Soc591 Syllabus: Introduction to Computational Social Science

Yongjun Zhang, PhD*

January 20, 2022

Time and Location: Monday 11:45 AM - 2:35 PM; Room N403
Email: Yongjun.Zhang@stonybrook.edu
Virtual Office Hours: Appointments as needed

Course Description

A multidisciplinary introduction to computational social science (CSS), emphasizing how social scientists develop and make use of computational-related social theory and methods to understand and analyze social behavior in the digital era. Topics include the CSS history and its latest development as well as how to use computational methods to collect, process, analyze, and visualize large-scale data from the real world to address social problems. This course also introduces state-of-the-art tools such as Python, R, and its scientific libraries for web-scraping, natural language processing, topic modeling, and machine learning techniques for text, image, and video data.

Course Learning Objectives

This course provides students a set of computational social science toolkits to acquire the knowledge or skills necessary to achieve the following learning outcomes:

1. Understand the history and development as well as the major concepts in computational social science.
2. Understand the methods of inquiry used by social scientists to explore social and behavioral phenomena.
3. Skillfully interpret and form educated opinions on social science issues.
4. Master the ability to apply computational tools and knowledge to problem-solving.
5. Design and build computational systems to explore and analyze some aspects of the human world.

*Yongjun Zhang is an assistant professor in the department of sociology and the institute for advanced computational science at the State University of New York at Stony Brook. He is also a research affiliate at New York University. Email: Yongjun.Zhang@stonybrook.edu

Textbooks and Other Useful Materials

Required:

1. Hadley Wickham and Garrett Grolemund. 2016. R for Data Science. https://r4ds.had.co.nz/
2. Hadley Wickham. Advanced R. https://adv-r.hadley.nz/
3. Kieran Healy. Data Visualization. https://socviz.co/

Optional:

1. Matthew J. Salganik. 2017. Bit by Bit: Social Research in the Digital Age. Princeton University Press. http://www.bitbybitbook.com
2. Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python. O'Reilly Media. https://www.nltk.org/book/

1 Course Requirements and Evaluation

The course will be broadly divided into three modules: CSS basics, Text as Data, and Image/Audio/Video as Data. We will focus on the first two parts. Each module will introduce the basic and latest methods as well as relevant social research using associated methods. Students can choose the specific module that is particularly helpful in their research to develop their final research proposal. For each meeting, it is a mix of mini-lecture, student mini-presentation, instructor or student-led discussions, and lab training. In the mini-lecture, the instructor will give an overview of the field and the development of research methods. In the mini-lab, the instructor will use R, Python, and relevant programming languages to walk through each method and prepare students with the necessary computational skills for their research. Students will be evaluated based on the following aspects:

Research Proposal.

Students are required to develop a short research proposal (or use an existing one) integrating at least one of the methods introduced in the course. Particularly students are required to use large-scale administrative and digital trace data hosted by Google BigQuery and other platforms or scraped by themselves.

Research Presentation.

Students are required to present the research proposal and final project in the class. Students are expected to use at least one of the CSS methods learned in the class.

Coding Challenge.

Students are required to complete four coding challenges. The instructor will release problem sets as the semester progresses. Students need to submit a code report as well as the R or python script.

Class Participation.

Students are required to attend every session and prepare the assigned readings before each session. Students are also required to participate in the class discussion. Students will receive no credits if they have over 3 times of absence.

Course Policies

Make-up exams are not allowed without prior arrangements and documentation of extenuating circumstances. Please speak with the instructor regarding any known absences or emergencies ASAP to avoid any issues regarding assignment days. Late assignments are not accepted without prior arrangements with the instructor.

We all share responsibility for maintaining an appropriate learning environment. For this reason, please mute yourself if attending zoom meetings or when others are speaking so that your peers are not distracted. Finally, all offline or online classroom behavior and discourse should reflect the values of respect and civility.

