101 or 105 permission to take the class, since it is not a math-heavy class.

What will I learn in this course?

By the end of this course, you will be able to:

1. Explain and evaluate the nature, strengths and weaknesses, and analytic approaches of digitized texts
2. Describe ethical and legal issues associated with working with online text data
3. Critically interpret the findings of exemplary studies in computational social science
4. Collect and analyze various sources of text data using Python
5. Formulate innovative types of research questions that can be answered with the new data sources and develop an independent research proposal to answer the questions

What will I do in this course?

Before each class session

- Read weekly reading materials

During each class session

- Lecture and discussion
- Lab and practice

Weekly assignments

- Post reading reflection (3 times throughout the semester)
- Complete homework assignments
- Submit final project-related assignments

How can I prepare for the class sessions to be successful?

Students should read all assigned materials before the class and consider it carefully. They should come to ask ready to discuss the material and ask any questions they have to the professor.

What required texts, materials, and equipment will I need?

Texts and Materials

- Instead of using a specific textbook, we will read relevant articles from related fields.
- All readings will be provided as pdfs or accessible for free online.
- Lecture slides and code will be posted on Canvas after each class session.
- If you are looking for a textbook for reference, see Speech and Language Processing by Jurafsky and Martin, which is not only the best NLP textbook, but also free!

Equipment and Computational Tools

- Python will be the primary programming language in this course. We may use R sometimes.
- Please bring your laptop to the class. You can borrow a laptop from DKU library: https://dukekunshan.edu.cn/en/academics/library/using

How will my grade be determined?

Summary of Graded Assignments

- **Reading reflections: 15%** (5% * 3) W2-7, choose 3 readings.
- **Homework exercises: 20%** (5% * 4) W2-5
- **Final project: 50%**
  - Proposal (5%)
  - Progress presentations (5% * 2 = 10%)
  - Peer feedback (5% * 2 = 10%)
  - Final paper (10%)
  - Final project oral exam (15%)
- **In-class participation: 15%**
  - Attendance & punctuality: 10%
  - Discussion Participation: 5%

Grade Scale

A+= 98% - 100%; **A** = 93% - 97.9%; **A-** = 90% - 92.9%; **B+** = 87% - 89.9%; **B** = 83% - 86.9%; **B-** = 80% - 82.9%; **C+** = 77% - 79.9%; **C** = 73% - 76.9%; **C-** = 70% - 72.9%; **D+** = 67% - 69.9%; **D** = 63% - 66.9%; **D-** = 60% - 62.9%; **F** = 59.9% and below.

For final grades, .05 is rounded up. For example, a 92.94999 is an A-, a 92.95 is an A. Grades are non-negotiable and can only be changed due to an error in calculation or transcription.

What are the course policies?

Academic Integrity
As a student, you should abide by the academic honesty standard of the Duke Kunshan University. Its Community Standard states: "Duke Kunshan University is a community comprised of individuals from diverse cultures and backgrounds. We are dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Members of this community commit to reflecting upon and upholding these principles in all academic and non-academic endeavors, and to protecting and promoting a culture of integrity and trust."

Generative AI Guidelines and Policy
Generative AI (including ChatGPT and Quillbot, as well as numerous tools in media and art) is a novel technology to which higher education is learning to adapt. In this course, **the use of generative AI is explicitly allowed**, but still governed by DKU's rules:

- Use of these tools is governed by DKU's Academic Integrity Policy, and students must employ this technology consistent with expectations of the instructor, course, or assessment.
- In any situations in which such tools are used (with or without permission) in the process of completing assignments, students are obliged to cite fully any use of generative AI tools in the formulation of their work, including by preserving a record of the use of the tool as original source material.
- Students should be encouraged to save all rough drafts and notes for papers, in case any concerns arise.

DKU also has a licensed version of ChatGPT that you can (but don't have to) use: https://app.dukekunshan.edu.cn/

Academic Policy & Procedures
You are responsible for knowing and adhering to academic policy and procedures as published in University Bulletin and Student Handbook. Please note, an incident of behavioral infraction or academic dishonesty (cheating on a test, plagiarizing, etc.) will result in immediate action from me, in consultation with university administration (e.g., Dean of Undergraduate Studies, Student Conduct, Academic Advising). Please visit the Undergraduate Studies website for additional guidance related to academic policy and procedures.

