Theories, Methods, and Practices in Computational Social Science

Course Overview

Social science research is now regularly conducted using data that are larger and more complex than the data conventional statistical tools were designed for. Examples include individual-level consumption records, data streams collected from social media, and archives of electronic/digitized government records. These fine-grained, voluminous, and high-dimensional data require a set of methods that are more sophisticated than the conventional toolkits of quantitative social science, which would allow us to learn patterns from the data, select variables with predictive power, reduce the dimensionality of data by grouping variables, and make predictions based on these variables. In this course, we will learn the basic principles of machine learning methods and how they are applied to solve social science problems. This course focuses on textual data – one of the most common data types in mainstream computational social science research.

Course Objectives

1. Develop a basic understanding of the theory behind machine learning methods
2. Understand how to explore, describe, and visualize data
3. Apply machine learning methods to real-life data, with a particular focus on text
4. Appreciate how machine learning methods are applied to solve social science problems

Prerequisites

This course is an intensive course for advanced undergraduate and postgraduate students interested in machine learning and automated text analysis. Students enrolled in this course are expected to have basic knowledge of statistics, quantitative methods, and programming. Students who want to develop a future research agenda in computational social science are especially welcome.

Assessment

- 10% - In-class participation
  - Active participation is required
- 40% - Individual coding exercise x 4
  - Solving specific tasks with methods learned in class
- 50% - Group Project (5% research prospectus; 15% presentation; 30% final report)
  - 2 – 3 students form a group
  - A group project that applies the methods learned in this class to solve an issue in students' own disciplines

Course Structure

The weekly class is divided into three 45-minute sessions. The first session covers basic theoretical concepts and walks through applying relevant methods in the assigned readings (if any). The second and third sessions are labs in which all students follow the instructor to run code and complete in-class coding exercises.

Weekly Schedule and Readings
The schedule below lists the topics covered in the first session, the content of the lab sessions, and the required readings and applications each week. Read the readings and applications (focusing on how different methods are applied) before class that day. Full citations are provided below in the references section. Students are expected to spend, on average, 3 hours per week reading the assigned materials and working on coding.

<u>Part A Introduction to Machine Learning, R and Text Analysis</u>

**Download R and install RStudio before the start of the semester*

Week 1 Introduction

- Issues: intuition and background of machine learning, prediction vs. explanation, the online community of Data Science
- Lab: basic setup of RStudio, the basic operation of the R language in RStudio, data types and structure, conditions and iteration
- Reading: Rhys (2020) Ch1; Shmueli et al. (2010)

**Apply for YouTube API, install Chromium*

Week 2 Data wrangling and visualization

- Issues: split-apply-combine/map-reduce, common charts and plots in social science
- Lab: using packages, read and write files, data wrangling with *dplyr*, data visualization with *ggplot*, R Markdown
- Reading: Rhys (2020) Ch2; Wickham & Grolemund (2016) Ch3 & Ch27

Week 3 Problems with Big Data and data collection

- Issues: problems with Big Data, how API works, web scraping
- Lab: retrieve YouTube data using API, web-scrape data with *RSelenium*
- Reading: Salganik (2017) Ch1 & Ch2
- Application: Teo & Fu (2021)

**HW 1 due*

Week 4 Text as Data and Text (pre)-processing

- Issues: text as data, bag-of-words as text representation, text preprocessing, sparsity & hashing, Zipf's Law
- Lab: regular expressions, quanteda package: preprocessing, document-feature matrix (dfm), tokens' co-occurrence
- Reading: Grimmer & Brandon (2013); Grimmer et al. (2022) Ch5
- Application: Luqiu & Lu (2021),

Week 5 Text visualization, dictionary-based analysis, and basics of classification

- Issues: bias-variance trade-off, labeling data, inter-coder reliability (ICR), cross-validation, performance metric
- Lab: word cloud and network graph, sentiment analysis, coding a training set, ICR Test
- Reading: Rhys (2020) Ch3; Silge & Robinson (2017) Ch8
- Application: Rheault et al. (2019)

<u>Part B Supervised Methods: Classification and Regression</u>

Week 6 Naïve Bayesian and Logistics Regression
- Issues: parametric methods: pros and cons, (naïve) Bayesian rule, Logistic Regression
- Lab: Naïve Bayesian, Logistics Regression
- Reading: Rhys (2020) Ch4 & Ch6 (6.1-6.3)
- Application: Chenoweth & Ulfelder (2017)