Composition of Final Grades

Research Proposal x 1	15
Research Paper x 1	30
Research Presentation x 1	20
Coding Challenge (4 x 5)	20
Class Participation	15
Total	100 points

Grade Scale:

95-122 = A	75-79 = B-	0-59 = F
90-94 = A-	70-74 = C+	
85-89 = B+	65-69 = C	
80-84 = B	60-64 = C-	

Schedule

Note: The instructor reserves the right to modify the schedule as deemed necessary. Flexibility throughout the semester will allow us to incorporate the latest computational social science methods into the course.

Welcome and Introduction to CSS

WEEK 1 (01-24-2022) A General Introduction to CSS

Topics: CSS, Big Data, and Data Science.

Assigned Readings:

1. Lazer et al. 2009. "Computational Social Science." Science.
2. David M. J. Lazer et al. 2020. "Computational social science: Obstacles and opportunities." Science, 369, 6507, Pp. 1060-1062. Publisher's Version Copy at https://j.mp/2YIuWdh
3. Edelman et al. 2020. "Computational Social Science and Sociology." Annual Review of Sociology. https://doi.org/10.1146/annurev-soc-121919-054621
4. David Donoho. 2015. 50 Years of Data Science. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Lab:

1. Github and version control; understand basic git commands, like git clone, git fetch, git pull, and git push;
2. install all necessary software like python, R, Rstudio, notebook, colab, github desktop, etc;
3. Understand how to use the command line, like how to run Python using terminal, etc.
4. Get Twitter Academic API for later use.

Module 1 CSS Basics

In this module, we will briefly discuss how other computational social scientists think about CSS, how to use R or Python from scratch, regular expression, machine learning, and data visualization.

WEEK 2 (01-31-2022) Conceptualizing CSS and Programming

Topics: methodological approach; algorithm bias; measurement bias; research ethics; etc.

Assigned Readings:

1. Laura K. Nelson. 2017. Computational Grounded Theory: A Methodological Framework. Sociological Methods and Research.
2. Obermeyer, Ziad, et al. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." Science 366.6464: 447-453.
3. Schwemmer, Carsten, et al. 2020. "Diagnosing gender bias in image recognition systems." Socius.
4. Wagner, Claudia, et al. 2021. "Measuring algorithmically infused societies." Nature 595.7866: 197-204.
5. Lazer, David, et al. 2021. "Meaningful measures of human society in the twenty-first century." Nature 595.7866: 189-196.
6. R for data science. Chapter 1-3. (Lab reading, Pls spend some time reading these chapters)

Lab:

1. Basic programming in R or python
2. Understand how to read/save files
3. Basic data wrangling using tidyverse etc.
4. Understand regular expression

WEEK 3 (02-7-2022) Web Scraping, API, and Big Data

Topics: Introducing web scraping and API.

Assigned Readings:

1. Sobel, Benjamin LW. "A New Common Law of Web Scraping." Lewis Clark L. Rev. 25 (2021): 147.
2. Li, Fumin, Yisu Zhou, and Tianji Cai. "Trails of data: Three cases for collecting web information for social science research." Social Science Computer Review (2019).
3. George, Gerard, et al. "Big data and data science methods for management research." Academy of Management Journal 59.5 (2016): 1493-1507.
4. Twitter Academic API: https://developer.twitter.com/en/products/twitter-api/academic-research
5. Google Cloud APIs: https://cloud.google.com/apis
6. Google API Key: https://developers.google.com/maps/documentation/maps-static/get-api-key
7. Google BigQuery: https://cloud.google.com/bigquery/docs/quickstarts
8. Google SDK Guide: https://cloud.google.com/sdk/docs/how-to

Lab:

1. Using R or python to scrape data from websites or social media platforms (e.g., twitter) (code challenge)
2. Understand how to use webdriver in data scraping
3. Understand how to use Google Cloud Service/Google API (e.g., How to use python to get data from bigquery)

WEEK 4 (02-14-2022) Machine Learning

Topics: Basics on supervised machine learning

Assigned Readings:

1. Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. "Machine Learning for Social Science: An Agnostic Approach." Annual Review of Political Science 24 (2021): 395-419.
2. Molina, Mario, and Filiz Garip. "Machine learning for sociology." Annual Review of Sociology 45 (2019): 27-45.
3. Optional:Athey, Susan, and Guido W. Imbens. "Machine learning methods that economists should know about." Annual Review of Economics 11 (2019): 685-725.
4. Optional:Max Kuhn and Kjell Johnson. Applied Predictive Modeling.
5. Optional:R caret Package: https://topepo.github.io/caret/