Academic Disruptive Behavior and Community Standard
Please avoid all forms of disruptive behavior, including but not limited to: verbal or physical threats, repeated obscenities, unreasonable interference with class discussion, making/receiving personal phone calls, text messages or pages during class, excessive tardiness, leaving and entering class frequently without notice of illness or other extenuating circumstances, and persisting in disruptive personal conversations with other class members. If you choose not to adhere to these standards, I will take action in consultation with university administration (e.g., Dean of Undergraduate Studies, Student Conduct, Academic Advising).

Academic Accommodations
If you need to request accommodation for a disability, you need a signed accommodation plan from Campus Health Services, and you need to provide a copy of that plan to me. Visit the Office of Student Affairs website for additional information and instruction related to accommodations.

Class Attendance
As a seminar class, it is vital that all students attend and participate. Class attendance is therefore mandatory and absences will only be excused for serious medical and personal matters. If you will be absent from a class for a university-sponsored activity, please make arrangements with me — beforehand — regarding any work you might miss.

Late Penalties
This course will move quickly so therefore it is imperative that you do not fall behind by submitting late homework. Therefore, homework will be accepted only until 48 after the due date and will be subject to a 50% penalty. Rescheduling of homework/final due dates will only be permitted for serious medical and personal matters, and requires advance notice. Unless stated otherwise, all assignments are due at 11:59pm China time.

What campus resources can help me during this course?

Academic Advising and Student Support
Please consult with me about appropriate course preparation and readiness strategies, as needed. Consult your academic advisors on course performance (i.e., poor grades) and academic decisions (e.g., course changes, incompletes, withdrawals) to ensure you stay on track with degree and graduation requirements. In addition to advisors, staff in the Academic Resource Center can provide recommendations on academic success strategies (e.g., tutoring, coaching, student learning preferences). Note, there is an ARC Sakai site for students and tutors. Please visit the Office of Undergraduate Advising website for additional information related to academic advising and student support services.

Writing and Language Studio
If you want additional help with academic writing and more generally with language learning you are welcome to go to Writing and Language Studio (WLS). You can register for an account, make an appointment, and learn more about WLS services, policies, and events on the WLS website. You can also find writing and language learning resources on the Writing & Language Studio Sakai site.

IT Support
If you are experiencing technical difficulties, please contact IT:

- China-based faculty/staff/students 400-816-7100, (+86) 0512- 3665-7100
- US-based faculty/staff/students (+1) 919-660-1810
- International-based faculty/staff/students can use either telephone option (recommend using tools like Skype calling)
- Live Chat: https://oit.duke.edu/help

- Email: service-desk@dukekunshan.edu.cn

What is the expected course schedule?

Dates	Topic	Readings	Assignments due
W1D1 Jan. 9	Course overview & Python	Computational Social Science	
W1D2 Jan. 11	Python	Adapting computational text analysis to social science (and vice versa) Preface: Big Data Is Not About The Data!	
W2D1 Jan. 16	Data collection – APIs	Spatial disparity of income-weighted accessibility in Brazilian Cities: Application of a Google Maps API	HW1 (Jan 17) Final project proposal (Jan 19)
W2D2 Jan. 18	Data collection – Web scraping	New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings	
W3D1 Jan. 23	Intro to machine learning	Beyond prediction: Using big data for policy problems	HW2 (Jan 24) Final project – progress presentation 1 (Jan 25)
W3D2 Jan. 25	Text-as-data – Bag of words	What Have Economists Been Doing for the Last 50 Years? A Text Analysis of Published Academic Research from 1960–2010	
W4D1 Jan. 30	Supervised ML for text classification	Automated Text Classification of News Articles: A Practical Guide	HW3 (Jan 31) Final project – peer feedback 1 (Feb 1)
W4D2 Feb. 1	Preprocessing	Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it	
W5D1 Feb. 6	Topic models	Measuring Philosophy in the First Thousand Years of Greek Literature	HW4 (Feb 7) Final project – progress presentation 2 (Feb 8)
W5D2 Feb. 8	Text similarity	Text as Policy: Measuring Policy Similarity through Bill Text Reuse	
W6D1 Feb. 20	Fightin’ words	Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict	Final project – peer feedback 2
W6D2 Feb. 22	Traditional NLP	Named Entity Recognition and Classification in Historical Documents: A Survey	
W7D1	Word embeddings	Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes	Final paper (Feb 28)

Feb. 27			
W7D2 Feb. 29	Modern NLP	Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation	
Finals Week			Final project oral exam (Mar 5)