**HW 2 due*

Week 7 Support Vector Machine (SVM), Decision Tree, and Random Forest,
- Issues: non-parametric methods: pros and cons, hyper-tuning, SVM, Decision Tree, Ensemble Methods like Random Forest
- Lab: SVM, Random Forest, XGboost
- Reading: Rhys (2020) Ch6 (6.4 and after), Ch7 & Ch8
- Application: Pan (2019)

**Research prospectus due*

Week 8 Regression and Regularization
- Issues: basic of regression, regularization: lasso, ridge, and elastic net
- Lab: Regression in machine learning, regularization
- Reading: Rhys (2020) Ch9 & Ch11
- Application: Mitts (2019)

<u>Part C Unsupervised Methods: Dimension Reduction</u>

Week 9 Principal Component Analysis (PCA) and Clustering
- Issues: PCA, various approaches to clustering
- Lab: PCA, clustering
- Reading: Rhys (2020) Ch4 13, Ch16 &17
- Application: Li (2022)

**HW 3 due*

Week 10 Topic Modeling
- Issues: Generative topic modeling, topic modeling with structure
- Lab: Latent Dirichlet Allocation (LDA), Structural Topic Modeling (STM)

- Reading: Silge & Robinson (2017) Ch6
- Application: Urman et al. (2021)

Part D Advanced Topic

Week 11 Neural Network

- Issue: Deep learning, basics of the neural network, Recurrent Neural Network (rNN), Convolutional Neural Network (cNN)
- Lab: Neural Network with non-textual data
- Reading: Günther & Fritsch (2010); Lectures 1 from "Introduction to Deep Learning." MIT: http://introtodeeplearning.com/
- Application: Zhang & Pan (2019)

*HW 4 due

Week 12 Natural Language Process (NLP), Word-embedding, Transfer learning

- Issue: NLP, word-embedding as text representation, transfer-learner: Bert
- Lab: using Bert in R
- Reading: Rodriguez & Spirling (2022); Jon Harmon – RBert - https://www.youtube.com/watch?v=1kV0fPDefBM&ab_channel=LanderAnalytics
- Application: Dai (2018)

Week 13 Generative AI

- Issue: Current development of Generative AI, prompt engineering, AI ethics
- Lab: Local use of large language model with ollama
- Reading: Palmer & Spirling (2023)
- Application: Rettenberger et al. (2025)

Week 14 Presentation

Reference

Chenoweth, E., & J. Ulfelder (2017). Can structural conditions explain the onset of nonviolent uprisings? *Journal of Conflict Resolution 61*(2), 298-324.

Dai, Y. (2018). Measuring populism in context: A supervised approach with word embedding models. *Working Paper*.

Grimmer, J and Brandon S. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents. *Political Analysis, 21*(3): 267-297.

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.

Günther, F. and S. Fritsch (2010). neuralnet: Training of neural networks. *The R journal, 2*(1), 30-38.

Li, L. (2022). Decoding political trust in China: A machine learning analysis. *The China Quarterly, 249*, 1-20.

Luqiu, L. R., & Lu, S. (2021). Bounded or boundless: A case study of foreign correspondents' use of twitter during the 2019 Hong Kong protests. *Social Media+ Society*, *7*(1), 2056305121990637.

Mitts, T. (2019). From isolation to radicalization: anti-muslim hostility and support for ISIS in the West. *American Political Science Review, 113* (1), 173- 194.

Pan, J. (2019). How Chinese officials use the internet to construct their public image. *Political Science Research and Methods,* 7 (2), 197-213.

Palmer, A., & Spirling, A. (2023). Large language models can argue in convincing ways about politics, but humans dislike AI authors: implications for governance. *Political Science*, *75*(3), 281-291.

Rettenberger, L., Reischl, M., & Schutera, M. (2025). Assessing political bias in large language models. *Journal of Computational Social Science*, *8*(2), 1-17.

Rheault, L., Rayment, E., & Musulan, A. (2019). Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, *6*(1), 2053168018816228.

Rhys, H. (2020). *Machine Learning with R, the tidyverse, and mlr*. Simon and Schuster

Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics, 84*(1), 101-115.

Salganik, Matthew J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Shmueli, G. (2010). To explain or to predict?. *Statistical Science, 25*(3), 289-310.

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly.

Teo, E., & Fu, K. W. (2021). A novel systematic approach to constructing protest repertoires from social media: comparing the roles of organizational and non-organizational actors in social movements. *Journal of Computational Social Science*, *4*(2), 787-812.

Urman, A., Ho, J. C. T., & Katz, S. (2021). Analyzing protest mobilization on Telegram: The case of the 2019 anti-extradition bill movement in Hong Kong. *Plos one*, *16*(10), e0256675.

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly.

Zhang, H., & Pan, J. (2019). CASM: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology, 49*(1), 1-57.