Lab:

1. Using R caret package to do some basic supervised machine learning
2. Train a model to predict the gender of U.S. baby names using SSA data (code challenge)
3. SSA baby name data: https://catalog.data.gov/dataset/baby-names-from-social-security-car

WEEK 5 (02-21-2022) Data Visualization

Assigned Readings:

1. Healy, K. and Moody, J., 2014. Data visualization in sociology. Annual review of sociology, 40, pp.105-128.
2. Healy, K., 2018. Data visualization: a practical introduction. Princeton University Press. Chapter 1-4
3. Wickham, Hadley. "Programming with ggplot2." ggplot2. Springer, Cham, 2016. 241-253. (Optional)

Lab:

1. Understand how to use ggplot2 in R
2. Replicate the Twitter Hate Speech graphs using ggplot2

Module 2 Text As Data

WEEK 6 (02-28-2022) Text as Data (1)-Representation, Discovery, Measurement, and Causal Inference

Topics: Natural Language Processing; Big Data and Parallel Computing

Assigned Readings:

1. Hirschberg, Julia, Manning, Christopher D. 2015. "Advances in Natural Language Processing." Science 349:261–66.
2. Grimmer, Justin, Stewart, Brandon M. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Political Analysis 21:267–97.
3. Monroe, Burt L., Colaresi, Michael P., Quinn, Kevin M.. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." Political Analysis 16:372–403.
4. Nardulli, Peter F., Althaus, Scott L., Hayes, Matthew. 2015. "A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data." Sociological Methodology 45:148–83. (Think about how to use text to generate data)

Lab:

1. Introduction to basic text analysis
2. How to run basic NLP tasks in R or Python
3. Read Chapters 1-3. Natural Language Processing with Python.

WEEK 7 (03-7-2022) Text as Data (2)-Topic Modeling

Topics: LDA and Structural Topic Model; Model Applications.

Assigned Readings:

1. Blei, David M. 2012. "Probabilistic Topic Models." Communications of the ACM 55:77–84. (LDA)
2. Mohr, John W., Bogdanov, Petko. 2013. "Introduction—Topic Models: What They Are and Why They Matter." Poetics 41 (6): 545–69.
3. DiMaggio, Paul, Nag, Manish, Blei, David. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." Poetics 41:570–606.
4. Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B. and Rand, D.G., 2014. Structural topic models for open-ended survey responses. American Journal of Political Science, 58(4), pp.1064-1082. (STM)
5. Roberts, M.E., Stewart, B.M. and Tingley, D. 2014. stm: R package for structural topic models. Journal of Statistical Software, 10(2), pp.1-40.****** (A package intro paper)

Lab:

1. Steps to implement topic models via R (R library stm and topicmodels).

WEEK 8 (03-14-2022) Spring Break No class

WEEK 9 (03-21-2022) Text as Data (3)-Word Embedding and Transformers

Topics: Word embedding and transformers

Assigned Readings:

1. Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. PNAS 201720347 (2018). doi: 10.1073/pnas.1720347115
2. Nelson, Laura K. "Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century US South." Poetics (2021): 101539.
3. Wankmüller, Sandra. "Neural Transfer Learning with Transformers for Social Science Text Analysis." arXiv preprint arXiv:2102.02111 (2021).
4. Optional:Kozlowski, A.C., Taddy, M., and Evans, J.A., 2019. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. American Sociological Review.
5. Optional:Rong, Xin. "word2vec parameter learning explained." arXiv preprint arXiv:1411.2738 (2014).
6. Optional:Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013. https://arxiv.org/pdf/1310.4546.pdf
7. Optional:Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Lab:

1. How to use word embedding/transformer in R or Python to achieve NLP tasks

WEEK 10 (03-28-2022) Text as Data (4)-Sentiment Analysis

Assigned Readings:

1. Paxton, Pamela, Kristopher Velasco, and Robert W. Ressler. "Does use of emotion increase donations and volunteers for nonprofits?." American Sociological Review 85.6 (2020): 1051-1083.
2. Flores, René D. "Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data." American Journal of Sociology 123.2 (2017): 333-384.
3. Hassan, Tarek A., et al. "Firm-level political risk: Measurement and effects." The Quarterly Journal of Economics 134.4 (2019): 2135-2202.
4. Naldi, Maurizio. "A review of sentiment computation methods with R packages." arXiv preprint arXiv:1901.08319 (2019).
5. Cook, Gavin, Junming Huang, and Yu Xie. "How COVID-19 has Impacted American Attitudes Toward China: A Study on Twitter." arXiv preprint arXiv:2108.11040 (2021).

Lab:

1. How to conduct sentiment analysis in R or Python.
2. Use Twitter data to train a sentiment analysis model (code challenge).

Module 3 Image/Audio/Video as Data

WEEK 11 (04-4-2022) Image as Data (1)

Topics: why image as data? How do we use images for social research?

Assigned Readings:

1. Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., and Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), pp.790-794.
2. Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., and Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), pp.790-794. Read the supplement: https://science.sciencemag.org/content/sci/suppl/2016/08/19/353.6301.790.DC1/Jean.SM.pdf
3. Han Zhang and Jennifer Pan. 2019. CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media. Sociological Methodology.
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248–255). Ieee.

Lab:

1. Basic knowledge about using R or Python to obtain and process images
2. Extra Resource: http://neuralnetworksanddeeplearning.com/

WEEK 12 (04-11-2022) Image as Data (2)

Topics: Methods to implement image recognition, extract useful image information, machine learning methods, transfer learning approach, etc.

Assigned Readings:

1. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. Nature, 521(7553), pp.436-444.
2. Li, Shan, and Weihong Deng. "Deep facial expression recognition: A survey." IEEE transactions on affective computing (2020).
3. Liu, Zhuang, et al. "A ConvNet for the 2020s." arXiv preprint arXiv:2201.03545 (2022).
4. Watch the video and read notes of Introduction to Convolutional Neural Network: http://cs231n.github.io/convolutional-networks/
5. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097–1105).
6. Pan, S.J. and Yang, Q., 2010. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), pp.1345–1359.

Lab: Image data storage, cleaning, and processing in Python

WEEK 13 (04-18-2022) Image as Data (3)

Topics: how to use Keras (or Pytorch) frameworks to analyze image data and train your neural network.

Assigned Readings:

1. Chollet, Francois. Deep learning with Python. Simon and Schuster, 2021. Chapter 2-3.
2. TensorFlow Tutorial: https://www.tensorflow.org/tutorials
3. Explore some code examples: https://keras.io/examples/
4. Pytorch Tutorial: https://pytorch.org/tutorials/beginner/deep_learning_60min_blitz.html
5. Check facenet project: https://github.com/davidsandberg/facenet
6. Check openface project:]urlhttp://cmusatyalab.github.io/openface/

Lab

1. Google Cloud Service; TensorFlow; Keras; R packages (torch or keras).
2. An example using pytorch to replicate the Jean et al's result: https://github.com/joshzyj/predicting-poverty-replication
3. Use google map api to obtain google images/Use Jean's model to predict economic outcomes/visualize and map the outcomes (code challenge)

WEEK 14 (04-25-2022) Audio and Video data

Topics: Introducing basic methods to process acoustic data and speech recognition; Introducing basic methods to process video data; introducing social research based on video data

Assigned Readings:

1. Dietrich, Bryce J., Matthew Hayes, and Diana Z. O'brien. "Pitch perfect: Vocal pitch and the emotional intensity of congressional speech." American Political Science Review 113.4 (2019): 941-962.
2. Dietrich, Bryce J. "Using motion detection to measure social polarization in the US House of Representatives." Political Analysis 29.2 (2021): 250-259.
3. Qin and Yang. 2019. What you say and how you say it matters. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy.
4. Choudhury, Prithwiraj, et al. "Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles." Strategic Management Journal 40.11 (2019): 1705-1732.

Lab:

1. Using Keras, Tensorflow, and DeepSpeech to build and train a speech recognition model and object detection model.
2. Explore Essentia: https://essentia.upf.edu/documentation.html and https://mtg.github.io/essentia-labs/news/tensorflow/2020/01/16/tensorflow-models-released/
3. Explore librosa:https://librosa.org/doc/latest/index.html
4. Stanford Cable TV News Analyzer: https://tvnews.stanford.edu/methodology
5. YOLO3: https://pjreddie.com/media/files/papers/YOLOv3.pdf
6. Explore YOLO3 (You Only Look Once–for Object Detection): https://www.youtube.com/watch?v=MPU2HistivI&feature=youtu.be&ab_channel=JosephRedmon
7. Here is the instruction of how to use YOLO3: https://pjreddie.com/darknet/yolo/

WEEK 15 (05-2-2022) Research Paper Presentation

Technical and Software Requirements

Students need a stable internet environment to access the Blackboard and Zoom. If you have any questions or difficulties, please let me know. You also need to set up Google Cloud for the use of free Google Colab and other cloud computing services. You should also install R, Rstudio, Python3, etc. Students will have instructions to install those free software and associated packages or modules.

If you need technical assistance at any time during the course or to report a problem with Blackboard you can:

1. Phone: 631-632-9800 (client support, Wi-Fi, software and hardware)
2. Submit a help request ticket: https://it.stonybrook.edu/services/itsm
3. If you are on campus, visit the Walk-Up Tech Support Station in the Educational Communications Center (ECC) building.
4. For laptop loans: https://www.stonybrook.edu/commcms/studentaffairs/studentsupport/ for IT support: https://it.stonybrook.edu/services/itsm

Student Accessibility Support Center Statement

If you have a physical, psychological, medical or learning disability that may impact your course work, please contact the Student Accessibility Support Center, ECC (Educational Communications Center) Building, Room 128, (631)632-6748. They will determine with you what accommodations, if any, are necessary and appropriate. All information and documentation are confidential.

Academic Integrity Statement

Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person's work as your own is always wrong. Faculty are required to report any suspected instances of academic dishonesty to the Academic Judiciary. Faculty in the Health Sciences Center (School of Health Technology Management, Nursing, Social Welfare, Dental Medicine) and School of Medicine are required to follow their school-specific procedures. For more comprehensive information on academic integrity, including categories of academic dishonesty please refer to the academic judiciary website at http://www.stonybrook.edu/commcms/academic_integrity/index.html

Critical Incident Management

Stony Brook University expects students to respect the rights, privileges, and property of other people. Faculty are required to report to the Office of Student Conduct and Community Standards any disruptive behavior that interrupts their ability to teach, compromises the safety of the learning environment, or inhibits students' ability to learn. Until/unless the latest COVID guidance is explicitly amended by SBU, during Spring 2022 "disruptive behavior" will include refusal to wear a mask during classes. For the latest COVID guidance, please refer to: https://www.stonybrook.edu/commcms/strongertogether/latest.php

Copyright Notice

Other Useful Resources

1. Student Success Resources: A helpful resource is the "For Students" section linked from the Stony Brook homepage: http://www.stonybrook.edu/for-students/ as well as the Division of Undergraduate Education website: http://www.stonybrook.edu/due.

2. Academic Success and Tutoring Center: This important program opened in September 2013. Please be sure that your students are aware of the available services. Information can be found at: http://www.stonybrook.edu/commcms/academic_success/

3. Support for Online Learning: https://www.stonybrook.edu/online/

4. Writing Center: Students are able to schedule face-to-face and online appointments. https://www.stonybrook.edu/writingcenter/

5. Career Center: The Career Center's mission is to support the academic mission of Stony Brook University by educating students about the career decision-making process, helping them plan and attain their career goals, and assisting with their smooth transition to the workplace or further education. Phone: 631-632-6810; email: sbucareer-center@stonybrook.edu; website: http://www.stonybrook.edu/career-center/

6. Counseling and Psychological Services: CAPS staff are available by phone, day or night. http://studentaffairs.stonybrook.edu/caps/